A COMPARISON OF IMPUTATION METHODS IN THE 2012 BEHAVIORAL RISK FACTOR SURVEILLANCE SURVEY

By

Philip Andrew Moll

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

April 2014

Department of Public Health & Preventative Medicine

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's thesis of

Philip A. Moll

has been approved

Dongseok Choi, PhD (Thesis Advisor)

Eun Sul Lee, PhD (Committee Member)

David Degras, PhD (Committee Member)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank my committee, in particular Dr. Choi, for his encouragement, direction, and advice. I can still remember the excitement I felt seeing his signature on my graduate school acceptance letter. It turns out the feeling was warranted, as Dr. Choi has been the best academic mentor I have ever had. I would also like to thank Dr. Lee and Dr. Degras, who offered their time and help reading my thesis and offered suggestions essentially as volunteers. Dr. Lee's expertise in survey design and analysis was particularly helpful, while Dr. Degras' insight and direction substantially improved the quality of this thesis. Thanks are also due to the outstanding faculty at OHSU with whom I have had the privilege of studying, especially Dr. Dawn Peters. I would also like to thank my friend Andrew Michael Roberts, for his proofreading help. Thank you as well to my family, for their support and encouragement, in particular my parents John and Dorothy Moll and parents-in-law, Steve and Susan Wetherell, and my brother Joe, whose academic achievements will always be an inspiration. And last, but definitely not least, thanks are due to my wife, Leanne, and my dog Samuel Gompers. One of whom helped me immensely with proofreading and moral support, and both of whom are sweethearts.

ABSTRACT

The U.S. government Behavioral Risk Factor Surveillance (BRFSS) survey is an important source of demographic and health data. As with many surveys, BRFSS has missing data resulting from non-response. Because it is impossible to know the true value of missing data, the accuracy of imputation methods for real missing data cannot be known. To solve this problem, I created artificially missing data for two demographic variables for which the originally missing amounts were relatively small: age and race/ethnicity. Proportion estimates for imputation methods at 5%, 10%, and 20% artificially missing were compared against proportion estimates for the same variables from other governmental surveys and against the baseline imputation estimates made at the originally missing amounts, which were between 1% and 3%. I compared and contrasted no imputation, BRFSS imputation methods, multiply imputed hotdeck, and multiply imputed model-based imputation. At each level, missing data were artificially created where the missingness depended on the missing value, where it depended on the value of covariates, and where it did not depend on anything measured by the survey. I found that no imputation was by some measures no worse and even marginally better than any imputation method compared. This thesis has limited scope, however, and caution is recommended before researchers using BRFSS or other survey data forego any attempt at using an imputation method.

1 Introduction

This thesis is an investigation into methods of handling missing information in the 2012 Behavioral Risk Factor Surveillance Survey (BRFSS), a U.S. government administered telephone survey. In particular, it is an investigation into statistical methods of handling missing race/ethnicity and missing age information, and of estimating age and race/ethnicity proportions. Although race and ethnicity constructs that have little or no biological meaning (Park, 1999), that does not mean that they do not exist. It is precisely because of cultural distinctions and disparities that race/ethnicity proportions merit an attempt at accurate estimation. Age meanwhile, cultural while not a purely concept, is nevertheless an important demographic variable for which there are crucial cultural distinctions between groups. Accurate race/ethnicity and age proportion estimation is therefore important, and it should not be taken for granted that proportion estimates from surveys with missing data are accurate.

In this work, I compared four methods for handling missing age values and race/ethnicity values. The methods compared and contrasted included ignoring missing data, using BRFSS imputed values to make proportion estimates, using multiply imputed hotdeck imputation to make proportion estimates, and using multiply imputed model-based imputation to make proportion estimates. I investigated optimal approaches depending on population and sampling method of state-level BRFSS data, percent missing, and missingness mechanism. The computation for this work was done using the statistical software package Stata 12.1.

This research will contribute to the ability of researchers to make informed decisions about how to address missing BRFSS age and race/ethnicity information, and may provide guidance for handling other missing BRFSS information or data from other surveys.

2 Background

Not accounting for missing data in health research is a potentially serious statistical issue. In the worst case scenario, if observations with missing data differ in some fundamental way from observations without missing data, biases will result. In the best case scenario, missing data decreases efficiency and more observations will be required to achieve a given accuracy. In using data in which there is missingness, researchers can either use complete case analysis, which means disregarding observations for which there are missing items, or they can attempt to estimate or account for the missing data in some way.

In order to give some context to the imputation methods studied in this thesis, it will help to understand the 2012 BRFSS survey design. In the BRFSS survey design, the sampling frame is a list of every landline or cell phone number that is a possible household. All phone numbers in the sampling frame come from a Telecordia Technologies database (CDC, 2013 A). The idea is that this sampling frame enables every household with a telephone to have a nonzero chance of selection in the survey. The sampling frame is geographically stratified by a U.S. State or territory or subdivision of a U.S. State or territory (CDC, 2013 A). The decision on whether or not to geographically substratify a state or territory is made at the state or territory level. In addition, the landline numbers are stratified into listed and unlisted phone numbers. The sampling frame is stratified into those numbers that are dedicated as cellular, and those that are not. Then the landline numbers in the sampling frame are cross-referenced with a list of listed numbers, and identified as either listed or unlisted (CDC, 2013 B). The cell numbers are not identified as listed or unlisted.

The states and territories set the target number of completed interviews for each geographic strata within their boundary. The goal of BRFSS is to support at least 4000 interviews per state each year (CDC, 2013 B). For most states, the proportion of listed landline numbers called in a given geographic stratum is 1.5 times the proportion of unlisted landline numbers called in a given geographic stratum. For example, suppose a geographic stratum had a target of 1,750 interviews. Suppose further that there were 50,000 listed landline numbers, and 100,000 unlisted landline numbers in the stratum. Then, in order to meet the requirement that the proportion of listed numbers be 1.5 times the proportion of unlisted, one would sample 750/50,000 = .015 from the listed numbers and 1000/100,000 = .01 from the unlisted numbers. States call as many cell numbers as required to meet their target of completed interviews. If a cell phone is not in the physical location designated by the phone area code, then the subject is added to the sample of the state or territory in which he or she is physically located (CDC, 2013 B). For the cellular phone numbers, a stratified sample is collected in which "An interval, K, is formed by dividing the population count of telephone numbers in the frame, N, by the desired sample size, n. The frame of telephone numbers is divided into n intervals of size K telephone numbers. From each interval, one 10-digit telephone number is drawn at random" (CDC, 2013 A).

Although there are several stratifications of the BRFSS sampling frame, the study design can be thought of as a single stage design with one stratification because, essentially, the stratifications are successively finer divisions of the sampling frame of household phone numbers. The survey design divides the sampling frame of household numbers into geographic strata, strata for listed and unlisted phone numbers, and strata for cell or not cell numbers. In other words, the sampling frame is divided into strata that combine information on geographic area, cell or landline, and listed or unlisted, and every one of the strata is sampled. There is no partial sampling of strata in stages, such as one might see in a multi-stage stratification design. For example, the BRFSS design does not first sample a subset of geographic strata, and then stratify the sampled geographic strata by listed or unlisted numbers and sample from those strata. In the BRFSS design, there is a partitioning of the sampling frame into disjoint strata, all of which are then sampled.

The values for the race/ethnicity categories in the 2012 BRFSS are determined in the following way. Respondents are asked “Are you Hispanic or Latino?” The interviewer then marks either yes, no, don’t know/not sure, or refused. Respondents are then asked “Which one or more of the following would you say is your race: White, Black or African American, Asian, Native Hawaiian or other Pacific Islander, or American Indian or Alaska Native or Other?” If a respondent chooses more than one race, they are then asked “Which one of these groups would you say best represents your race: White, Black or African American, Asian, Native Hawaiian or other Pacific Islander, American Indian or Alaska Native?” On the basis of the responses to these questions, respondents are classified with a single race/ethnicity value. Notice that if someone is Hispanic or Latino they can be of any race and are still classified as Hispanic. The categories are not mutually exclusive, and no options are given for ethnicity except Hispanic or not Hispanic (CDC, 2012 A).

The values for the age group categories in the 2012 BRFSS are determined in the following way. Respondents are asked “What is your age?” On the basis of their response, age groups are made for 18-24, 25-34, 35-44, 45-54, 55-64, and 65 and up.

A distinction is made between unit non-response, in which every variable is missing for a given observation, and item non-response, in which an observation has some missing values and some non-missing (Little & Rubin, 2002). In this thesis, I am mostly concerned with issues involved with item non-response. This is because in BRFSS survey data, unit non-response is dealt with in the weighting process, and researchers who do not have information about unitnonresponse have no way of accounting for it except through the BRFSS weighting scheme (CDC, 2013 B).

The BRFSS weighting process attempts to account for unit non-response with a post-stratification adjustment called iterative proportional fitting (IPF), or "raking." In the IPF algorithm, survey demographic totals are standardized to population marginal totals. The population marginal totals are from U.S. Census population estimates and other population data from Claritas, Current Population Survey data (CPS,) and Public Use Micro-data Samples (PUMS) (Town, 2009). The idea in using the IPF algorithm for unit non-response is that non-missing data are weighted so that the proportions of certain subgroups of observations for which it is assumed that there is little variability match the proportions of those subgroups from known population data (Lemeshow & Levy, 2008). In the BRFSS IPF scheme, there are "8 margins (age group by gender, race/ethnicity, education, marital status, tenure, gender by race/ethnicity, age group by race/ethnicity, [and] phone ownership). If geographic regions are included there are four additional margins (region, region by age group, region by gender, region by race/ethnicity)" (CDC, 2012 B). In addition to unit non-response, the IPF algorithm attempts to account for survey non-coverage (people without telephones), and cell/landline overlap. BRFSS methodology uses the IPF iterative algorithm to standardize proportions from BRFSS variables until they reach some desired level of convergence to population estimates from the Census and other population data. In the BRFSS survey design, the weights from the IPF scheme are combined with the survey design weights to form the final weights for researchers to use when making estimation and inference from the data. The survey design weight takes into account the probability of a household being selected. The idea with survey design weights is that, for example, if a household has two phones then it has twice the probability of being selected as a household with only one phone. It should therefore be weighted to count only half as much as a household with one phone.

The BRFSS IPF scheme does not attempt to account for item non-response. Any accounting for item non-response must be dealt with by users of BRFSS data.

When dealing with either unit non-response or item non-response, some care is required depending on the mechanism of the missingness. Donald Rubin and Roderick Little, in their *Statistical Analysis with Missing Data* text, describe the three fundamental ways in which data are missing (Little & Rubin 2002). Data can be:

1) Missing completely at random (MCAR),
2) Missing at random (MAR), or
3) Not missing at random (NMAR).

These concepts have formal mathematical definitions, but can be understood intuitively with examples. Suppose the government gave a health survey to a group of individuals. NMAR data would occur if whether or not a value was missing depended on what the value was. For example, if only bald people refused to answer a question about the extent of their hair loss, those data would be NMAR. On the other hand, MAR data would occur if the question of whether or not a value was missing depended on some measured quantity other than the missing value. For example, if men (either bald or not) were more likely than women to refuse to answer the question about the extent of their hair loss, then the missingness is MAR. Data missing completely at random, or MCAR, would occur if the question of whether or not the data was missing did not depend on either the missing value or any other measured quantity. Imagine, for example, that the ink on the answer to the question about the extent of hair loss was smudged to the point of unreadability on several surveys for some reason that was independent of the survey respondents.

Fritz Scheuren, writing in *The American Statistician*, maintains that in his experience all three mechanisms for missing data are usually present in one data set. According to Scheuren, "something like 10% to 20%" of missingness is MCAR, while MAR "is about half of the problem" (Scheuren, 2005). Trying to understand the mechanism for missingness is important because the method of imputation will be based on some assumption about the distribution of the missingness. The simplest imputation methods assume the missing data are MCAR, while more sophisticated imputation methods may assume the data are MAR. For data that are NMAR, by definition we have no information on the distribution of the missingness, which makes imputation problematic.

As previously stated, the mechanism of missingness dictates what statistical techniques should be used. For item non-response that the researcher is willing to assume is MAR, there are methods that use the non-missing data to estimate the missing data. The key idea here is that for data that are MAR, the things that are known about a subject give information about the things that are not known. For example, suppose one knew a subject was male and had left the question about hair loss blank. Then one would have some idea of the probability of the subject's hair loss. For item non-response that the researcher is willing to assume is MCAR, the things that are known about other subjects can be used to estimate the missing values for a given subject. The idea is that if the missingness does not depend on anything measured or unmeasured, then non-missing information from other subjects gives a decent idea of how to estimate the missing data for subjects with missing information. For example, if one knows that the ink was smudged to the point of unreadability for certain items in a way that did not depend on anything measured or unmeasured, one could still estimate the missing items for given individuals based on the non-missing values from other subjects. Item non-response that the researcher believes is NMAR is the most difficult case, and there may not always be any good methods. If the researcher has any additional information to work with, however, there may be steps that can be taken. Rubin and Little describe a method in which the item non-response is NMAR, but there is a non-missing covariate that partitions the range of values for the quantity attempting to be measured. In their example, survey respondents refused to give an exact annual income amount but were willing to identify an income range. In that case, they showed that a “maximum likelihood estimate can be derived” (Rubin & Little, 2002). They also offer the example of censoring, in which the mechanism for missing time-to-event data is NMAR but is known to depend on the time until “termination of data collection” (Rubin & Little, 2002). The IPF method to deal with unit non-response implicitly assumes the data is MAR. For unit non-response that is NMAR, some techniques may still be some available. For example, Rubin and Little describe a pattern-set mixture model and an iterative maximum likelihood method (Rubin & Little, 2002).

In many cases, researchers will have good reason to suspect one missing data mechanism over another. With income data, for example, there is a psychological reason warranting a suspicion that people with certain incomes may not wish to divulge the information. When researchers do not have a good idea of the missingness mechanism there are informative statistical tests. One method is to compare the means of every other measured variable between missing and nonmissing groups. If the difference in means is significant for any of the covariates that would be reason to believe that the missingness was not MCAR. But as Little points out, if there are many covariates this will result in “multiple comparison problems” (Little, 1988). Depending on the number of covariates, a Bonferroni correction may result in a p-value threshold that is too conservative to detect a real difference. Little introduced a single test statistic to test for MCAR (Little, 1988).

Previous researchers investigating missing data in BRFSS survey data have advocated “a systematic search for items that may be correlated with the key response measure” (Frankel et al., 2012). The idea being that if one can find variables that help predict missingness, then they should be included in an imputation model in order to improve imputation based estimates.

Imputing a single value for each missing item can result in an over-fitted model. Over-fitted in this context means that there won’t be as much variability or statistical noise as there would be if one had actual measurements. As Frank Sulloway notes in his book, *Born to Rebel*, the data in some sense will fit “too good” (Sulloway, 1996). Donald Rubin gets around the problem of single imputation by using multiple imputation. In the multiple imputation approach, several imputed data-sets are created and then estimates from each data-set are averaged or otherwise combined to introduce some statistical variability.

Missing data imputation can accomplish several objectives. Researchers can avoid dropping information they do have for a subject simply because they are missing some information for that subject (Little & Rubin, 2002). Researchers can estimate the decrease in precision (wider confidence intervals) from estimates that they were able to make already based on non-missing data (Sulloway, 1996). And researchers can potentially reduce any bias that would result from analyzing only the non-missing data (Horton & Kleinman, 2007).

3. Methods

This work analyzed the effect of the missingness mechanism, percentage of missing data, and state level survey design on race/ethnicity and age proportion estimates for different imputation methods in missing BRFSS data.

Imputation-based estimates for race/ethnicity and age group proportions in BRFSS data were compared to estimates of those variables from other government surveys from which the BRFSS IPF weighting scheme attempted to match proportions to, including 2010 Census data and 2012 American Community Survey (ACS) data. For age group proportion estimation, proportions from the 2012 ACS 3-year survey estimates were taken as the true proportions for the purposes of comparison. For race/ethnicity proportion estimation, the average of estimates from the 2010 Census and the 2012 ACS data were taken as the true proportions of the population for the purposes of comparison. The true proportions were then compared to proportion estimates made using a multiply imputed hotdeck method, a multiply imputed model-based method, non-imputed data (ignoring subjects with missing items), and BRFSS imputation methods. The American Indian and Alaska Native category was combined with the Native Hawaiian or other Pacific Islander category. For the age group imputation estimates, the exact BRFSS imputation methodology was used where possible, but in some cases a BRFSS-like method was used because the exact method was not known. Estimates were calculated with the final survey weights and survey design taken into account.

The BRFSS race imputation method is a single imputation method that imputes the most common race in the geographic substrata for which the subject with a missing race value is located. The BRFSS age imputation method is a single imputation method that imputes a mean age from the sample (CDC, 2013 C). One caveat is that the age group proportion estimates labeled “BRFSS” did not technically use the exact BRFSS age imputation method except at the originally missing level because I do not know the exact method used to obtain a mean age value for imputation. The BRFSS documentation on age imputation states only that “the value of the imputed age will be an average age computed from the sample if the respondent refused to give an age” (CDC, 2013 C). There is no further specification of what subset of age values the mean comes from. The method I used for the artificially missing levels computed a mean from the same geographic strata used in the race imputation method. Nevertheless, because of the grouping of imputed age values into age groups, the method I used is a similar type of method to the exact, though unknown, BRFSS age imputation scheme.

Hotdeck imputation refers to a group of imputation methods that impute missing values using non-missing values from subjects in the same dataset, as opposed to cold deck imputation which uses values from another dataset (Levy & Lemeshow, 2008). The hotdeck multiple imputation method I used is a Stata user developed add-on written by Adrian Mander and David Clayton of Cambridge (Clayton & Mander, 1999). Their method applied to this work imputed a race or age group value of a subject with a non-missing race or age group value and some specified set of covariates the values of which match those of the subject with the missing age or race value. This was done five times (multiple imputation) and then proportion estimates were based on an average of the five data sets. For example, I imputed missing race values with a non-missing race value from another subject in the dataset with the same values for gender, 5-year age group, race, and income group as those of the subject with the missing race or age value. After creating five data sets in this way, an average race or age proportion was obtained from the five data sets. Confidence intervals in the Mander and Clayton hotdeck procedure are calculated based on both between imputation variance and within imputation variance (Mander & Clayton, 2007), using the method of Rubin and Little (Little & Rubin 2002).

