EVALUATION OF IMPUTATION METHODS FOR THE NATIONAL SURVEY ON DRUG USE AND HEALTH

Substance Abuse and Mental Health Services Administration
Center for Behavioral Health Statistics and Quality
Rockville, Maryland

April 2017

This page intentionally left blank

EVALUATION OF IMPUTATION METHODS FOR THE NATIONAL SURVEY ON DRUG USE AND HEALTH

RTI Authors:

Kimberly Ault
Peter Frechtel
Jeremy Aldworth
Kortnee Barnett-Walker
Kristen Brown
Lisa Carpenter
Chris Cummiskey
Lanting Dai
Susan Edwards
Celia Eicheldinger
Glynis Ewing
Jeff Laufenberg
Peilan Martin
Andrew Moore
Victoria Scott
Jean Wang

RTI Project Director:

David Hunter

SAMHSA Project Officer:

Peter Tice

SAMHSA Authors:

Rebecca Ahrnsbrak
Jonaki Bose
Joe Gfroerer
Sarra Hedden
Art Hughes

For questions about this report, please e-mail Peter.Tice@samhsa.hhs.gov.

Prepared for Substance Abuse and Mental Health Services Administration, Rockville, Maryland

Prepared by RTI International, Research Triangle Park, North Carolina

April 2017

Recommended Citation: Center for Behavioral Health Statistics and Quality. (2017). *Evaluation of Imputation Methods for the National Survey on Drug Use and Health*. Substance Abuse and Mental Health Services Administration, Rockville, MD.

Acknowledgments

This report would not be possible without the guidance and input of staff from the Substance Abuse and Mental Health Services Administration (SAMHSA), Center for Behavioral Health Statistics and Quality. In particular, Michael Jones (formerly with SAMHSA) provided useful comments. Special thanks are also due to several current and former RTI International (a registered trademark and a trade name of Research Triangle Institute) staff members. Heather Archambault (formerly with RTI), Stephen Black, Elizabeth Copello (formerly with RTI), and Bing Liu (formerly with RTI) assisted with SAS® program (SAS Institute Inc., 2002) submissions and quality control checks. James Chromy (formerly with RTI), Ralph Folsom (formerly with RTI), and Phillip Kott assisted with developing specifications for the imputation methods evaluated in this report, performing quality control checks on methodology, and reviewing the report. Amanda Lewis-Evans assisted with the flowcharts that appear in Chapter 5 (Figures 5.1 through 5.4) and Chapter 10 (Figure 10.1). James Chromy and Phillip Kott assisted with technical material in Appendix D. Kevin Adams, Neeraja Sathe, and Feng Yu assisted with the modifications to the report associated with limiting the risk of disclosure of personally identifiable information. Other contributors to this report at RTI include Claudia Clark, Valerie Garner, Roxanne Snaauw, Debbie Bond, and Teresa Bass.

Table of Contents

Table of Contents (continued)

Chapter **Page**

Table of Contents (continued)

Table of Contents (continued)

Chapter **Page**

Table of Contents (continued)

Chapter **Page**

Appendix

List of Tables

List of Tables (continued)

List of Tables (continued)

List of Tables (continued)

List of Tables (continued)

Table **Page**

List of Figures

Figure **Page**

This page intentionally left blank

1. Introduction

The National Survey on Drug Use and Health (NSDUH) provides national, state, and substate data on substance use and mental health in the civilian, noninstitutionalized population aged 12 or older. The overall purpose of imputing data in the NSDUH is to replace missing data with plausible values and to provide a completed data file for analysis purposes that contains the most accurate information possible about drug use among U.S. residents aged 12 or older. The redesign of the 1999 survey to computer-assisted interviewing (CAI) reduced the opportunities for respondents to have missing data or to report inconsistent answers, but CAI did not completely eliminate these problems. The procedure for editing data is referred to as the "flag and impute" procedure. Under these procedures, potential inconsistencies that exist between variables are identified and flagged, and inconsistent values are set to missing.[1] These inconsistencies are handled by statistically imputing final values with consistent data.

Since 1999, the predictive mean neighborhood (PMN) method has been used to impute missing values for many of the analytical variables in the NSDUH. Details on the PMN methodology and its application to the NSDUH, for which it was designed, can be found in Singh, Grau, and Folsom (2002) and in the 2011 imputation report of the NSDUH methodological resource book (MRB; Frechtel et al., 2013). Although PMN imputation as currently implemented has a number of advantages, including the ability to use a large number of similar variables to determine the imputed value and to provide individual record consistency among very complex variable relationships, the goal of this study was to evaluate this method compared with other options, especially in the context of the redesign of the NSDUH.