Model-based imputation refers in general to imputation procedures that “analyze data having missing values by modeling the likelihood function of the incomplete data and using maximum likelihood procedures” (Levy & Lemeshow, 2008). The model-based multiple imputation method I used imputed missing race or age values based on an Expectation Maximization (EM) iteration of a multinomial logistic model of race or age and some specified set of covariates. Similar to the multiple hotdeck imputation, a multiple model-based imputation was produced. Five datasets were created and then proportion estimates were obtained by averaging the estimates from each of the datasets. For example, a multinomial logistic model was produced with race as the dependent variable, and income group, age, and gender as independent variables. Race or age values were imputed based on this model for five different datasets and then the proportion estimates were averaged. The model-based procedure I used was a Stata 12 command that did not use the method of Rubin and Little to estimate variance by combining between imputation variance and within imputation variance (Appendix 7.1).

For each of the above imputation methods, proportion estimates were obtained based on imputation of originally missing data amounts, as well as artificially created missing data amounts. Imputation for artificially created missing data amounts was done for 5%, 10%, and 20% missing. For each percent of missing, the artificially created missing values simulated data that were MCAR, MAR where the missingness depends on a covariate used in the hotdeck and model-based imputation models, MAR where the missingness depends on a covariate not used in the hotdeck and model-based imputation methods, and NMAR. The idea was to attempt to determine the level and mechanism at which the imputation proportion estimates stopped approximating estimates from the 2012 ACS survey and 2010 Census estimates. A secondary goal was to examine the effect of using multiple imputation with and without accounting for between imputation variance.

MCAR, MAR, and NMAR data were artificially simulated using a uniform random number generator. A value between 0 and 1 was assigned to every non-missing race and age value. For MCAR data, all race and age values with an assigned random number less than 0.05, 0.1, and 0.2, depending on the level being simulated, were artificially designated as missing. Similarly, for MAR data, race and age values were designated as missing using a random number generator to artificially create missingness that depended on gender and marital status, separately. And, finally, NMAR data were simulated at each level, separately for age and race, using a random number generator to artificially create missing values where the missingness depended on white status, and on whether or not a subject was in the 65 and up age group.

Imputation estimates were made for nationwide data (excluding Guam and Puerto Rico) and for data from 4 individual states: NY, NJ, OR, and WA. The rationale for comparing proportion estimates from these states was to attempt to account for population size and number of sub-geographic sampling strata per state while holding demographics constant. Washington sub-stratifies into 40 sub-geographic strata, while Oregon does not substratify geographically. Similarly, New Jersey substratifies into 23 sub-geographic strata while New York substratifies into two. The implicit assumption made here was that Oregon and Washington, and New York and New Jersey have similar age and race demographics, and that by comparing the imputation estimates between the two pairs of states I could analyze the effect of number of strata and population size on imputation.

4. Race Imputation Results

4.1 Summary of all cases considered

At each level of missing, there were 120 estimates compared to the true proportion. These consisted of six race categories, five subsets of the BRFSS survey, and four methods for handling missing data. Three metrics were used to measure the accuracy of race proportion estimates:

1) Percent of estimates whose 95% confidence intervals contained the true value;
2) Total absolute difference between each estimate and the true value; and
3) Distance between each estimate and the originally missing estimate for that method.

Measured by the average total distance from the true value, the accuracy of complete case method, multiple hot deck imputation, and multiple model-based imputation methods appear to be stable until the 20% artificially missing level. At the 20% MCAR and MAR levels, the model-based method is slightly less accurate than the hotdeck and complete case methods. At the 20% NMAR level, the hotdeck, complete case, and model-based methods all drop substantially in accuracy. The BRFSS imputation, meanwhile, is markedly less accurate than the other three methods at every level except for NMAR data. The accuracy of the BRFSS NMAR estimates is due to the mechanism I used to create artificial NMAR data, which was tailored for the BRFSS imputation. The BRFSS method would not be as accurate for other types of NMAR data. Since the BRFSS imputation method imputes the most common race in the geographic strata for which the respondent with the missing value is located, all geographic strata for which white is the most common race are imputed with the correct race under the BRFSS imputation method. For Oregon, this means every artificial NMAR data point will be imputed correctly under the BRFSS imputation scheme. For New York, this means almost every artificial NMAR data point will be imputed correctly under the BRFSS imputation scheme. It is clear, therefore, that the performance of the BRFSS scheme would be substantially worse had I created artificially missing NMAR data where the missingness was not made up of only originally non-missing white respondents (Figure 1).

As a result of the arbitrary choice of the true proportion and the inherent difficulty in estimating race proportions, many imputation estimates differ from the true proportion at every level of missingness mechanism and level of missingness, even the relatively small originally missing amounts. In addition, due to the fact that a confidence interval either contains the true value, or it does not, some confidence intervals that miss the true value just barely are considered inaccurate by this metric, and it causes the accuracy measurement to be more erratic than the mean absolute error. Also, because of the probable underestimates of the variance for the model-based multiple imputation (Appendix 7.2), the percent of model-based method confidence intervals that contain the true value was reduced. Even at the 1% originally missing level the percent of confidence intervals that contain the true value do not come close to the nominal value of 95% for any method.

Because of the issues with 95% confidence intervals as an accuracy metric, I used the accuracy of the race proportion estimates at the original level of missing as a kind of baseline accuracy level to compare to future levels. After all, the real values of the artificially missing in the BRFSS survey (if not in the population) are known, and if the imputation methods are not at least as accurate as at the originally missing level, it means that the methods are not imputing the artificially missing in the correct place. Because of the relatively small originally missing proprotions, one can think of the proportion estimates at the originally missing level as a standard by which to judge artificially missing imputation estimates.

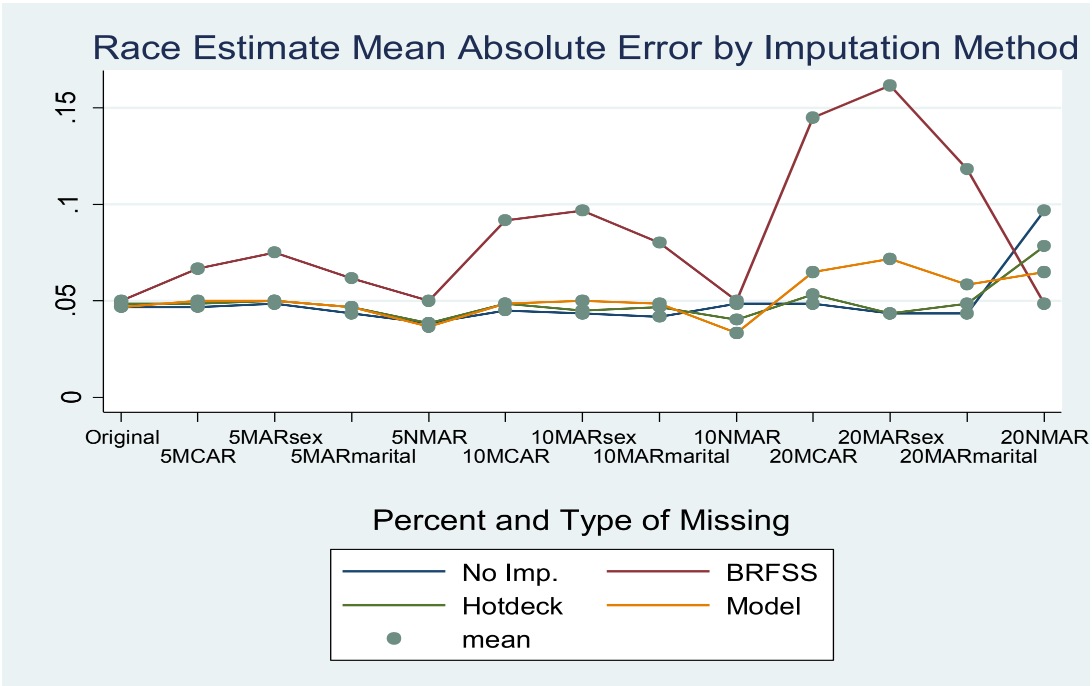

Figure 1: Mean absolute error of all race imputation estimates by missing level. Each point represents the average difference between the imputation estimates and the true value for a given missing level and mechanism. The x-axis labels are, from left to right, originally missing, 5% MCAR, 5% MAR where missingness depends on sex, … , 20% NMAR.

When separated by imputation method, the accuracy of the imputation methods as measured by the proportion of 95% confidence intervals that contain the true proportion give the same general results as when accuracy is measured by the mean absolute error. The hotdeck, model-based, and complete case methods perform similarly until the 20% missing level. At the 20% MCAR and MAR levels, the model-based method is noticeably less accurate than the hotdeck and complete case methods. At the 20% NMAR level, all three of the hotdeck, model-based, and complete case methods drop off in accuracy. The BRFSS method, meanwhile, is substantially less accurate than the other methods except for NMAR data (Figure 2).

Also of consideration is that both the hotdeck and model-based imputation methods imputed using listwise deletion of the selected imputation covariates. For the hotdeck method, any survey respondent with a non-missing race value but a missing value for one or more of gender, income group, or age was not eligible to have his or her race value imputed into any subject with a missing race value. For the model-based method, any survey respondent with a non-missing race value but a missing value for one or more of gender, income group, or age was not eligible to be used in the model on which the imputation estimates are based. The percent of survey respondents with missing values for one or more of gender, income group, or age is displayed in Table 1. The majority of these missing are due to income group, which had 66,745 missing values.

Table 1: Percent of imputation model covariates missing by BRFSS subset.

Survey Respondents with Missing Age, Sex, or Income Group	NJ	NY	OR	WA	U.S. BRFSS
Weighted Percent Missing	18.56	16.7	13.48	13.98	13.68
Percent Missing	17.48	14.85	14.86	13.72	14.16
Total Sampled	15,761	6,060	5,302	15,319	467,333

Confidence intervals for model-based imputation proportion estimates were computed using a first order Taylor Series approximation to estimate the variance of the proportion estimate, and then a t-distribution to estimate the confidence limits. The hotdeck imputation confidence intervals were computed in a similar way, except that they accounted for between imputation variance using the method of Rubin and Little (Rubin & Little, 2002), while the model-based confidence intervals did not. For a complete description of how confidence intervals were computed, see Appendix 7.1.

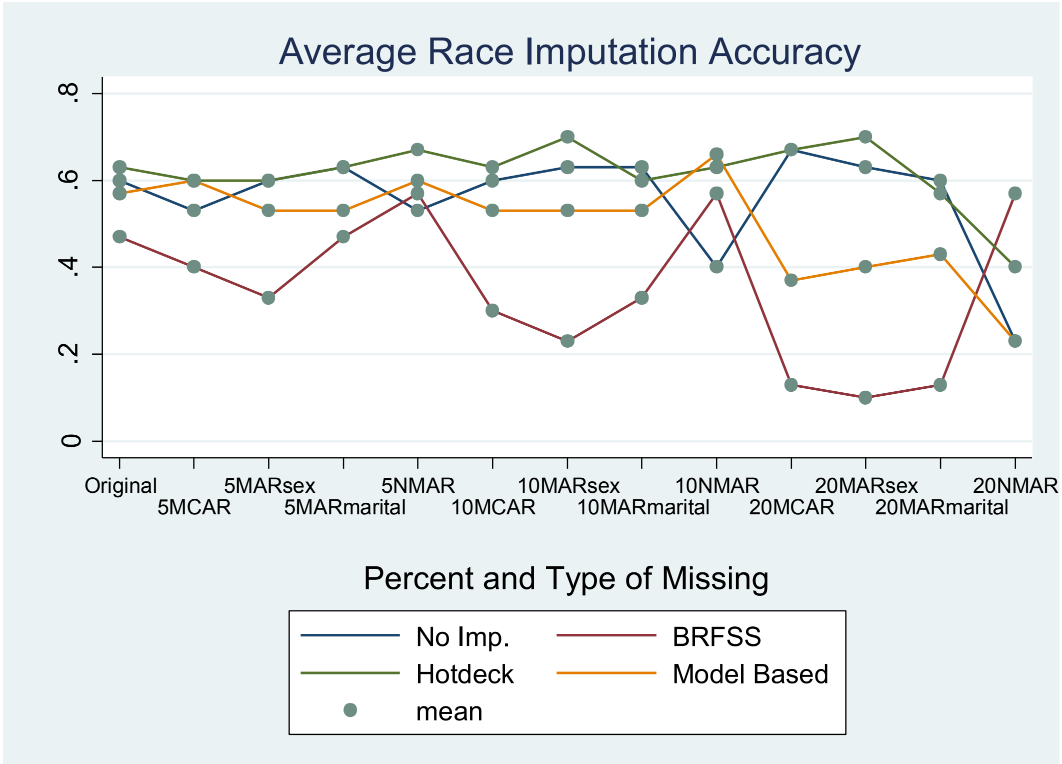

Figure 2: Each point represents the average proportion of race estimate confidence intervals that contained the true proportion for the given level of missing and mechanism of missingness. The x-axis labels are, from left to right, originally missing, 5% MCAR, 5% MAR where missingness depends on sex, … , 20% NMAR.

4.2 Originally missing imputed race proportion estimates

The percent of race values that were originally missing were all between 1% and 3% for OR, WA, NY, NJ, and the entire BRFSS survey. However, because the final BRFSS weights have a substantial effect on estimates, it is also informative to consider the weighted percent missing. For the originally missing, these did not differ substantially (Table 2).

Table 2: Weighted and unweighted missing percent by BRFSS subset.

Survey Respondents with Missing Race/Ethnicity	NJ	NY	OR	WA	U.S. BRFSS
Weighted Percent Missing	1.51	2.59	1.29	1.27	1.17
Percent Missing	1.81	2.49	1.51	1.40	1.33
Total Sampled	15,761	6,060	5,302	15,319	467,333

At this level of missing, the proportion estimates for all imputation methods and for no imputation do not differ markedly from the average of the 2012 ACS proportion estimates and the 2010 U.S. Census proportion estimates (hereafter referred to as the true proportion). One exception was the "Other" and "Native" categories, which are harder to estimate than are "White" (Figure 3).

Table 3: Performance of originally missing race imputation estimates by imputation method.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model-based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	60%	46%	63%	56%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.275	0.295	0.289	0.277	

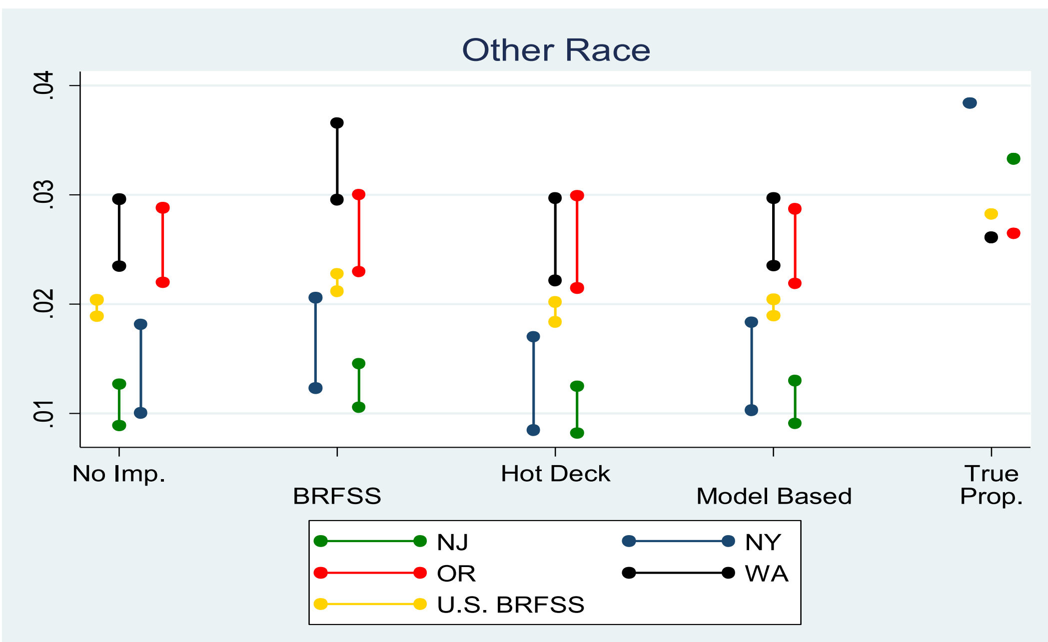

Figure 3: Originally missing imputed "Other" race proportion estimates by imputation method and subset of BRFSS. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the "true" proportion.

4.3 Originally missing plus 5% artificially created MCAR imputed race proportion estimates

In addition to the originally missing, 5% artificial MCAR race values resulted in approximately 6.2% missing and 6% weighted missing for the entire U.S. portion of the BRFSS survey. At this level, the weighted percent missing did not differ substantially from the percent missing for any of the five subsets of the BRFSS survey analyzed (Appendix 7.2). With the exception of the BRFSS method, all the methods were close to their baseline accuracy (Table 4).

Table 4: Performance of 5% MCAR race imputation estimates by imputation method.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model-based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	53%	40%	60%	60%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.281	0.404	0.294	0.296	
Difference between absolute difference at current level and baseline absolute difference	0.006	0.109	0.005	0.019	

Again, the estimates for the “Other” race category were the least accurate. In particular, for higher population BRFSS subsets, the imputation estimates were underestimates; meanwhile the U.S. “Native” population estimates were all slight underestimates (Figure 4).

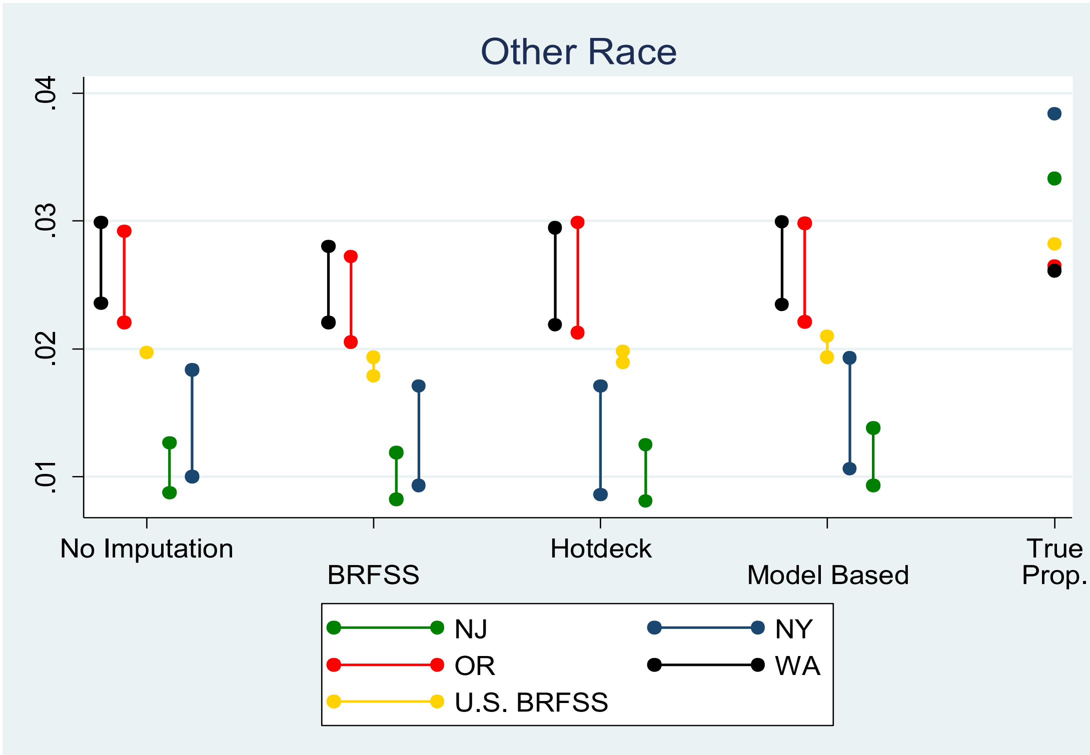

Figure 4: “Other” category originally missing plus 5% MCAR imputed race proportion estimates by imputation method and subset of BRFSS. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the “true” proportion.

4.4 Originally missing plus 5% artificially created MAR where missingness depends on a variable used as a covariate in the hotdeck and model-based imputed race proportion estimates

In addition to the originally missing, the 5% artificial MAR race values resulted in approximately 6.3% missing and 7.2% weighted missing for the entire survey. At this level, the weighted missing percents were similar to the unweighted (Appendix 7.2).

As can be seen in Table 5, at this level the BRFSS estimates do worse than the other methods, and the other methods perform similarly to the baseline performance at the originally missing level. Interestingly, the “Other” estimates perform worse than they did at the 5% MCAR level, even for the smaller populations of Oregon and Washington. Additionally, in contrast to the 5% MCAR level, the Oregon Native imputation estimates were substantial underestimates for every imputation method (Figure 5).

Table 5: Performance of 5% MAR race where missingness depends on gender imputation estimates by imputation method.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	60%	33%	60%	53%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.288	0.445	0.303	0.298	
Difference between sum of absolute differences at current level and baseline absolute differences	0.013	0.150	0.014	0.021	

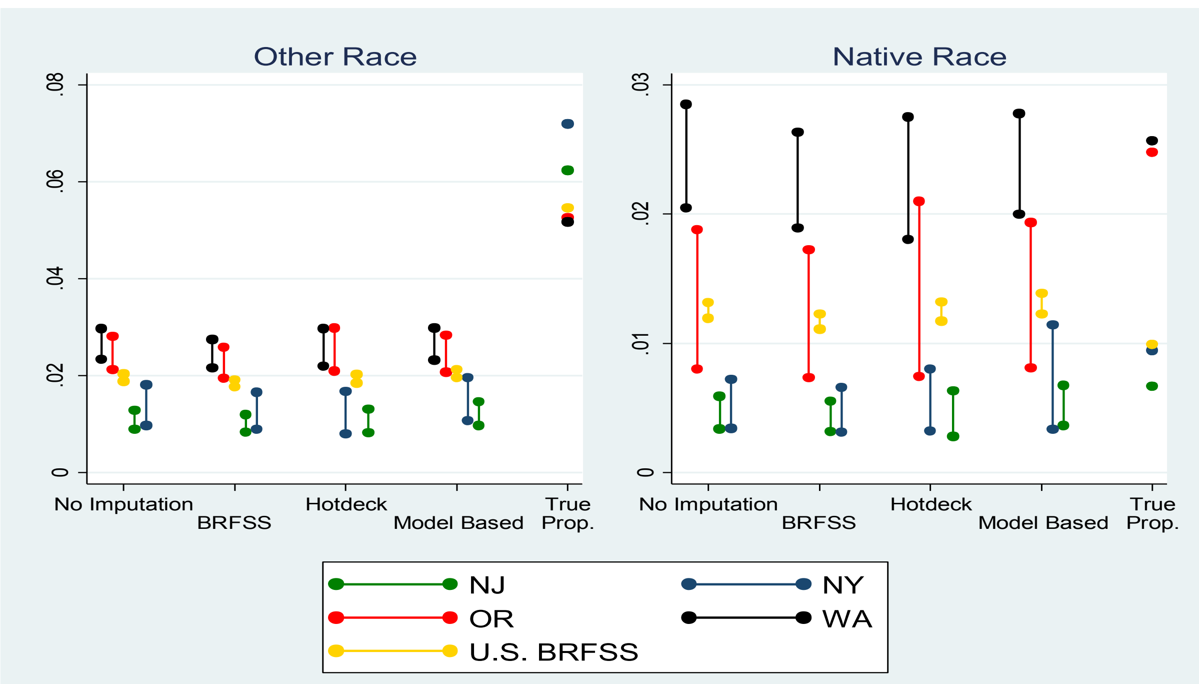

Figure 5: Originally missing plus 5% MAR imputed “Other“ and “Native” race proportion estimates by imputation method and subset of BRFSS where missingness depends on gender. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the “true” proportion.

4.5 Originally missing plus 5% artificially created MAR where missingness depends on a variable not used as a covariate in the hotdeck and model-based imputed race proportion estimates

In addition to the originally missing, the 5% MAR where the missingness depends on marital status resulted in approximately 6.2% missing and 5.8% weighted missing for the entire survey. The weighted and unweighted missing percentages did not differ substantially (Appendix 7.2).

Table 6: Performance of 5% MAR race where missingness depends on marital status imputation estimates by imputation method.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	63%	47%	63%	53%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.259	0.373	0.284	0.284	
Difference between sum of absolute differences at current level and baseline absolute differences	-0.016	0.078	-0.005	0.007	

As Table 6 indicates, at this level and mechanism of missing, once again the BRFSS method performed the worst. The model-based method performed at a similar accuracy to the baseline originally missing level, while the hotdeck and no imputation methods were actually more accurate than the baseline level. Similar to the 5% MAR level where missingness depends on gender, the “Other” estimates were all underestimates. Compare the “Other” estimates, for example, to the “Native” estimates, which were mostly accurate, except for OR, where every estimate was an underestimate (Figure 6).

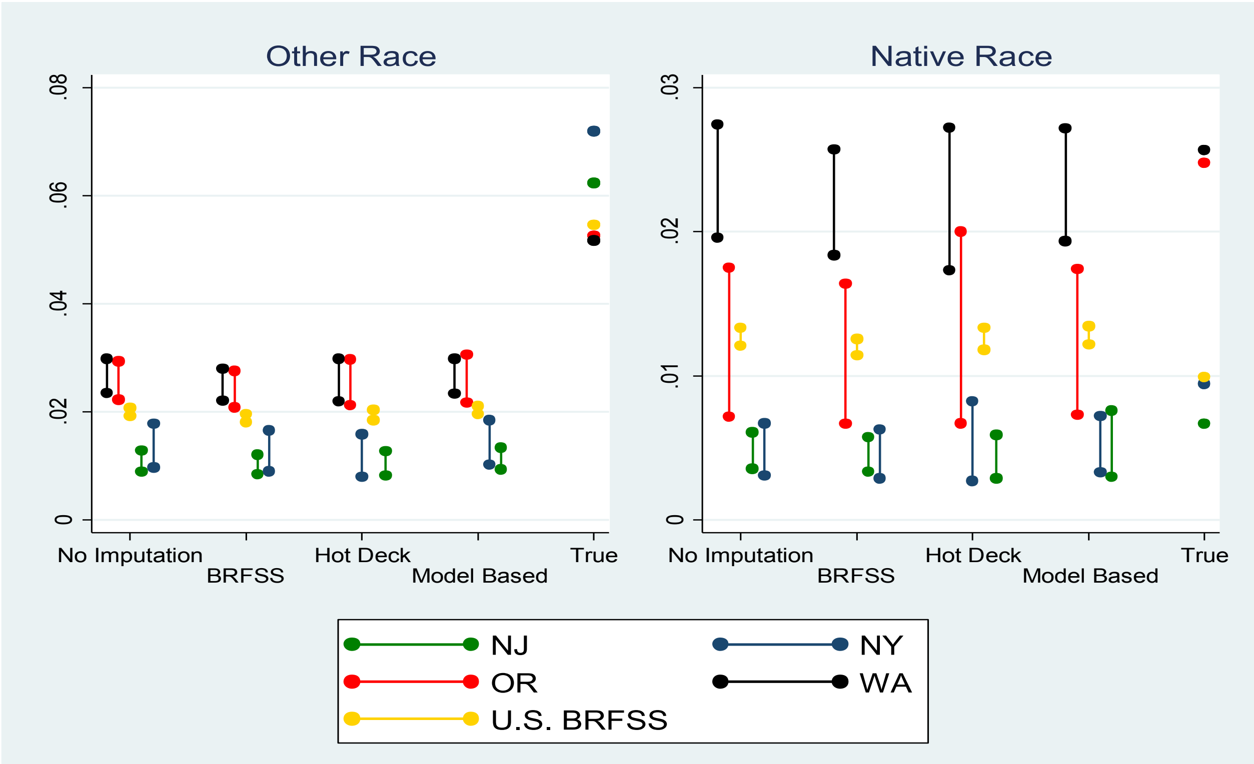

Figure 6: Originally missing plus 5% MAR imputed "Other" and "Native" race proportion estimates by imputation method and subset of BRFSS where missingness depends on marital status. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the "true" proportion.

4.6 Originally missing plus 5% artificially created NMAR imputed race proportion estimates

In addition to the originally missing, the artificially created 5% NMAR race values resulted in approximately 6.3% missing and 5.8% weighted missing for the entire survey. As with the previous levels of missingness, the 5% NMAR weighted and unweighted missing percentages were similar (Appendix 7.2).

At the 5% NMAR level, the no imputation, hotdeck, and model-based imputation methods performed better than at the baseline missing level. The BRFSS method was comparable to baseline accuracy due to the NMAR mechanism used, not the imputation method itself. As with the 5% MAR level where missingness depended on marital status, the "Other" race estimates were all substantial underestimates, the Oregon “Native” estimates were all underestimates (Figure 7), and the other race category estimates were all in the neighborhood of the true proportion. The “White” race category estimates are typical of the accuracy for non “Other” and “Native” categories, even though for the 5% NMAR level the missingness was all created from originally “White” respondents (Figure 8).

Table 7: Performance of 5% artificially created NMAR race imputation estimates by imputation method where missingness depends on white race.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	53%	56%	66%	60%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.234	0.299	0.233	0.224	
Difference between sum of absolute differences at current level and baseline absolute differences	-0.041	0.004	-0.056	-0.053	

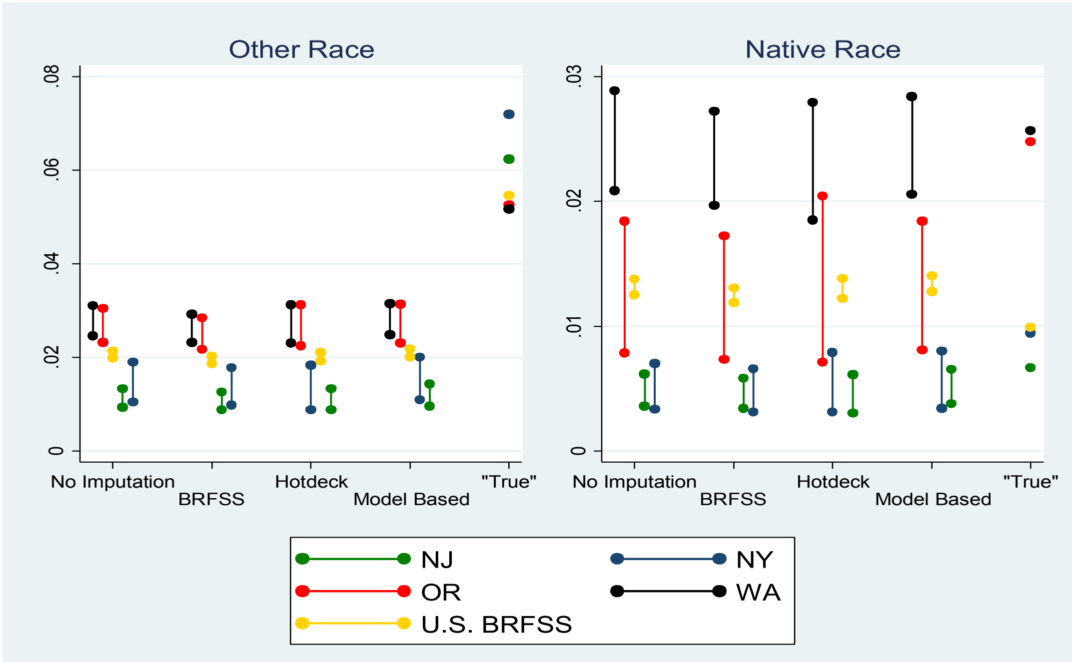

Figure 7: Originally missing plus 5% NMAR imputed "Native" and "Other" race proportion estimates by imputation method and subset of BRFSS where missingness depends on white status. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the "true" proportion.

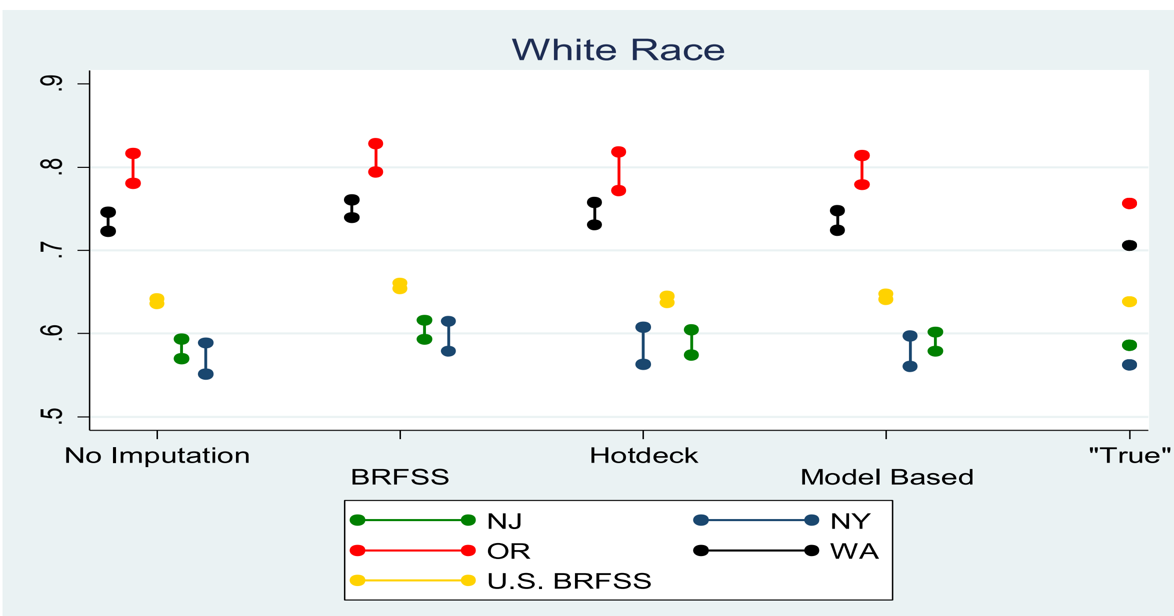

Figure 8: Originally missing plus 5% NMAR imputed "White" race proportion estimates by imputation method and subset of BRFSS where missingness depends on white status. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the "true" proportion.

4.7 Originally missing plus 10% artificially created MCAR imputed race proportion estimates

In addition to the originally missing, the 10% artificially created MCAR race values resulted in approximately 11.2% missing and 11% weighted missing for the entire survey. The weighted and unweighted missing percentages did not differ substantially for any subset of the survey (Appendix 7.2).

Table 8: Performance of 10% artificially created MCAR race values by imputation method.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	60%	30%	63%	53%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.267	0.547	0.285	0.291	
Difference between sum of absolute differences at current level and baseline absolute differences	-0.008	0.252	-0.004	0.014	

The 10% MCAR level had the same issues with accuracy for the "Other" and "Native" categories as did the 5% missingness and mechanism levels. In addition, as can be seen from Table 8, the BRFSS estimates performed even worse compared to the other methods, which more or less maintained their baseline accuracy. Once again the hotdeck and no imputation methods actually were slightly better than baseline, while the model-based method was slightly worse.

4.8 Originally missing plus 10% artificially created MAR where missingness depends on a variable used as a covariate in the hotdeck and model-based imputed race proportion estimates

In addition to the originally missing, the 10% artificially created MAR race values where missingness depended on gender resulted in approximately 11.2% missing and 12.8% weighted missing for the entire survey. The weighted and unweighted missing percents were not substantially different for any of the BRFSS subsets analyzed (Appendix 7.2).

Table 9: Performance of 10% artificially created MAR race values by imputation method where missingness depends on gender.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	63%	23%	70%	53%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.259	0.580	0.270	0.302	
Difference between sum of absolute differences at current level and baseline absolute differences	-0.016	0.285	-0.019	0.025	

As Table 9 shows, at this level of missing, the BRFSS imputation method performs noticeably worse than the other imputation methods. This is not a result of any single estimate being extremely off, however, but of many estimates being slightly more off than the other methods and the inaccuracy adding up over all 30 BRFSS estimates. For a typical example, see the “White” proportion estimates (Figure 9), where the BRFSS estimates are in the neighborhood of the true value, but are noticeably farther away than the other methods.

At this level, all of the imputation methods had the same inaccuracy issues with the estimates for the “Other” category and the estimates for the Oregon “Native” category that we saw for the other missingness levels and mechanisms. Once again the no imputation and hotdeck methods were slightly better in accuracy than at baseline, while the model-based method was slightly worse in accuracy than at baseline.