1.1 Goals of Evaluation

Although the use of CAI has reduced the occurrence of missing or inconsistent data in the NSDUH, the large number of variables that require imputation have imposed complex requirements to ensure that the imputation results maintain consistency across the variables. Consequently, the editing and imputation procedures for checking related variables for inconsistencies are time-consuming and costly. Given the large number of completed cases in each year's dataset and the relatively low levels of inconsistent and missing data,[2] more elegant and rigorous editing and imputation methods may not yield improvements over simpler methods that could be employed instead. This PMN imputation evaluation was developed to investigate the effects of alternate imputation methods on prevalence estimates and trends, relative to the current procedures. The main goals of this study were the following:

- Evaluate the general efficacy of PMN.
- Identify ways imputation for the NSDUH can be simplified while maintaining the quality of the resulting estimates.
- Identify any trade-offs associated with the alternative methods under consideration.

[1] The editing procedures for the 2007 NSDUH are summarized in Kroutil and Handley (2009) and Kroutil, Handley, Felts, Bradshaw, and Chien (2009).

[2] The 2011 MRB imputation report (Frechtel et al., 2013) provides a summary of the item nonresponse rates for the variables that are imputed on the NSDUH main study.

In evaluating the general efficacy of PMN, the advantages it provides over simpler hot-deck methods were quantified, and the question of whether those advantages are sufficient to justify its use was evaluated. To identify ways of simplifying the imputation process, this evaluation appraised the value of sequentially imputing drugs (i.e., using prior imputed variables as predictors for subsequent imputations) as currently performed in PMN. Also considered were alternative methods, including off-the-shelf software, that might produce estimates of comparable quality to those produced via PMN yet reduce staff burden and processing times relative to the current method. Finally, to quantify the variance inflation introduced when imputing missing values in the NSDUH, two multiple imputation (MI) methods were evaluated.

In addition, a few other aspects related to imputation on the NSDUH were evaluated for this study:

- testing alternative editing and imputation procedures for race and Hispanicity variables,
- determining the feasibility of using pair member data to assist with imputation,
- examining item nonresponse rates and testing alternative imputation methods for mental illness variables,
- examining item nonresponse rates for substance dependence and abuse variables, and
- examining item nonresponse rates and patterns for family income and related variables.

1.2 Overview of Report

Chapters 1 through 6 of this report focus on the feasibility of implementing the following four different imputation methods in the NSDUH and then presents a comparison of each of these methods with PMN:

- a simple approach for imputing drug variables using a weighted sequential hot-deck (WSHD) imputation method where a limited set of predictor variables was used, labeled simple WSHD for this report (Cox, 1980);
- a complex approach for imputing drug variables using WSHD where the set of predictor variables was expanded slightly over the simple approach, labeled complex WSHD for this report;
- an off-the-shelf software package that performs MI by using sequential regression models, called IVEware; and
- a modified version of PMN that can be described as a doubly protected MI method (Kott & Folsom, 2010) similar to IVEware, labeled modPMN-MI for this report.

The estimates presented are based on the 2007-2011 NSDUH data.[3] Chapter 1 presents the goals of the study and describes the NSDUH variables that were selected for evaluation including the percentage of imputed data based on PMN. Chapter 2 provides an overview of how PMN is currently implemented in the NSDUH and discusses key features of PMN that were used as comparison criteria for evaluating the alternative imputation methods that were considered.

[3] The analysis performed in this report using the 2007-2009 NSDUH data is based on the pre-March 2012 data files, and the analysis using the 2010-2011 data is based on the March 2012 revised data files.