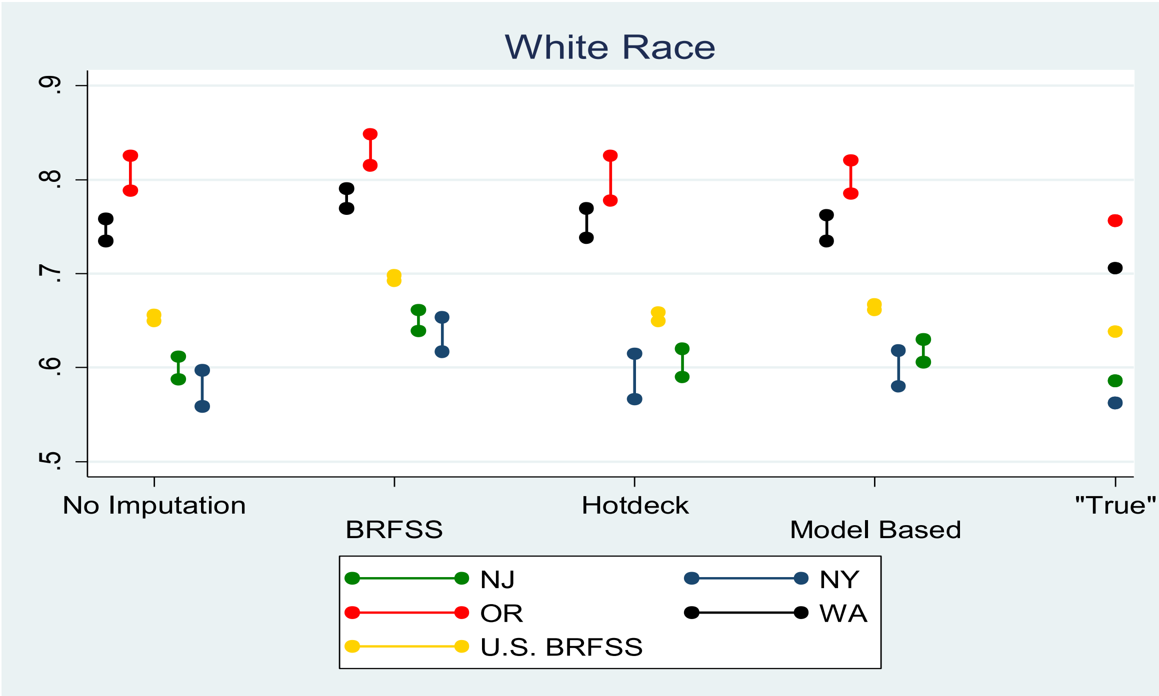

Figure 9: Originally missing plus 10% MAR imputed "White" race proportion estimates where missingness of race values depends on gender by imputation method and subset of BRFSS. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the "true" proportion.

4.9 Originally missing plus 10% artificially created MAR where missingness depends on a variable not used as a covariate in the hotdeck and model-based imputed race proportion estimates

In addition to the originally missing, the 10% artificially created MAR race values where the missingness depended on gender resulted in approximately 11.6% missing and 10.6% weighted missing for the entire survey. At this level, the weighted and unweighted missingness percentages did not differ substantially (Appendix 7.2).

At this level of missing, the BRFSS was the only imputation method that performed substantially worse than the baseline accuracy (Table 10). Once again, this is not a result of any single estimate being extremely off, but rather of many estimates being slightly more off than the other methods and the inaccuracy adding up over all 30 BRFSS estimates. Again, there were substantial accuracy issues with all imputation methods for the "Other" race category, and for the Oregon "Native" estimates. Also once again, the hotdeck and no imputation methods were better than baseline in accuracy, while the model-based method was slightly worse.

Table 10: Performance of 10% artificially created MAR race values by imputation method where missingness depends on marital status.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	63%	33%	60%	53%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.250	0.478	0.276	0.292	
Difference between sum of absolute differences at current level and baseline absolute differences	-0.025	0.183	-0.013	0.015	

4.10 Originally missing plus 10% artificially created NMAR imputed race proportion estimates

In addition to the originally missing, the 10% artificially created NMAR race values resulted in approximately 6.3% missing and 5.8% weighted missing for the entire survey. At this level and mechanism of missingness, the weighted and unweighted missing proportions were approximately the same (Appendix 7.2).

At this level of missing, the BRFSS and no imputation methods performed worse than the baseline accuracy (Table 10). The BRFSS method performed worse than baseline in spite of the fact that it is tailor made for this type of NMAR data. This is not a result of any single estimate being extremely off, but rather of many estimates being slightly more off than the other methods and the inaccuracy adding up over all 30 BRFSS estimates. As usual, there were also substantial accuracy issues with all imputation methods for the "Other" race category, and for the Oregon "Native" estimates. At this level, interestingly, the hotdeck and model-based methods were better than baseline in accuracy, while the no imputation method was slightly worse.

Table 11: Performance of 10% artificially created NMAR race values by imputation method where missingness depends on white status.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	40%	56%	63%	66%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.291	0.298	0.238	0.199	
Difference between sum of absolute differences at current level and baseline absolute differences	0.016	0.003	-0.051	-0.078	

4.11 Originally missing plus 20% artificially created MCAR imputed race proportion estimates

In addition to the originally missing, the 20% artificially created MCAR race values resulted in approximately 21% missing and 20.8% weighted missing for the entire survey. At this level, the weighted and unweighted missing percents did not differ substantially (Appendix 7.2).

As can be seen from Table 12, at this level of missing the BRFSS imputation estimates perform much worse than any other imputation method. Once again, all methods were off for the "Other" category, and for the Oregon "Native" category. Additionally, the model-based multiple imputation, hotdeck multiple imputation, and no imputation method all performed marginally worse than at their baseline accuracy. For the model-based method, the absolute differences are not enormously different, and the reduced percentage of estimates whose 95% confidence intervals did not contain the true proportion are because of several estimates being slightly outside the limits, rather than because of any being a great deal outside the limits.

Table 12: Performance of 20% artificially created MCAR race values by imputation method.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	66%	13%	66%	36%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.294	0.866	0.316	0.387	
Difference between sum of absolute differences at current level and baseline absolute differences	0.019	0.571	0.027	0.110	

4.12 Originally missing plus 20% artificially created MAR imputed race proportion estimates where missingness depends on a variable used as a covariate in the hotdeck and model-based imputed age proportion estimates

In addition to the originally missing, the artificially created MAR race values resulted in approximately 21% missing and 25% weighted missing for the entire survey. At this level of missing, we see an approximately 3%-4% difference between the weighted and unweighted missing percents for all subsets analyzed (Appendix 7.2).

At this level, the hotdeck and no imputation methods performed slightly better than at baseline, while the model-based method performed slightly worse than baseline. The BRFSS method, meanwhile, had the worst level of accuracy. Figure 10 shows the imputation proportion estimates for the hispanic race/ethnicity category, which shows a typical scenario at this level of missing.

Table 13: Performance of 20% artificially created MAR race values estimates by imputation method where missingness depends on gender.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	63%	10%	70%	40%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.260	.972	0.263	0.428	
Difference between sum of absolute differences at current level and baseline absolute differences	-0.015	0.677	-0.026	0.151	

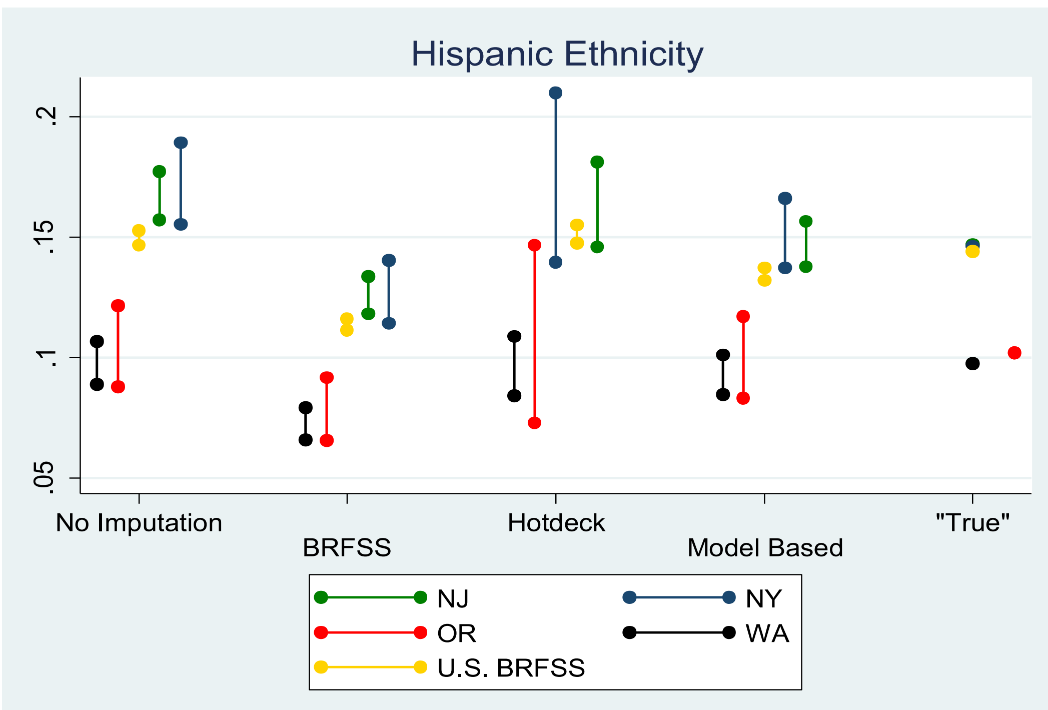

Figure 10: Originally missing plus 20% MAR imputed "Hispanic" race/ethnicity proportion estimates by imputation method and subset of BRFSS where missingness depends on gender. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the "true" proportion.

4.13 Originally missing plus 20% artificially created MAR imputed race proportion estimates where missingness depends on a variable not used as a covariate in the hotdeck and model-based imputed age proportion estimates

In addition to the originally missing, the artificially created MAR race values where missingness depends on marital status resulted in approximately 20.9% missing and 19.9% weighted missing for the entire survey. These proportions were approximately the same for every subset of the survey, except for New York, which had 29.8% missing and 20.1% weighted missing (Table 14).

Table 14: Weighted and unweighted age value missing percent by BRFSS subset where missingness is MAR and depends on marital status.

Survey Non-Respondents plus 20% artificial MAR Missing Race Values where missing depends on marital status	NJ	NY	OR	WA	U.S. BRFSS
Weighted Percent Missing	20.4%	20.1%	20.2%	20.3%	19.9%
Percent Missing	21.1%	29.8%	20.8%	22%	20.9%
Total Sampled	15761	6060	5302	15319	467,333

Table 15: Performance of 20% artificially created MAR race values estimates by imputation method where missingness depends on marital status.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	60%	13%	56%	43%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.256	.714	0.285	0.347	
Difference between sum of absolute differences at current level and baseline absolute differences	-0.019	0.419	-0.004	0.070	

As can be seen in Table 15, the BRFSS imputation method at this level of missing performs the worst of all methods analyzed. The model-based method performs slightly less accurately in general than the baseline estimates, while the hotdeck and no imputation methods actually perform better than baseline.

4.14 Originally missing plus 20% artificially created NMAR imputed race proportion estimates

In addition to the originally missing, the 20% artificially created NMAR race values resulted in approximately 21.3% missing and 17.8% weighted missing for the entire survey. At this level of artificially created missing the weighted and unweighted missing percentages were within 3% to 4% of each other.

Table 16: Weighted and unweighted race value missing percent by BRFSS subset where missingness is NMAR and depends on white status.

Survey Non-Respondents plus 20% artificial NMAR Missing Race Values	NJ	NY	OR	WA	U.S. BRFSS
Weighted Percent Missing	16.8%	18.1%	21.6%	19.3%	17.8%
Percent Missing	19.1%	22.0%	23.0%	22.5%	21.3%
Total Sampled	15761	6060	5302	15319	467,333

Table 17: Performance of 20% artificially created NMAR race values estimates by imputation method where missingness depends on white status.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	23%	56%	40%	23%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.580	.287	0.473	0.387	
Difference between sum of absolute differences at current level and baseline absolute differences	0.305	-0.008	0.184	0.110	

At this level of missing, the BRFSS imputation was the most accurate, performing better than baseline due to the NMAR mechanism being tailor made for the BRFSS method. No other method was as accurate as at the baseline level. The no imputation method was the worst of the other three, at an average of approximately .01 off per estimate. Interestingly, the perennially inaccurate Oregon "Native" estimates were marginally better for the hotdeck and model-based estimates (Figure 11).

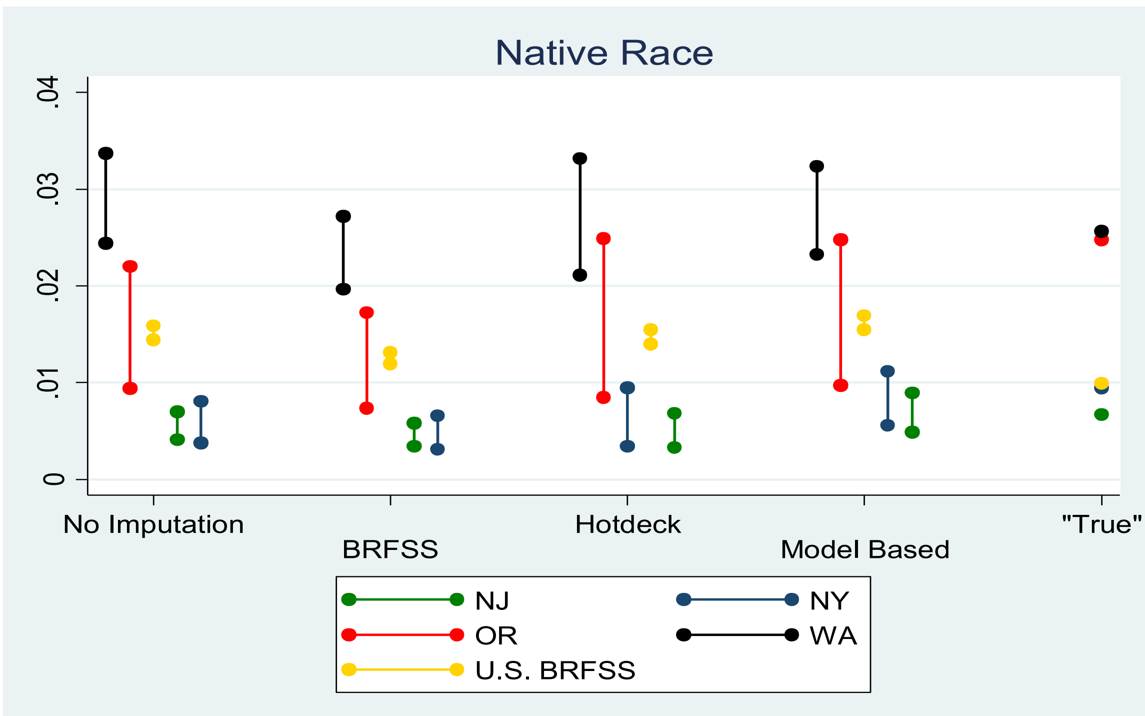

Figure 11: Originally missing plus 20% NMAR imputed race proportion estimates by imputation method and subset of BRFSS where missingness depends on white status. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the "true" proportion.

5. Age Imputation Results

5.1 Summary of all cases considered

As with the race imputation, two metrics were used to measure the accuracy of age proportion estimates:

1) Percent of estimates whose 95% confidence intervals contained the true value;
2) Total absolute difference between each estimate and the true value; and
3) Distance between each estimate and the originally missing estimate for that method.

Based on the average total distance from the true value, the accuracy of complete case method, multiple hot deck imputation, and multiple model-based imputation methods appear to break down at the 20% artificially missing level, while the BRFSS imputation method behaves erratically before that (Figure 13). As in the case of the race estimate 95% confidence intervals, there is no range of error represented by the percent of age estimate 95% confidence intervals that contain the true value. A confidence interval either contains the true value, or it does not. This means that some confidence intervals that miss the true value just barely are considered inaccurate by this metric, and it causes the accuracy measurement to be more erratic than the mean absolute error (Figure 12).

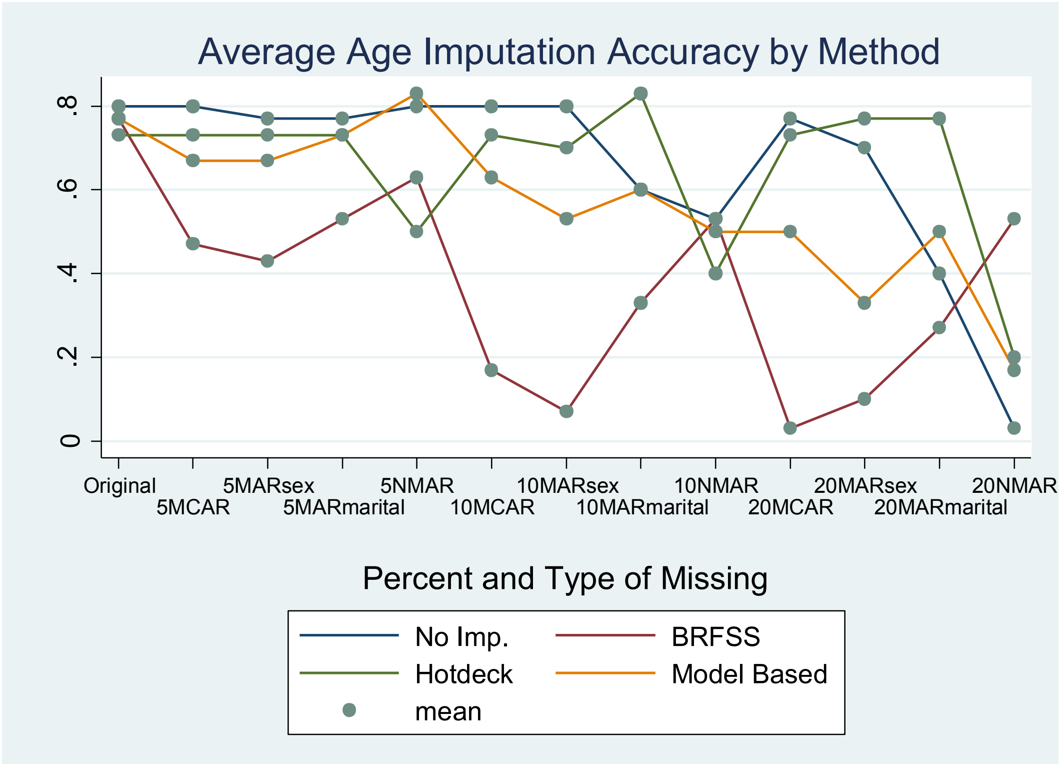

Figure 12: Each point represents the average percent of confidence intervals that contained the true proportion.