Additionally, Chapter 2 includes results from an investigation into the use of fewer predictor variables for the response propensity models. Chapters 3, 4, and 5 discuss the steps for implementing each alternative imputation method evaluated in this report and identify the advantages and disadvantages of each method as compared with PMN. Chapter 3 explains the simple and complex WSHD methods. Appendix A presents the results of the Chi-square Automatic Interaction Detection (CHAID) analysis for the two WSHD methods and the model summaries for PMN, IVEware, and modPMN-MI. Chapter 4 discusses the implementation of IVEware, and Chapter 5 describes modPMN-MI. Appendix B contains two tables (one for IVEware and one for modPMN-MI) complementary to Chapters 4 and 5 that show the added variance, percentage imputed, relative increase in variance due to imputation, and other information related to MI. Associated with Chapter 5, Appendix C presents a discussion of alternative weighting procedures that compensate for item nonresponse in the estimation of predicted means that were implemented only in the modPMN-MI method. Chapters 3, 4, and 5 also discuss how each of the alternative methods might reduce costs and the time required for implementing them, relative to the costs and time associated with PMN. Chapter 6 summarizes the results of the comparisons, and Appendix D describes the methodology used for comparing estimates based on the four different imputation methods with those from PMN. Related to Chapter 6 are Appendices E and F, which include tables presenting the results of this evaluation. Appendix E presents before and after imputation distributions for demographic and drug variables across each imputation method. These tables help demonstrate the effects of imputation on the estimates and show how they differ by imputation method. Appendices F and G present the results of the comparisons of methods. Specifically, Appendix F displays the estimates for the demographic and drug variables by imputation method and the p-values from the statistical tests, and Appendix G illustrates the statistical testing results by summarizing the significant differences reflected in the pairwise comparisons of the imputation methods. Appendix H includes estimates of two-way drug comparisons by method.

Chapters 7 through 10 and the remaining appendices present examinations based loosely on the results of the comparison of PMN to the four alternative imputation methods discussed in Chapters 1 through 6. Chapter 7 describes possible new procedures for imputing race and Hispanicity and includes a literature review of how other national surveys impute these variables. Chapter 8 presents the results of an investigation of how to use pair member data to assist with imputation. Appendix I, complementary to Chapter 8, presents variable-specific feasibility assessments for using the other pair member's value in imputation. Chapter 9 discusses the imputation method used for mental health variables and presents results of using an alternative method of mean imputation. Appendix J presents item nonresponse rates for selected mental health variables and estimates of mental health by imputation class. Chapter 10 summarizes item nonresponse rates and patterns for substance dependence and abuse variables and presents the results of a sensitivity analysis using a simple imputation method. Appendix K presents item nonresponse rates and patterns of item nonresponse for substance dependence and abuse variables. Appendix L explores item nonresponse for family income in more detail, in part to follow up on the conclusion in Chapter 8 that these variables are good candidates for other-pair-member imputation. Finally, Chapter 11 presents conclusions and provides options for next steps that could lead to potential simplifications and improvements for imputing data in the NSDUH.

1.3 Selection of Variables for Evaluation

A critical component of this evaluation was to identify which variables would be selected for testing the different imputation procedures. Consequently, the focus was on two types of variables: (1) those that are mentioned frequently in reports, and (2) those that require a fair amount of imputation because of large amounts of missing data. It was determined that this evaluation should involve the core demographic and key drug variables. The data used in this report are edited based on rules implemented in the 2007 NSDUH. Baseline estimates presented in subsequent chapters are based on the 2007 NSDUH data that were imputed using the PMN imputation method. The 2007 NSDUH contained a total of 67,800 completed interviews.[4]

1.3.1 Demographic Variables

The core demographic variables were chosen because, along with being shown frequently in NSDUH tables, they are good predictor variables for the drug variables. The four core demographic variables imputed for this evaluation—marital status (IRMARIT), Hispanic/Latino origin (IRHOIND), race (IRRACE2), and education level (EDUCCAT2)—are shown in Table 1.1. Note that the race variable (EDRACEFINAL) selected for this study was an edited variable with four levels similar to the imputed race variable (IRRACE2).[5] Similarly, the education level variable (EDUCCAT2) selected was based on a multilevel imputed variable that was collapsed into five levels: (1) less than high school and aged 18 or older, (2) high school graduate and aged 18 or older, (3) some college and aged 18 or older, (4) college graduate and aged 18 or older, and (5) 12 to 17 years old. Both of these collapsed-level variables were selected because they are typically used in national reports and because categorical variables with many levels can be problematic when implementing some imputation methods.

Also shown in Table 1.1 is the unweighted frequency and weighted percentage imputed or logically assigned[6] with PMN for each demographic variable. The ANALWT variable is the analysis weight variable utilized to construct the weighted percentages reported, and it is the final person-level weight that is used in all analyses of the 2007 data presented in this report. In calculating the weighted percentages, only the subset of respondents for whom the variable is relevant or applicable[7] contributed to the sum of the weights included in the corresponding

[4] To be considered a completed case (or interview) for purposes of analysis, a respondent had to provide "yes" or "no" answers to the cigarette usage gate question and to at least 9 of the following additional drug usage gate questions: (1) chewing tobacco, (2) snuff, (3) cigars, (4) alcohol, (5) marijuana, (6) cocaine (in any form), (7) heroin, (8) hallucinogens, (9) inhalants, (10) pain relievers, (11) tranquilizers, (12) stimulants, and (13) sedatives.