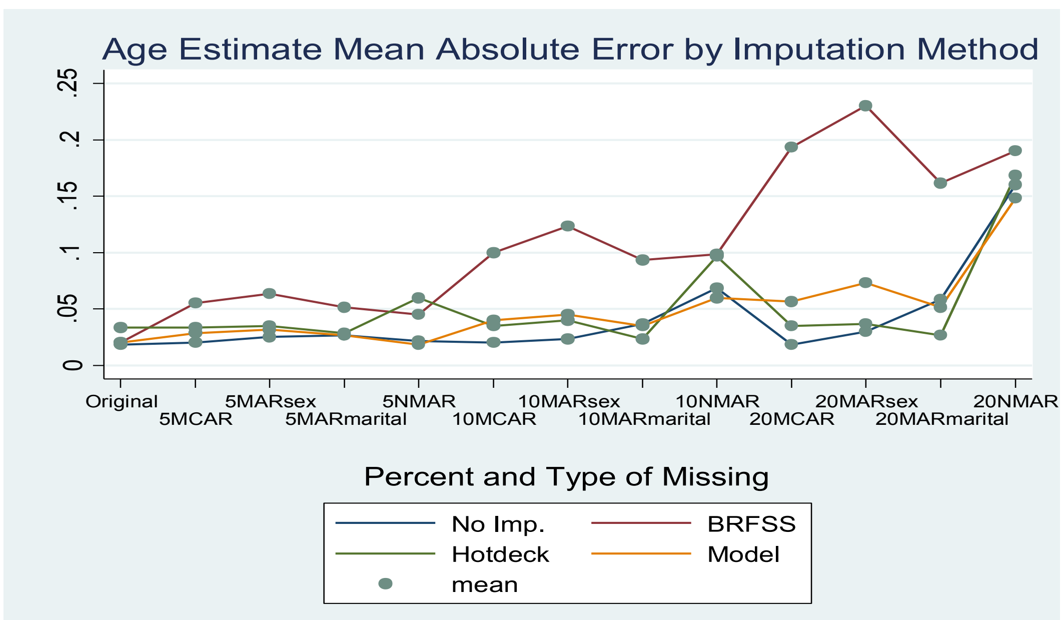

Figure 13: Mean absolute error of all age imputation estimates by missing level and method. The x-axis labels are, from left to right, originally missing, 5% MCAR, 5% MAR where missingness depends on sex, … , 20% NMAR.

5.2 Originally missing imputed age proportion estimates

The percent of age values that were originally missing were all between 0.96% and 1.57% for OR, WA, NY, NJ, and the entire BRFSS survey. The weighted missing percentages were similar to the non-weighted, varying between 0.69% and 1.24% (Table 18).

Table 18: Weighted and unweighted missing percent by BRFSS subset.

Survey Respondents with Missing Age Values	NJ	NY	OR	WA	U.S. BRFSS
Weighted Percent Missing	1.13%	1.24%	0.95%	0.86%	.69%
Percent Missing	1.40%	1.57%	1.13%	1.12%	.96%
Total Sampled	15,761	6,060	5,302	15,319	467,333

The difference between the imputation proportion estimates and the ACS proportions for the originally missing age values was generally quite small. The average absolute difference was .003, and the average absolute value of the percent change from imputation estimate to true proportion was approximately 2%. The 18-24 year old NJ and WA hotdeck estimates were the worst, having two estimates that were off by approximately 17%,—visibly noticeable in Figure 14. In contrast to the race proportion estimates for the race categories with extremely small proportions in absolute terms (e.g. “Other” and “Native”), the age categories did not have proportions that were quite so hard to estimate.

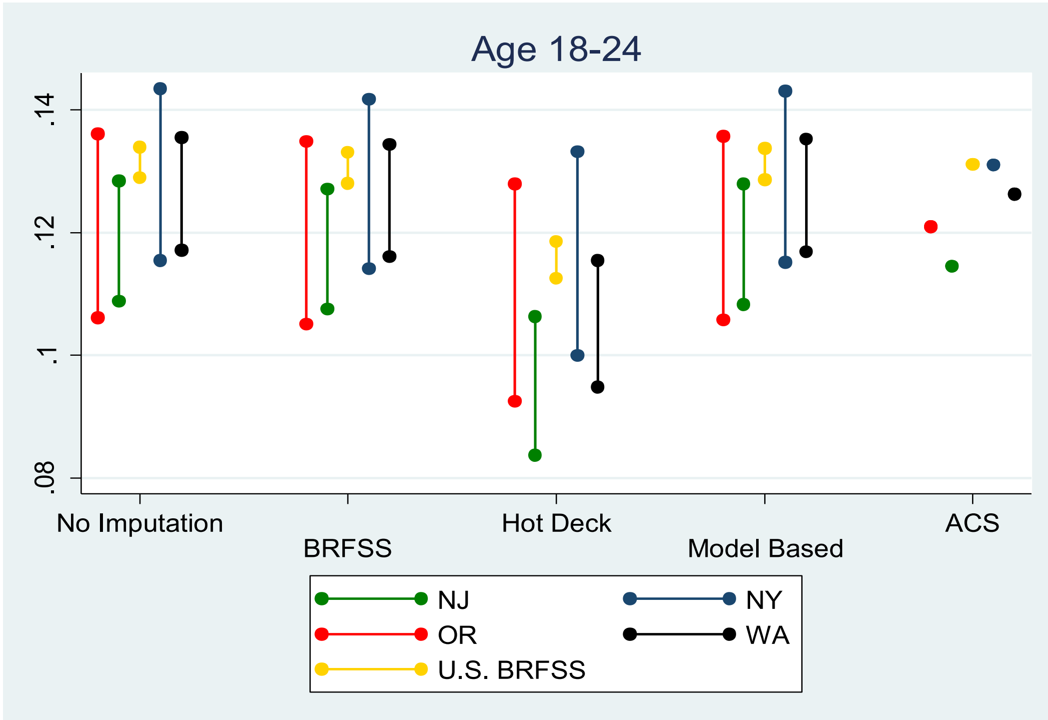

Figure 14: Originally missing imputed age 18-24 group proportion estimates by imputation method and subset of BRFSS.

As in the case of the race imputation estimates, we can use the originally missing age imputation estimates as a kind of baseline to compare levels of artificially missing imputations to. Although none of the imputation method estimates had 95% of their confidence intervals span the true proportion, the methods performed better than for race at baseline (Table 19).

Table 19: Performance of originally missing age imputation estimates by imputation method.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	80%	76%	73%	76%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.112	0.118	0.198	0.117	

5.3 Originally missing plus 5% artificially created MCAR imputed age proportion estimates

In addition to the originally missing, the 5% artificial MCAR resulted in approximately 5.8% missing and 5.5% weighted missing for the entire survey. The weighted and unweighted missing percentages did not differ substantially for any subset analyzed (Appendix 7.2). At this level, the BRFSS imputation method performed substantially worse than at the originally missing level (Table 20). The model-based method was slightly worse than the baseline accuracy, while the hotdeck and no imputation methods performed approximately the same as at baseline. Figure 15 shows the typical accuracy of estimates at this level of missing.

Table 20: Performance of 5% MCAR age imputation estimates by imputation method.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	80%	46%	73%	66%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.119	0.333	0.200	0.172	
Difference between sum of absolute differences at current level and baseline absolute differences	0.007	0.215	0.002	0.055	

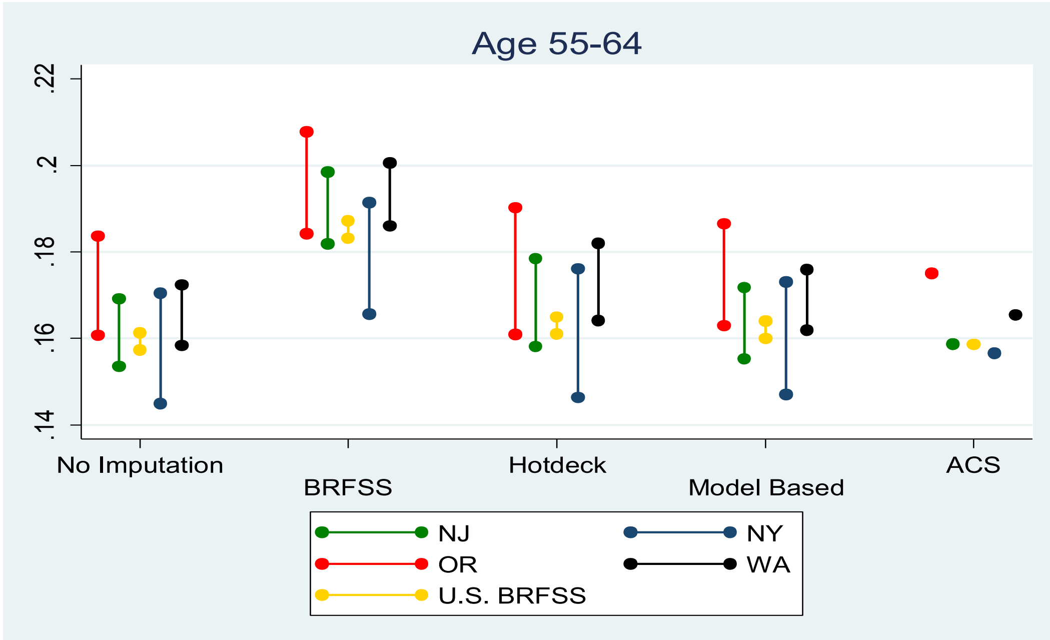

Figure 15: Originally missing plus 5% MCAR imputed age proportion estimates by imputation method and subset of BRFSS. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the "true" proportion.

5.4 Originally missing plus 5% artificially created MAR where missingness depends on a variable used as a covariate in the hotdeck and model-based imputed age proportion estimates

In addition to the originally missing, the artificially created 5% MAR age values where the missingness depends on gender resulted in approximately 6% missing and 6.7% weighted missing for the entire survey. The weighted and unweighted missing percentages did not differ substantially for the other subsets analyzed (Appendix 7.2).

At this level, the BRFSS performed the worst overall of all methods. The hotdeck method performed at approximately the same accuracy as at baseline, while no imputation and the model-based method were marginally worse than at baseline. The age 65 and up group model-based estimates were less accurate than the model-based estimates for other groups (Figure 16).

Table 21. Performance of 5% MAR age where missingness depends on gender imputation estimates by imputation method.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	76%	43%	73%	66%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.148	0.381	0.205	0.185	
Difference between sum of absolute differences at current level and baseline absolute differences	0.036	0.263	0.007	0.068	

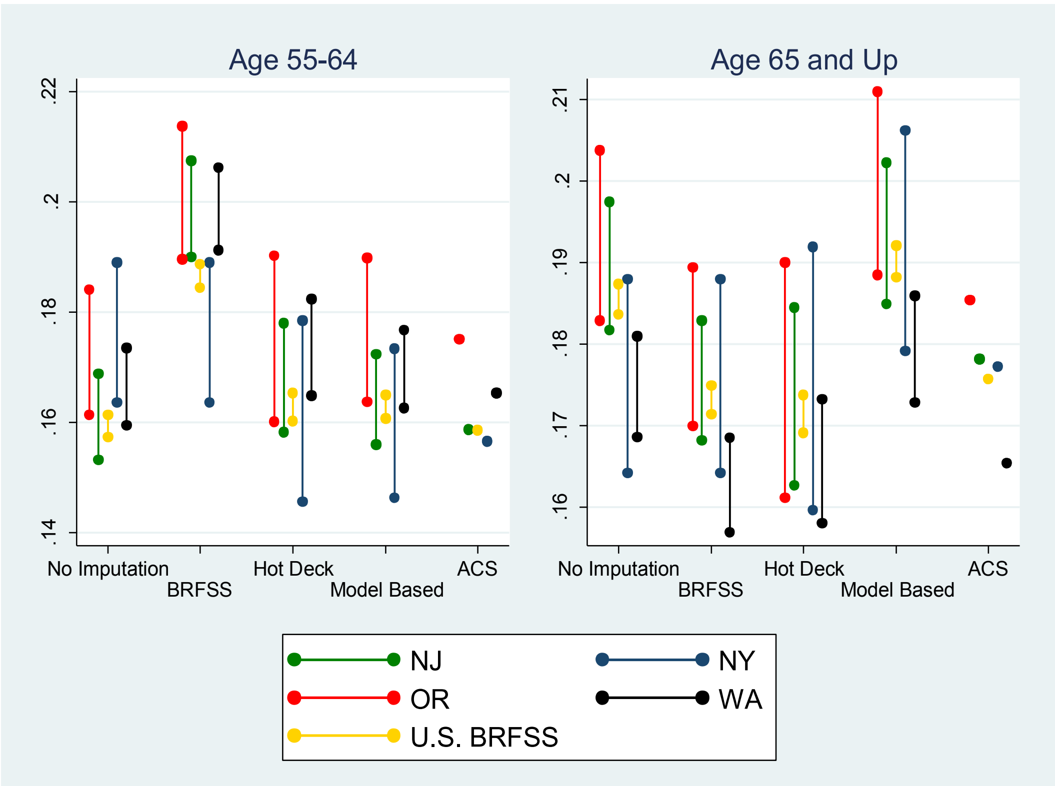

Figure 16: Originally missing plus 5% MAR imputed age proportion estimates by imputation method and subset of BRFSS where missingness depends on gender. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the "true" proportion.

5.5 Originally missing plus 5% artificially created MAR where missingness depends on a variable not used as a covariate in the hotdeck and model-based imputed age proportion estimates

In addition to the originally missing, the artificially created 5% MAR age missingness resulted in approximately 5.9% missing and 5.4% weighted missing for the entire survey. At this level, the weighted and unweighted missing percentages did not differ substantially for any subset analyzed (Appendix 7.2).

Table 22: Performance of 5% artificially created MAR age values where missingness depends on marital status imputation estimates by imputation method.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	76%	53%	73%	73%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.157	0.310	0.165	0.164	
Difference between sum of absolute differences at current level and baseline absolute differences	0.045	0.192	-0.033	0.047	

Once again, the BRFSS method performed the worst. The model-based and no imputation methods were slightly less than their baseline accuracy, while the hotdeck method outperformed its baseline accuracy (Table 22). As in the case of the MAR 5% data where missingness depended on gender, the model-based method performed poorly for the age 65 and up age group, overestimating for every subset analyzed (Figure 17). Nevertheless, the model-based method performed well enough at estimating the other age groups that it was almost as good as at the baseline originally missing accuracy level.

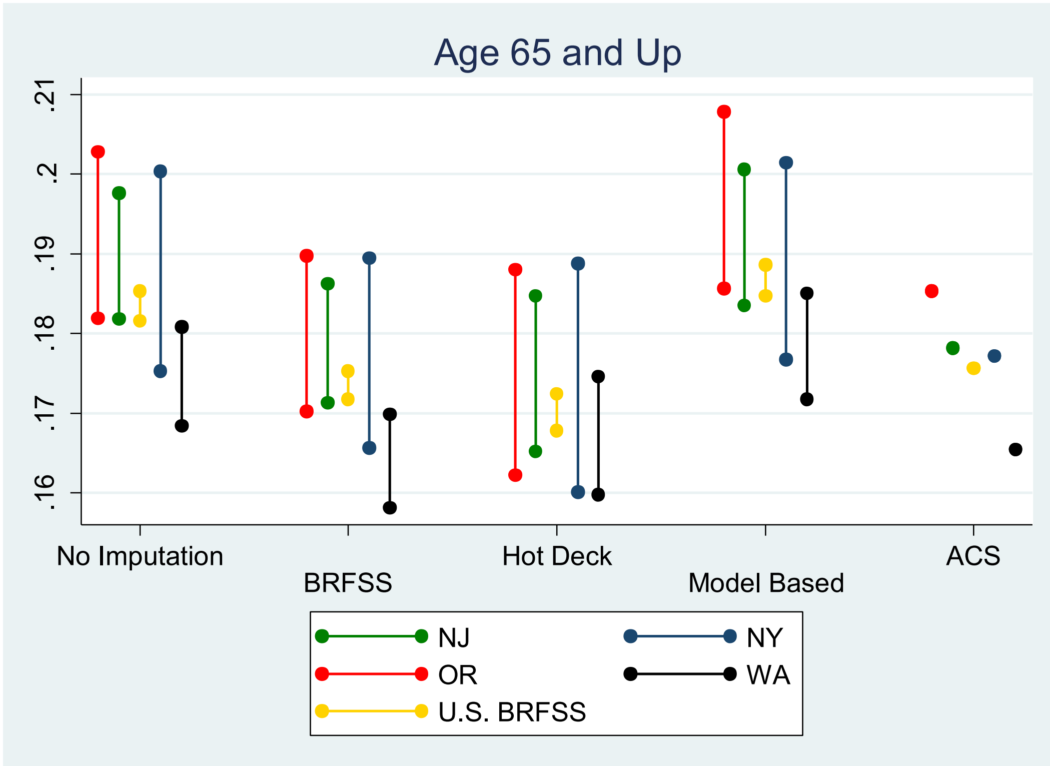

Figure 17: Originally missing plus 5% MAR imputed age proportion estimates by imputation method and subset of BRFSS where missingness depends on marital status. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the "true" proportion.

5.6 Originally missing plus 5% artificially created NMAR imputed age proportion estimates

In addition to the originally missing, the 5% artificially created missing age values resulted in approximately 5.9% missing and 3.4% weighted missing for the entire survey. The weighted percent missing was approximately 2% less than the percent missing for all subsets analyzed at this level of missing. This can be explained by the fact that the artificial NMAR mechanism used created missingness that was more likely for age group 65 and up. This age group was weighted less relative to the others in the survey, and therefore the weighted percentages missing are less than the percentages missing (Table 23).

Table 23: Weighted and unweighted missing percent by BRFSS subset where missingness depends on age.

Survey Non-Respondents plus 5% artificial NMAR Age Values	NJ	NY	OR	WA	U.S. BRFSS
Weighted Percent Missing	3.4%	3.3%	3.5%	3.5%	3.4%
Percent Missing	5.7%	5%	5.8%	6.1%	5.9%
Total Sampled	15,761	6,060	5,302	15,319	467,333

Table 24: Performance of 5% artificially created NMAR age values where missingness depends on age group imputation estimates by imputation method.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	80%	63%	50%	83%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.127	0.271	0.355	0.111	
Difference between sum of absolute differences at current level and baseline	0.015	0.153	0.157	-0.006	

Interestingly, and in contrast to the 5% MAR cases, the model-based imputation method performed relatively well for the age 65 and up group (Figure 18). This contributed to the model-based method performing better than baseline at this level. The hotdeck method performed noticeably worse at this level than at the baseline originally missing (Table 24). This appears to be a result of many estimates being marginally inaccurate, as opposed to one age group or one survey subset having substantially inaccurate estimates. The no imputation method, meanwhile, performed only slightly less accurately than baseline. The BRFSS method was approximately as accurate as the hotdeck method.

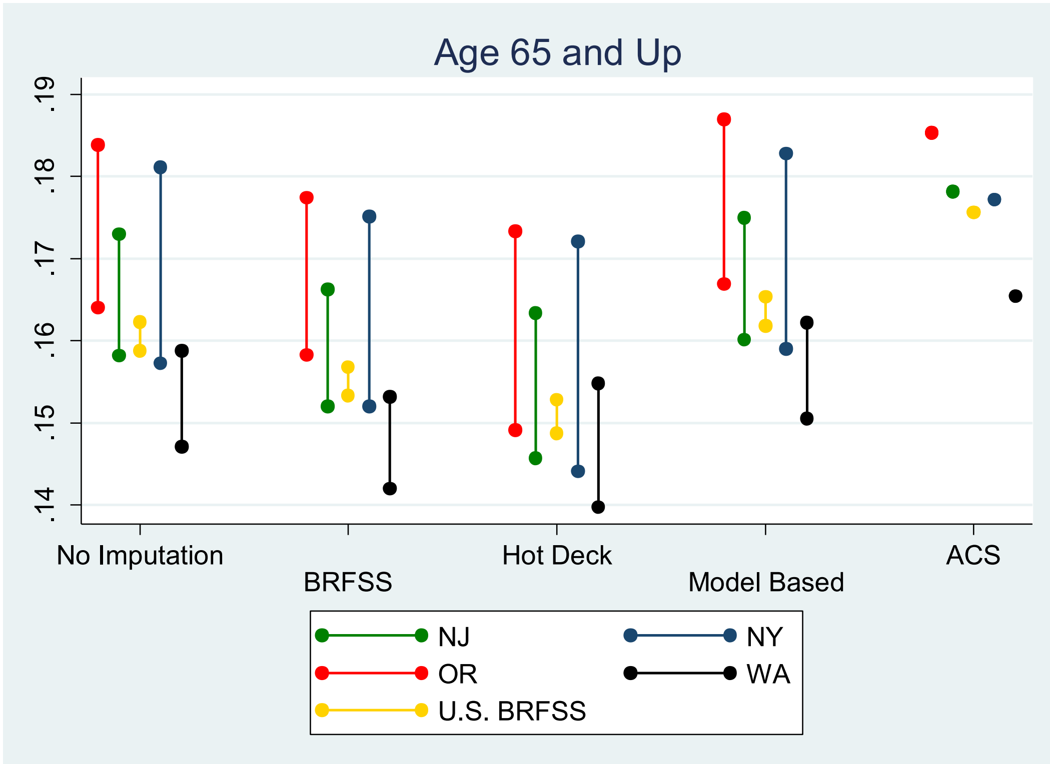

Figure 18: Originally missing plus 5% NMAR imputed age proportion estimates by imputation method and subset of BRFSS where missingness depends on age group 65 and up status. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the "true" proportion.

5.7 Originally missing plus 10% artificially created MCAR imputed age proportion estimates

In addition to the originally missing, the 10% artificially created MCAR age values resulted in approximately 10.8% missing and 10.5% weighted missing for the entire survey. The missing and weighted missing percentages did not differ substantially at this level for any subset analyzed (Appendix 7.2).

Once again, the BRFSS method performed the worst of all methods analyzed. Interestingly, the model-based method performed marginally worse than its baseline originally missing accuracy. This is due at least partially due to poor performance for age group 65 and up

(Figure 19). The hotdeck and no imputation methods, meanwhile, were approximately as accurate as at baseline.

Table 25: Performance of 10% artificially created MCAR age values by imputation method.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	80%	16%	73%	63%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.116	0.599	0.207	0.238	
Difference between sum of absolute differences at current level and baseline absolute differences	0.004	0.481	0.009	0.121	

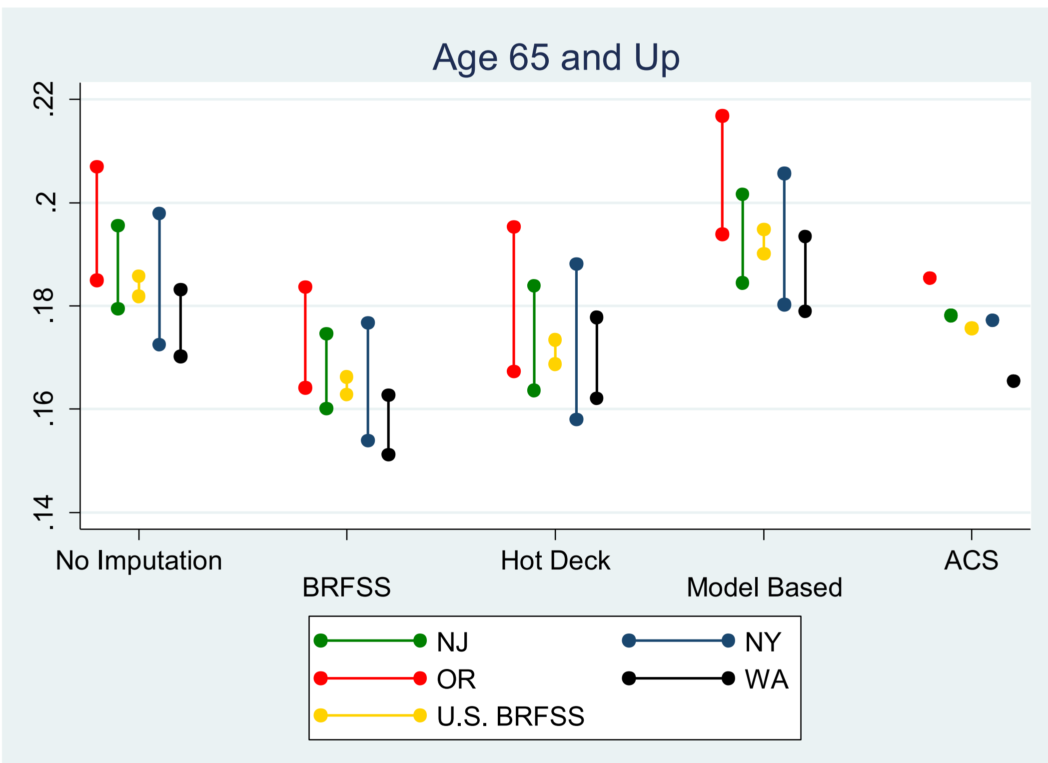

Figure 19: Originally missing plus 10% MCAR imputed age proportion estimates by imputation method and subset of BRFSS. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the "true" proportion.

5.8 Originally missing plus 10% artificially created MAR where missingness depends on a variable used as a covariate in the hotdeck and model-based imputed age proportion estimates

In addition to the originally missing, the 10% artificially created MAR age values resulted in approximately 11.2% missing and 12.8% weighted missing for the entire survey. At this level of missing, the weighted and unweighted missing percentages were similar for all subsets (Appendix 7.2).

Table 26: Performance of 10% artificially created MAR age values by imputation method where missingness depends on gender.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	80%	6%	70%	53%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.135	0.741	0.244	0.267	
Difference between sum of absolute differences at current level and baseline absolute differences	0.023	0.623	0.046	0.150	

Once again, the BRFSS imputation method performed the worst, while the hotdeck, model-based, and no imputation methods performed slightly less than precisely than at their baseline accuracy. The relatively reduced performance of the model-based imputation estimates at this level is, again, due at least in part to the estimates for age group 65 and up, which were off for every subset of the survey analyzed (Figure 20).

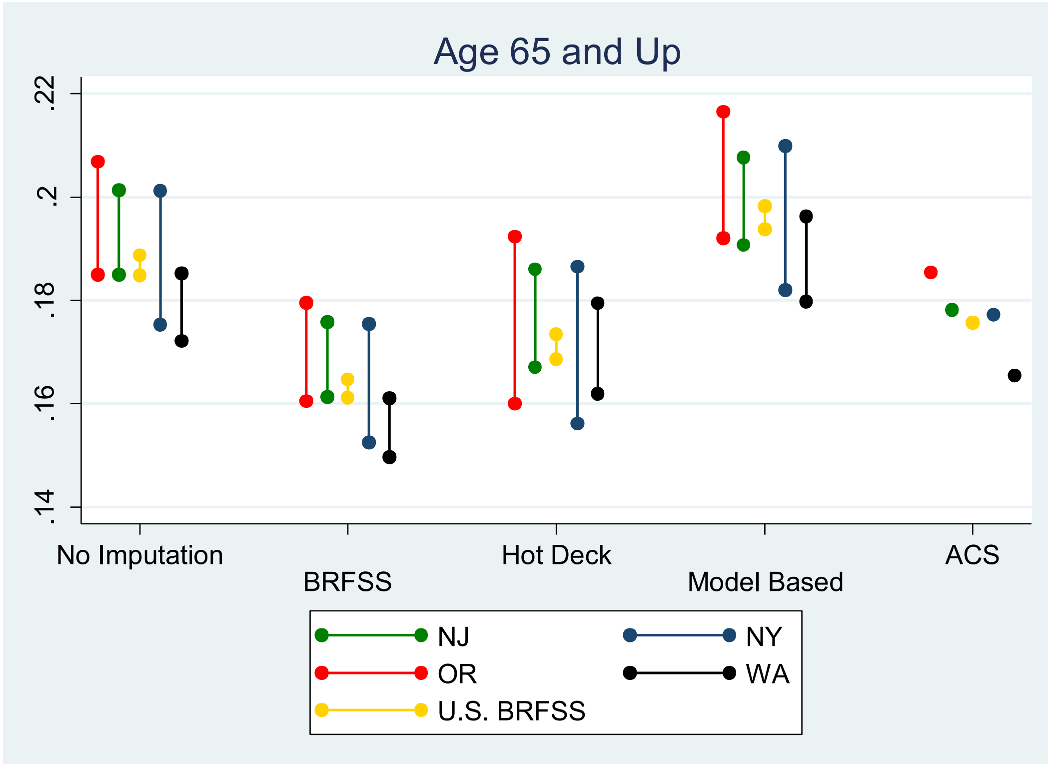

Figure 20: Originally missing plus 10% MAR imputed age proportion estimates where missingness of age values depends on gender by imputation method and subset of BRFSS. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the "true" proportion.

5.9 Originally missing plus 10% artificially created MAR where missingness depends on a variable not used as a covariate in the hotdeck and model-based imputed age proportion estimates

In addition to the originally missing, the 10% artificial MAR data where the missingness depended on marital status resulted in approximately 11.2% missing and 12.8% weighted missing for the entire survey. The weighted and unweighted missing percentages did not differ substantially for any subset of the survey analyzed.

Once again the BRFSS method performed the worst. Interestingly, however, the complete case method and the model-based method both performed marginally less accurately than their baseline accuracy levels. For both methods, this is due in part to their performance estimating the age 65 and up group. The hotdeck method performed better than its baseline accuracy.

Table 27: Performance of 10% artificially created MAR age values by imputation method where missingness depends on marital status.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	60%	33%	83%	60%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.222	0.564	0.140	0.208	
Difference between sum of absolute differences at current level and baseline absolute differences	0.110	0.446	-0.058	0.091	

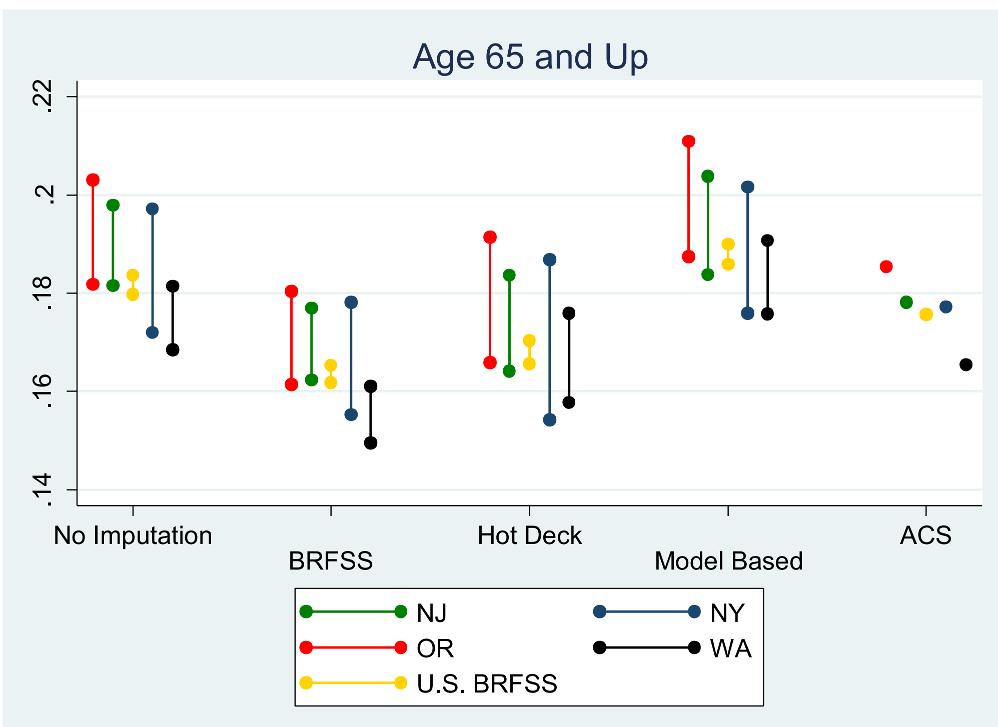

Figure 21: Originally missing plus 10% MAR imputed age proportion estimates where missingness of age values depends on marital status by imputation method and subset of BRFSS. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the "true" proportion.

5.10 Originally missing plus 10% artificially created NMAR imputed age proportion estimates

In addition to the originally missing, the 10% artificially created NMAR age values resulted in approximately 10.8% missing and 6.3% weighted missing for the entire survey. The weighted and unweighted missing percentages differed by between 3% and 4% for this case, across all survey subsets analyzed (Table 28). Effectively, this made the imputation estimates only effect approximately 7% of the sample sizes.

Table 28: Weighted and unweighted race value missing percent by BRFSS subset where missingness is NMAR and depends on white status.

Survey Non-Respondents plus 10% artificial NMAR Missing Age Values where missingness depends on age 55-64 group status	NJ	NY	OR	WA	U.S. BRFSS
Weighted Percent Missing	7.5%	7.1%	6.9%	6%	6.3%
Percent Missing	10.5%	10.4%	11.9%	10.7%	10.8%
Total Sampled	15,761	6,060	5,302	15,319	467,333

Table 29: Performance of 10% artificially created NMAR age values by imputation method where missingness depends on age 65 and up status.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	53%	53%	40%	50%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.411	0.587	0.579	0.359	
Difference between sum of absolute differences at current level and baseline	0.299	0.469	0.381	0.242	

At this level of missing every imputation method performed noticeably worse than their baseline accuracy. In particular, no method accurately estimated the age 65 and up group. For this NMAR level, in contrast to the inaccurate model-based estimates at the MAR levels, the reason for the inaccuracy is because the NMAR mechanism used to create the artificially missing made missingness more likely for the age 65 and up age group. The imputation methods simply have no way to "know" where to put the missing NMAR data.

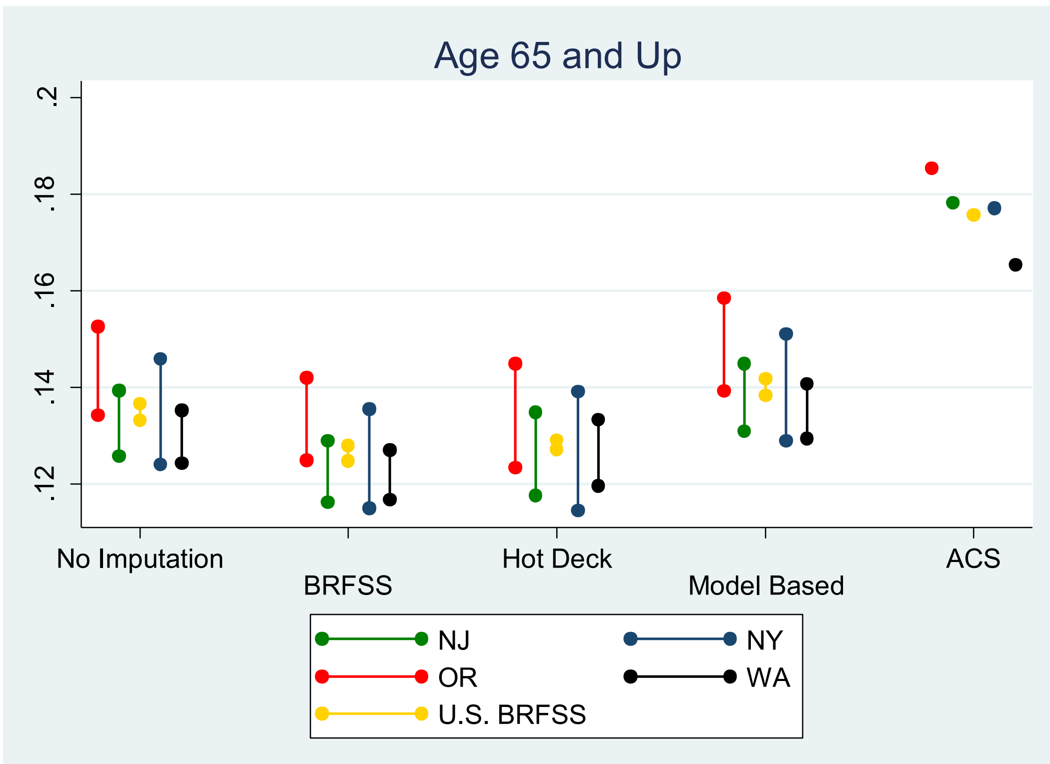

Figure 22: Originally missing plus 10% NMAR imputed age proportion estimates by imputation method and subset of BRFSS where missingness of age values depends on aged 65 and up status. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the "true" proportion.

5.11 Originally missing plus 20% artificially created MCAR imputed age proportion estimates

In addition to the originally missing, this resulted in approximately 20.7% missing and 20.5% weighted missing for the entire survey. The weighted and unweighted missing percentages at this level did not differ substantially for any subset analyzed (Appendix 7.2).

Once again, the BRFSS method did the worst. Although for the age 45-54 group the estimates did well for NY and NJ (Figure 23), presumably because the mean imputed by that method just happened to fit the demographics of those states.

Table 30: Performance of 20% artificially created MCAR age values estimates by imputation method.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	76%	3%	73%	50%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.109	1.163	0.207	0.339	
Difference between sum of absolute differences at current level and baseline absolute differences	-0.003	1.045	0.009	0.222	

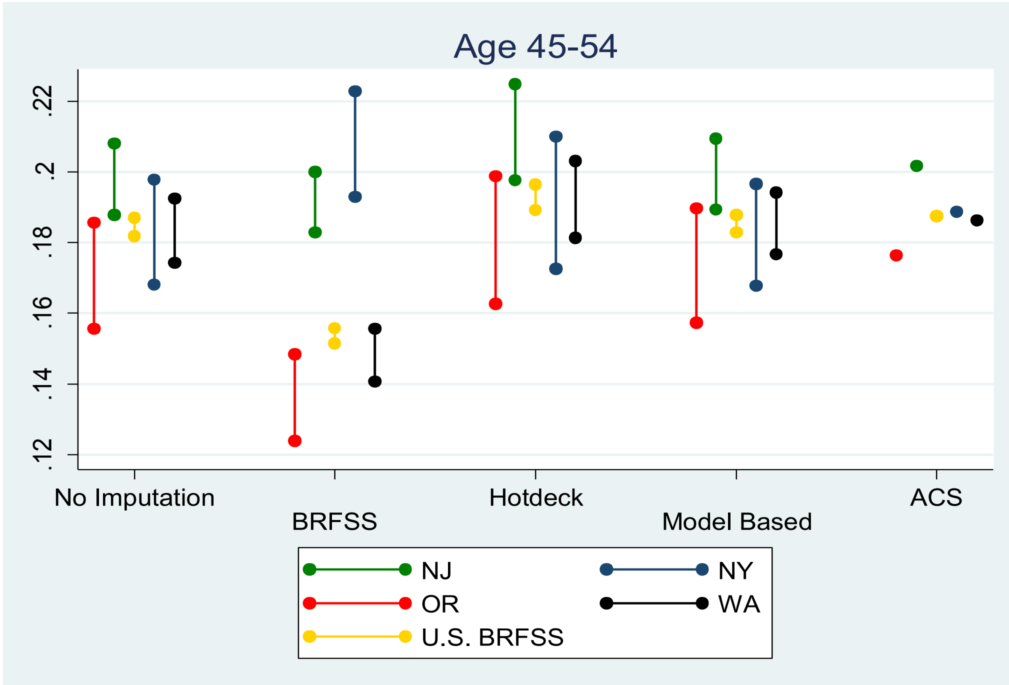

Figure 23: Originally missing plus 20% MCAR imputed age 45-54 proportion estimates by imputation method and subset of BRFSS. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the “true” proportion.

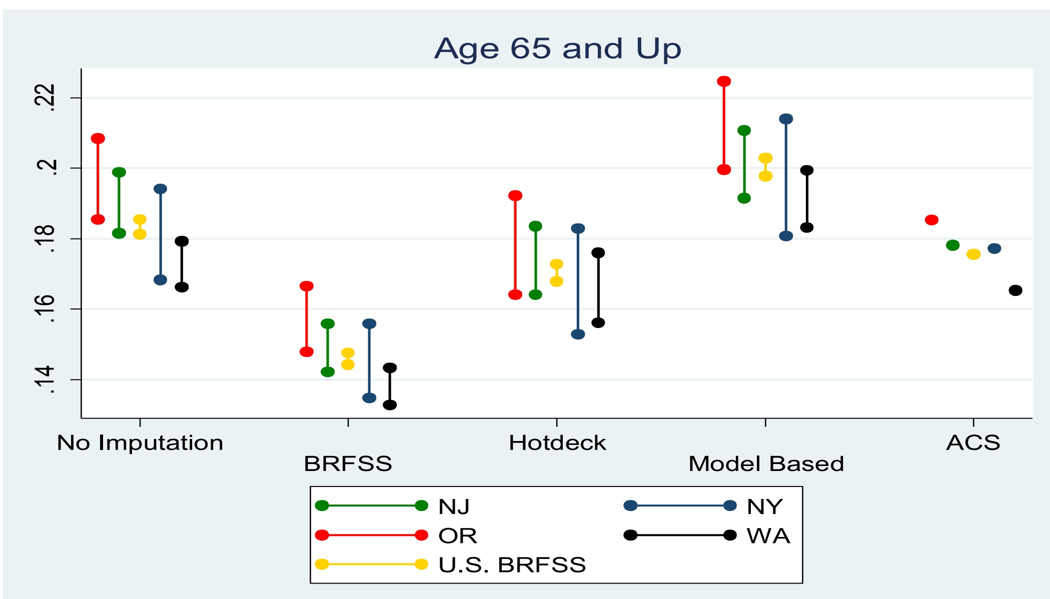

Figure 24: Originally missing plus 20% MCAR imputed age 65 and up proportion estimates by imputation method and subset of BRFSS. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the “true” proportion.

At this level, the complete case method outperformed its baseline accuracy, while the hotdeck method was within striking distance of its baseline accuracy. The model-based method fell off slightly in accuracy at this level of missing, for example in the age 65 and up group (Figure 24).

5.12 Originally missing plus 20% artificially created MAR imputed age proportion estimates where missingness depends on a variable used as a covariate in the hotdeck and model-based imputed age proportion estimates

In addition to the originally missing, the 20% artificially created MAR age values where missingness depends on gender resulted in approximately 21% missing and 24.8% weighted missing for the entire survey. For this level of missing, the weighted and unweighted percentages differed by approximately 3% to 4% for all subsets. The result is effectively approximately 25% missing.

Once again the BRFSS method performed the worst. Perhaps surprisingly, the model-based method was noticeably inaccurate. See, for example, age 25-34 proportions (Figure 25).

Table 31: Weighted and unweighted age value missing percent by BRFSS subset where missingness is MAR and depends on gender.

Survey Non-Respondents plus 20% artificial MAR Missing Age Values where missing depends on gender	NJ	NY	OR	WA	U.S. BRFSS
Weighted Percent Missing	25.3%	24.8%	25%	25.6%	24.8%
Percent Missing	21.5%	21.5%	21.3%	22%	21%
Total Sampled	15,761	6,060	5,302	15,319	467,333

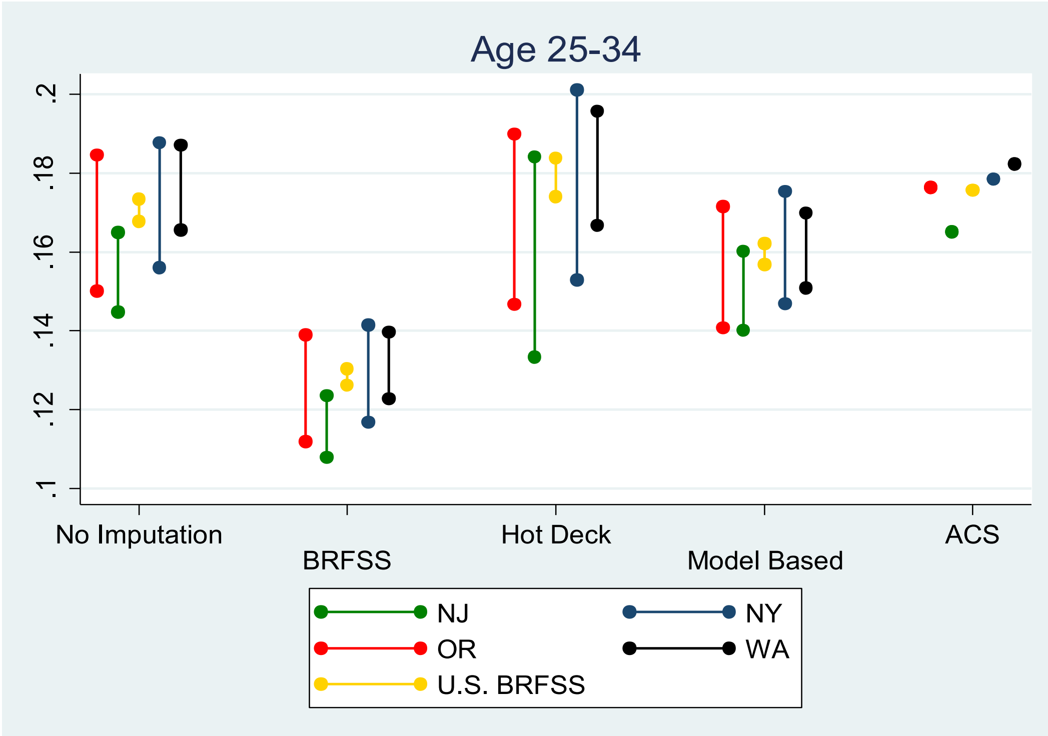

Figure 25: Originally missing plus 20% MAR imputed age 25-34 proportion estimates by imputation method and subset of BRFSS where missingness depends on gender. The dots connected by lines represent the upper and lower confidence limits of the imputation estimates, while the single dots represent the "true" proportion.

Table 32: Performance of 20% artificially created MAR age values estimates by imputation method where missingness depends on gender.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	70%	10%	76%	33%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.183	1.379	0.221	0.436	
Difference between sum of absolute differences at current level and baseline absolute differences	0.071	1.261	0.023	0.319	

5.13 Originally missing plus 20% artificially created MAR imputed age proportion estimates where missingness depends on a variable not used as a covariate in the hotdeck and model-based imputed age proportion estimates

In addition to the originally missing, the 20% artificially created MAR age values where missingness depends on marital status resulted in approximately 20.6% missing and 19.4% weighted missing for the entire survey. At this level, the weighted and unweighted missing percentages did not differ substantially (Appendix 7.2).

Table 33: Performance of 20% artificially created MAR age values estimates by imputation method where missingness depends on gender.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	40%	26%	76%	50%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.348	.972	0.162	0.314	
Difference between sum of absolute differences at current level and baseline	0.236	0.854	-0.036	0.197	

At this level, the complete case method finally slipped in accuracy, while the hotdeck method outperforms its baseline level. The model-based method did slightly better than no imputation and the BRFSS method was not very accurate at all.

5.14 Originally missing plus 20% artificially created NMAR imputed age proportion estimates

In addition to the originally missing, the 20% artificial NMAR data resulted in approximately 20.7% missing and 12% weighted missing for the entire survey. The difference in weighted and unweighted missing at this level is due to the NMAR mechanism used, which created artificially missing among survey respondents in the age 65 and up group. This group was weighted less relative to the other age groups, and therefore when they were designated as missing, the weighted percent missing was less than the actual percent missing. As a result, the approximately 12% weighted missing at this level is more reflective of the effect of estimates than is the non-weighted approximately 20% missing. At this level of missing all methods are inaccurate, although the model-based method is marginally better than the others.

Table 34: Weighted and unweighted age value missing percent by BRFSS subset where missingness is NMAR and depends on age group 65 and up status.

Survey Non-Respondents plus 20% artificial NMAR Missing Age Values	NJ	NY	OR	WA	Total BRFSS
Weighted Percent Missing	12.5%	12.4%	13.3%	11.7%	12%
Percent Missing	19%	19.1%	23%	21.2%	20.7%
Total Sampled	15,761	6,060	5,302	15,319	467,333

Table 35: Performance of 20% artificially created NMAR age values estimates by imputation method where missingness depends on age group 65 and up status.

	No Imputation	BRFSS Imputation	Hot Deck Multiple Imputation	Model Based Multiple Imputation	Nominal
Percent of estimates whose 95% CI contain the true proportion	3%	53%	20%	16%	95%
Sum (across all estimates) of absolute differences between imputation estimate and true proportion	0.964	1.139	1.014	0.892	
Difference between sum of absolute differences at current level and baseline absolute differences	0.852	1.021	0.816	0.775	

6. Discussion

6.1 Summary of proportion imputation estimate findings

With the exception of the BRFSS imputation method at NMAR levels, which had a demographic explanation, the methods compared had the following order in terms of accuracy:

1) Hotdeck and no imputation were the best methods, having approximately similar accuracy.
2) Model-based imputation was second in accuracy, slightly underperforming the hotdeck and no imputation methods.
3) The BRFSS race imputation method was a distant last, performing worse in accuracy than every other method except in cases where the NMAR data was tailor made for the method.

Perhaps the most intuitively attractive metric to measure accuracy is to compare imputation estimates at artificially missing levels to the baseline originally missing. Because the true value of race proportions in the population is not known, the best one can expect out of any imputation method using BRFSS data is to estimate as accurately as possible using the given final weights. By this metric, the hotdeck method was the best, frequently performing more accurately than the baseline originally missing estimates. No imputation was a close second and model-based imputation was just behind.

Depending on the mechanism of the missingness, the specific demographics, and the imputation method, there will always be instances where certain imputation methods perform better than others. For this reason, I mostly assessed the accuracy of the tested imputation methods in the aggregate, averaging over all estimates at each level and mechanism of missingness. On the other hand, future researchers may be interested to explore the accuracy of each method averaged over all estimates for a given category. Consider, for example, the distribution of the difference from the baseline age 65 and up estimates averaged over all missingness levels and mechanisms (Figure 26).

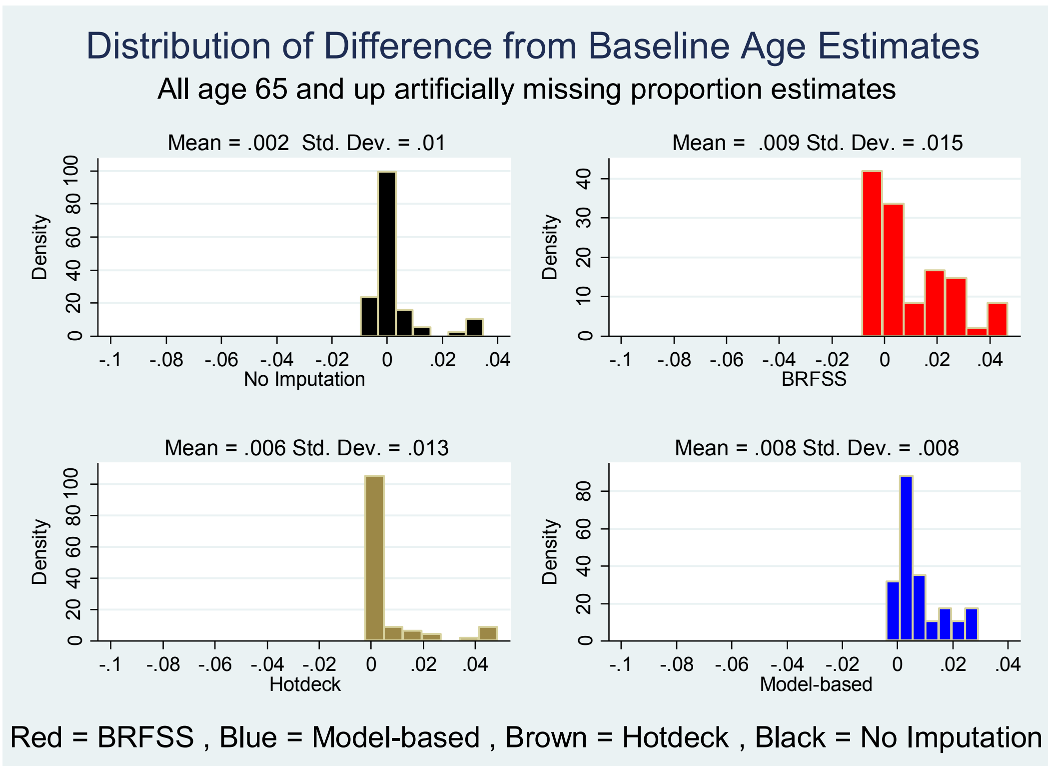

Figure 26: The distribution of the difference between all artificial levels and mechanisms of missingness and the baseline age 65 and up estimates for each tested imputation method.

There is also value in comparing imputation proportion estimates to proportion estimates from the American Community Survey and U.S. Census estimates. These were the proportions that the 2012 BRFSS survey design Iterative Proportional Fitting or “raking” algorithm attempted to match proportions to and may very well be accurate representations of the actual population proportions. By this metric, no imputation, hotdeck imputation, and model-based imputation were always in the relative ballpark of the true value.

The percent of 95% confidence intervals that contain the true proportion was not an ideal measurement of accuracy. Strictly speaking, confidence intervals measure precision. I used them here as a kind of proxy for accuracy, as well as a way to compare methods of calculating confidence intervals. In addition, the fact that an estimate can be outside an interval just barely and still not count as accurate, the arbitrary choice of the true proportion, and the differing methods of computing confidence intervals all make it the least desirable assessment of accuracy. Nevertheless, this metric offers insight into the preferability of calculating confidence intervals in multiple imputation by accounting for between imputation variance. For example, the hotdeck multiple imputation method had a higher percent of confidence intervals that contained the true proportion than the model-based multiple imputation method at every level of missingness and mechanism of missingness, and beat or tied the complete case method on all but two out of thirteen levels tested.

6.2 Comparisons of hotdeck and model-based proportion estimate accuracy for MAR data where missingness depended on a variable used as a covariate in the imputation model vs. MAR data where missingness depended on a variable not used as a covariate in the imputation model

For the hotdeck and model-based imputation methods, the accuracy of estimates for MAR data where the missingness depends on a covariate used in the imputation model (gender) was no better than where the missingness did not depend on a covariate used in the imputation model (marital status). For the model-based imputation, the estimates were actually farther on average from the true proportion, and the hotdeck imputation method was only slightly better on average for artificially created MAR data where the missingness depended on gender compared to MAR data where the missingness depended on marital status.

6.3 The effect of state level survey design on race proportion estimates

Based on the assumption that OR and WA, and NY and NJ, respectively, have similar demographics and the fact that they have different sample sizes and number of survey strata, a comparison of imputation estimates gives some insight into the effect of state level survey design on imputation estimates. The question of interest was whether the accuracy of the BRFSS imputation race proportion estimates for OR and WA, and for NY and NJ was significantly different between states. Because the BRFSS imputation method imputed the most common race by geographic strata, presumably the imputation estimates would be less accurate for states that had a limited number of geographic strata (i.e., NY with 2 and OR with 1), and more accurate for states that had a substantial number of geographic strata (i.e., NJ with 23 and WA with 34).

To evaluate the degree of this effect, I calculated the total difference between the race proportion imputation estimates and the true proportion for BRFSS imputation estimates and missing data level by state. Although for both pairs of states, the state with more geographic strata and larger sample size had proportion estimates that on average were more accurate; neither difference was substantial.

6.4 Summary of findings

A number of unexpected results came out of this research, and some that were expected. Perhaps expectedly, the estimation of small population proportions is difficult to do accurately. Also predictably, the BRFSS imputation method performed the worst out of all imputation methods tested. At the same time, however, the BRFSS imputation estimates were not substantially worse for states with only one or two geographic strata compared to states with many geographic strata and a higher sample size but similar demographics. This is contrary to what seemed sensible a priori.

Of the remaining three imputation estimates tested, all seemed to perform not only similarly to each other (with some minor exceptions), until the 20% NMAR level, but also similarly to the other levels and types of missing. Ahead of time, it seemed likely that the imputation methods would get progressively worse as the percent of missing increased and for the NMAR levels. This mostly did not happen. The complete case estimates, model-based estimates, and hotdeck method estimates all performed similarly, not only for a given level of missingness and missingness mechanism, but also compared to other levels of missingness and missingness mechanism. They dropped off substantially at the 20% NMAR level. With the exception of the 10% NMAR age estimates, this was true for both age and race proportion estimates. It is unclear why the complete case method doesn't break down for the 5% and 10% NMAR levels. It is also unclear why the model-based method should perform worse than no imputation at the 20% MCAR, and MAR levels, even if only moderately so.

It was unexpected that complete case estimation was as accurate, or more accurate, than both the hotdeck and model-based multiple imputation levels. This was true even for the 5% and 10% NMAR data. Additionally, it was expected that the hotdeck and model-based estimates would perform worse for artificial MAR data where the missingness depended on marital status than for artificial MAR data where the missingness depended on gender; however, that was not the finding of this research. The reason for this result is not readily obvious, and this is a potential area of investigation for future researchers.

6.5 Conclusions, implications, and recommendations

It is important to remember the goals of imputation in the context of this research in order to decide what lessons to draw, where to make recommendations for BRFSS or other survey data users, and what potential avenues exist for future research. Imputation enables researchers to keep from disregarding non-missing data for a subject simply because some data is missing for that subject (Little & Rubin, 2002). Researchers can estimate the decrease in precision (wider confidence intervals) from estimates that they were able to make already based on non-missing data (Sulloway, 1996). Researchers can also potentially reduce any bias that would result from analyzing only the non-missing data (Horton & Kleinman, 2007).

When considering the general accuracy, all methods tested except for the BRFSS method performed relatively well. Reducing bias is most relevant to the question of accuracy for NMAR data. At the 20% NMAR level, the least accurate non-BRFSS imputation method, hotdeck, was different from the true proportion by an average of approximately 1.57%, compared to 1.93% per estimate for no imputation. While this result is not anything to dismiss, especially if estimating small proportions, the difference is small in absolute terms.

A cursory look at the results from this work shows that no imputation is as accurate as more sophisticated imputation methods in estimating proportions, but there is of course more to consider. One of the goals of imputation is to avoid disregarding all data for a subject just because some data is missing. If researchers are only estimating proportions, perhaps complete case analysis is a viable method. Imagine, on the other hand, that one wants to estimate the total population using data from a survey. Depending on the weights of the missing data, no imputation will be off by approximately the amount of missing data. At the 20% NMAR level described above, for example, no imputation would underestimate the total by much more than

1% or 2%. Or imagine investigating risk factors for some health outcome and finding that the total percentage of subjects missing information for at least one covariate is 20%. Rather than disregard that data using the complete case method, one can use multiple imputation to estimate the decrease in precision of estimates.

Estimating the decrease in precision (wider confidence intervals) of estimates is an important goal of imputation. In this work, for example, a multiple imputation method that took into account the between-imputation variance, the multiple hotdeck method, had a higher percentage of confidence intervals that contained the true value than the method that did not account for between imputation variance, the multiple model based method.

Ultimately, deciding how to handle missing survey data requires careful weighing of one's research goals. For estimating discrete demographic proportions at missing levels less than 10%, using no imputation may suffice, and one may not need to worry about determining what the missingness of MAR data depends on. Generally speaking, however, this work provides evidence to support using some kind of multiple imputation method. This work also offers evidence that confidence intervals from imputation estimates are too narrow, and that supports estimating confidence intervals using methods that account for between-imputation variance.

7 Appendix

7.1 Model-based imputation, complete-case, and BRFSS imputation confidence interval calculation method

Model-based imputation proportion estimate confidence intervals for both race and age were computed using the Stata command *mi estimate : svy : proportion*. Complete case and BRFSS imputation proportion estimate confidence intervals were computed using the Stata command *svy: proportion*. These commands use a first-order Taylor series approximation of the variance of the estimate, and then a t-distribution to determine the confidence intervals. The precise methodology used by the software to calculate confidence intervals is described in the Stata manual (Statacorp, 2011):

"Ratios and other functions of survey data

Shah (2004) points out a simple procedure for deriving the linearized variance for functions of survey data that are continuous functions of the sampling weights. Let θ be a (possibly vector-valued) function of the population data and $\hat{\theta}$ be its associated estimator based on survey data.

1. Define the jth observation of the score variable by

$$z_j = \frac{\partial \hat{\theta}}{\partial w_j}$$

   If $\hat{\theta}$ is implicitly defined through estimating equations, z_j can be computed by taking the partial derivative of the estimating equations with respect to w_j.

2. Define the weighted total of the score variable by

$$\hat{Z} = \sum_{j=1}^{m} w_j z_j$$

3. Estimate the variance $V(\hat{Z})$ by using the design-based variance estimator for the total $\hat{Z}$. This variance estimator is an approximation of $V(\hat{\theta})$.

Revisiting the total estimator

As a first example, we derive the variance of the total from a stratified single-stage design. Here you have $\hat{\theta} = \hat{Y}$, and deriving the score variable for $\hat{Y}$ results in the original values of the variable of interest.

$$z_j(\hat{\theta}) = z_j(\hat{Y}) = \frac{\partial \hat{Y}}{\partial w_j} = y_j$$

Thus you trivially recover the variance of the total given in (1) and (2).

The ratio estimator

The estimator for the population ratio is

$$\hat{R} = \frac{\hat{Y}}{\hat{X}}$$

and its score variable is

$$z_j(\hat{R}) = \frac{\partial \hat{R}}{\partial w_j} = \frac{y_j - \hat{R} x_j}{\hat{X}}$$

Plugging this into (1) or (2) results in a variance estimator that is algebraically equivalent to the variance estimator derived from directly applying the delta method (a first-order Taylor expansion with respect to y and x)

$$\widehat{V}(\widehat{R}) = \frac{1}{\widehat{X}^2}\{\widehat{V}(\widehat{Y}) - 2\widehat{R}\,\widehat{\mathrm{Cov}}(\widehat{Y},\widehat{X}) + \widehat{R}^2\,\widehat{V}(\widehat{X})\}$$

….Confidence intervals

In survey data analysis, the customary number of degrees of freedom attributed to a test statistic is $d = n-L$, where n is the number of *PSUs* and L is the number of strata. Under regularity conditions, an approximate 100(1−α)% confidence interval for a parameter θ (for example, θ could be a total, ratio, or regression coefficient) is

$$\widehat{\theta} \pm t_{1-\alpha/2,d}\{\widehat{V}(\widehat{\theta})\}^{1/2}$$

Cochran (1977, sec. 2.8) and Korn and Graubard (1990) give some theoretical justification for using $d = n-L$ to compute univariate confidence intervals and *p*-values. However, for some cases, inferences based on the customary $n-L$ degrees-of-freedom calculation may be excessively liberal; the resulting confidence intervals may have coverage rates substantially less than the nominal $1-\alpha$. This problem generally is of the greatest practical concern when the population of interest has a skewed or heavy-tailed distribution or is concentrated in a few *PSUs*. In some of these cases, the user may want to consider constructing confidence intervals based on alternative degrees-of-freedom terms, based on the Satterthwaite (1941, 1946) approximation and modifications thereof; see, for example, Cochran (1977, sec. 5.4) and Eltinge and Jang (1996).

Sometimes there is no information on n or L for datasets that contain replicate-weight variables but no *PSU* or strata variables. Each of *svy*'s replication commands has its own default behavior when the design degrees of freedom are not *svyset* or specified using the *dof()* option. *svy brr:* and *svy jackknife:* use $d = r-1$, where r is the number of replications. *svy bootstrap:* and *svy sdr:* use $z_{1-\alpha/2}$ for the critical value instead of $t_{1-\alpha/2,d}$."

7.2 Hotdeck imputation confidence interval calculation method

The Mander and Clayton multiple imputation hotdeck procedure used the method of Rubin and Little to calculate variances for confidence intervals. The method of Rubin and Little is to add the average of the within-imputation variance and the average of the between-imputation variance and then average the sum. Quoting directly from the Rubin and Little text:

"Let $\hat{\theta}_d$, W_d , $d = 1,\ldots,D$ be D complete-data estimates and their associated variances for an estimated parameter θ, calculated from D repeated imputations under one model. The combined estimate is

$$\bar{\theta}_D = \frac{1}{D}\sum_{d=1}^{D}\hat{\theta}_d$$

….The variability associated with this estimate has two components: the average within-imputation variance,

$$\bar{W}_D = \frac{1}{D}\sum_{d=1}^{D} W_d$$

and the between-imputation component.

$$B_D = \frac{1}{D-1}\sum_{d=1}^{D}(\hat{\theta}_d - \bar{\theta}_D)^2$$

….The total variability associated with $\bar{\theta}_D$ is

$$T_D = \bar{W}_D + \frac{D+1}{D} B_D$$

Where (1+1/D) is an adjustment for finite D."

The method then uses a t-distribution to approximate confidence intervals (Rubin & Little, 2002).

Table 36: Percent race values artificially missing weighted and unweighted at each level of missing and mechanism of missingness.

Missing and missing mechanism level	Survey Respondents with Missing Race/Ethnicity	NJ	NY	OR	WA	U.S. BRFSS
	Total Sampled	15,761	6,060	5,302	15,319	467,333
Originally Missing	Weighted Percent Missing	1.5%	2.6%	1.3%	1.3%	1.2%
	Percent Missing	1.8%	2.5%	1.5%	1.4%	1.3%
5% Artificial MCAR	Weighted Percent Missing	6.0%	6.9%	6.8%	6.2%	6.2%
	Percent Missing	6.4%	7.0%	6.8%	6.4%	6.0%
5% Artificial MAR given gender	Weighted Percent Missing	7.2%	8.7%	8.1%	7.5%	7.2%
	Percent Missing	6.7%	7.8%	6.8%	6.8%	6.3%
5% Artificial MAR given marital status	Weighted Percent Missing	6.4%	6.6%	6.2%	6.2%	5.8%
	Percent Missing	6.7%	6.4%	6.8%	6.4%	6.2%
5% Artificial NMAR	Weighted Percent Missing	5.5%	6.2%	6.4%	5.8%	5.3%
	Percent Missing	6.0%	7.4%	7.1%	6.6%	6.3%
10% Artificial MCAR	Weighted Percent Missing	11.2%	12.7%	12.4%	10.8%	11.0%
	Percent Missing	11.5%	12.6%	11.8%	11.3%	11.2%
10% Artificial MAR given gender	Weighted Percent Missing	13.1%	13.5%	13.0%	13.1%	12.8%
	Percent Missing	11.8%	12.2%	11.6%	11.8%	11.2%
10% Artificial MAR given marital status	Weighted Percent Missing	11.2%	10.7%	10.9%	11.1%	10.6%
	Percent Missing	11.5%	10.3%	11.9%	11.7%	11.1%
10% Artificial NMAR	Weighted Percent Missing	8.4%	9.5%	10.4%	10.7%	9.4%
	Percent Missing	10.2%	11.2%	11.3%	12.4%	11.3%
20% Artificial MCAR	Weighted Percent Missing	21.1%	23.0%	21.0%	21.5%	20.8%
	Percent Missing	21.5%	21.8%	21.1%	21.8%	21.0%
20% Artificial MAR given gender	Weighted Percent Missing	25.2%	25.3%	25.7%	25.8%	25.0%
	Percent Missing	21.5%	22.6%	21.3%	22.5%	21.2%
20% Artificial MAR given marital status	Weighted Percent Missing	20.4%	20.1%	20.2%	20.3%	19.9%
	Percent Missing	21.1%	29.8%	20.8%	22.0%	20.9%
20% Artificial NMAR	Weighted Percent Missing	16.8%	18.1%	21.6%	19.3%	17.8%
	Percent Missing	19.1%	22.0%	23.0%	22.5%	21.3%

Table 37: Percent age values artificially missing weighted and un-weighted at each level of missing and mechanism of missingness.

Missing and missing mechanism level	Survey Respondents with Missing Age Values	NJ	NY	OR	WA	U.S. BRFSS
	Total Sampled	15761	6060	5302	15319	467,333
Originally Missing	Weighted Percent Missing	1.1%	1.2%	1.0%	0.9%	0.7%
	Percent Missing	1.4%	1.6%	1.1%	1.1%	1.0%
5% Artificial MCAR	Weighted Percent Missing	5.5%	5.5%	6.4%	5.8%	5.5%
	Percent Missing	6.0%	6.0%	6.4%	6.1%	5.8%
5% Artificial MAR | gender	Weighted Percent Missing	7.4%	6.7%	7.0%	7.0%	6.7%
	Percent Missing	6.5%	6.6%	6.2%	6.4%	6.0%
5% Artificial MAR | marital status	Weighted Percent Missing	5.7%	5.5%	6.4%	6.0%	5.4%
	Percent Missing	6.2%	5.8%	6.5%	6.5%	5.9%
5% Artificial NMAR	Weighted Percent Missing	3.4%	3.3%	3.5%	3.5%	3.4%
	Percent Missing	5.7%	5.0%	5.8%	6.1%	5.9%
10% Artificial MCAR	Weighted Percent Missing	10.7%	10.7%	11.2%	11.2%	10.5%
	Percent Missing	11.0%	11.3%	11.4%	11.1%	10.8%
10% Artificial MAR |gender	Weighted Percent Missing	12.7%	13.0%	13.2%	13.0%	12.8%
	Percent Missing	11.3%	11.3%	11.3%	11.5%	11.0%
10% Artificial MAR |marital status	Weighted Percent Missing	10.6%	9.6%	11.1%	11.3%	10.0%
	Percent Missing	11.0%	10.0%	11.4%	11.8%	11.0%
10% Artificial NMAR	Weighted Percent Missing	7.5%	7.1%	6.9%	6.0%	6.3%
	Percent Missing	10.5%	10.4%	11.9%	10.7%	10.8%
20% Artificial MCAR	Weighted Percent Missing	21.7%	19.8%	20.2%	20.0%	20.5%
	Percent Missing	21.4%	20.1%	20.4%	20.5%	20.7%
20% Artificial MAR |gender	Weighted Percent Missing	25.3%	24.8%	25.0%	25.6%	24.8%
	Percent Missing	21.5%	21.5%	21.3%	22.0%	21.0%
20% Artificial MAR |marital status	Weighted Percent Missing	20.5%	18.4%	20.4%	19.9%	19.4%
	Percent Missing	20.7%	18.3%	21.2%	21.4%	20.6%
20% Artificial NMAR	Weighted Percent Missing	12.5%	12.4%	13.3%	11.7%	12.0%
	Percent Missing	19.0%	19.1%	23.0%	21.2%	20.7%

8. References

CDC, 2013 A. (July 15, 2013). *Behavioral Risk Factor Surveillance System Overview: BRFSS 2012*. Retrieved from http://www.cdc.gov/brfss/annual_data/2012/pdf/Overview_2012.pdf

CDC, 2013 B. (August 15, 2013). *BRFSS Data User Guide*. Retrieved from http://www.cdc.gov/brfss/data_documentation/PDF/UserguideJune2013.pdf

CDC, 2013 C. (July 15, 2013). *Behavioral Risk Factor Surveillance System 2012 Codebook Report Land-Line and Cell-Phone Data*. Retrieved from http://www.cdc.gov/brfss/annual_data/2012/pdf/CODEBOOK12_LLCP.pdf

CDC, 2012 A. (January 6, 2012). *Behavioral Risk Factor Surveillance System 2012 Questionnaire*. Retrieved from http://www.cdc.gov/brfss/questionnaires/pdf-ques/2012_BRFSS.pdf

CDC, 2012 B. (n.d.). *Weighting the Data*. Retrieved from http://www.cdc.gov/brfss/annual_data/2012/pdf/Weighting%20the%20Data_webpage%20content%2020130709.pdf

Clayton, D. & Mander, A. (September 3, 2007) B. *Hotdeck Imputation*. Retrieved from http://fmwww.bc.edu/repec/bocode/h/hotdeck.ado

Clayton, D. & Mander, A.. (January 18, 1999) A. Hotdeck: Stata module to impute missing values using the hotdeck method. [Computer software]. Chestnut Hill, MA: Boston College. Retrieved February 2014 from http://www.stata.com/stb/stb51

Cochran, W. G. 1977. *Sampling Techniques*. 3rd ed. New York: Wiley.

Eltinge, J. L., and D. S. Jang. 1996. Stability measures for variance component estimators under a stratified multistage design. *Survey Methodology* 22: 157-165.

Frankel, M. R., et al. (2012). When data are not missing at random: implications for measuring health conditions in the Behavioral Risk Factor Surveillance System. *British Medical Journal Open 2*(4). Retrieved from http://europepmc.org/articles/PMC3400062?pdf=render

Horton & Kleinman. (2007). Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*. *61*(1), 81-90.

Korn, E. L., and B. I. Graubard. 1990. Simultaneous testing of regression coefficients with complex survey data: Use of Boferroni *t* statistics. *American Statistician* 44: 270-276.

Lemeshow, S. & Levy, P. (2008). *Sampling of Populations Methods and Applications*. Hoboken, NJ: John Wiley & Sons.

Little, R. (1988). A test of missing completely at random for multivariate data with missing values. *American Statistical Association*. *83*(404), 1198-1202.

Little, R. & Rubin, D. (2002). *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley & Sons.

Park, M. A. (1999). *Biological Anthropology*. Mountain View, CA: Mayfield.

Satterthwaite, F. E. 1941. Synthesis of variance. *Psychometrika* 6: 309-316.

—. 1946. An approximate distribution of estimates of variance components. *Biometrics Bulletin* 2: 110-114.

Scheuren, F. (2005). Multiple imputation: how it began and continues. *The American Statistician 59*(4), 315-319.

Shah, B. V. 2004. Comment [on Demnati and Rao (2004)]. *Survey Methodology* 30: 29.

StataCorp. 2011. *Stata 12 Base Reference Manual*. College Station, TX: Stata Press.

Sulloway, F. (1996). *Born to Rebel*. New York: Vintage Books.

Town, M. (October 28, 2009). Weighting BRFSS Dual Frame Data. Retrieved from http://claude.com/PDF_Files/WeightingDual%20FrameRegional%20Training09.pdf