[5] IRRACE2 was a four-level race variable with 1 = American Indian/Alaska Native, 2 = Asian/Other Pacific Islander, 3 = Black/African American, and 4 = White, where the multiple race respondents were assigned to one of these four categories based on models using the data from the 2000-2002 NSDUHs. However, racial demographics in the United States have changed since the 2002 survey, and more recent data needed to update these models are not available. As a result, IRRACE2 was no longer created after 2007. A five-level race variable with a separate multiple race category replaced IRRACE2 starting in 2008. For more details on this change, refer to Chapter 3 of the 2008 MRB imputation report (Ault et al., 2010).

[6] When values of nonmissing variables can be used to determine the value of the missing variable, that value was said to be "logically assigned" instead of "imputed."

[7] A "domain" is defined as the set of respondents who received a value other than a skip code for the imputed variable of interest. In other words, a domain is the subset of respondents for whom the variable of interest is relevant or applicable.

denominators. Similarly, the unweighted frequencies represent the number of imputed or logically assigned cases for only those respondents for whom the variable is relevant. The item nonresponse rates presented in this report include the logically assigned cases because originally a logically assigned value had a missing value. The number of logically assigned cases is typically very small, and the inclusion of these cases in the item nonresponse rate calculation does not change the conclusions in the report, because this proportion stays constant in the before and after imputation variable distributions.

For marital status, Hispanic/Latino origin, and education level, the percentages imputed or logically assigned are considerably lower (0.05 percent or less). For race, the percentage imputed or logically assigned is much higher (2.52 percent). Clearly the choice of imputation method affects the race variable considerably more than the other demographic variables.

Table 1.1 Demographic Variables Selected for Evaluation

Demographic Variable	Number of Imputed or Logically Assigned Cases	Weighted Percentage Imputed or Logically Assigned with PMN
Marital Status	18	0.03
Hispanic/Latino Origin	109	0.05
Race	1,860	2.52
Education Level	10	0.05

PMN = predictive mean neighborhood.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

1.3.2 Drug Variables

To help identify a subset of drug variables to select for the evaluation, a review of the percentage of missing data for each of the drug variables was performed. Additionally, summary statistics based on the complexity of the questionnaire skip patterns and the level of consistency between variables (later referred to as logical and likeness constraints for the PMN imputation method) were evaluated. Based on these two criteria, the drug variables selected for inclusion in this evaluation were cigarettes, alcohol, inhalants, marijuana, pain relievers, cocaine, and heroin. This set of drug variables contains variables with higher rates of imputation and variables with lower rates of imputation that are nonetheless important predictors of other drug use measures. In addition to exhibiting a wide range of percentages imputed, the selected drugs reflect the full range of complexity of variable consistency rules typically used in PMN. With the exception of the treatment of parent-child drug relationships,[8] the selected set of drug variables represents the many different types of situations that are encountered in the NSDUH.

For each of these drugs, all measures were imputed:

- lifetime usage,
- recency of drug use,

[8] The term "parent-child drug relationships" is used to refer to the reporting of drug classes and individual drug class members. As an example, crack is a form of cocaine. A respondent reporting use of crack should also report use of cocaine since crack is a form of cocaine.

- frequency of drug use (12-month and 30-day), and
- age at first use.

The recency variables have either four or five levels depending on the drug, where the levels indicate past month use, past year use, and nonuse. The 30-day frequency variables are estimates for days of use within the past month. The 12-month frequency variables are estimates for days of use within the past year.

For each of the drug variables included in this evaluation, the weighted percentage imputed or logically assigned via PMN is displayed in Table 1.2. On average, the weighted percentage of imputed data is less than 1 percent. The drug variables with the highest percentages imputed or logically assigned include alcohol and pain relievers. The Evaluation of Imputation Methods for the NSDUH Analytic Data Files Codebook provides the variable names, frequencies, and means for the variables presented in Tables 1.1 and 1.2.

Table 1.2 Drug Variables Selected for Evaluation

Drug Variable	Number of Imputed or Logically Assigned Cases	Weighted Percentage Imputed or Logically Assigned with PMN
Cigarettes Recency	469	0.27
Cigarettes 30-Day Frequency	211	0.20
Cigarettes Age at First Use	532	0.66
Alcohol Recency	879	0.90
Alcohol 12-Month Frequency	2,276	2.12
Alcohol 30-Day Frequency	875	1.06
Alcohol Age at First Use	802	1.21
Inhalants Recency	439	0.30
Inhalants 12-Month Frequency	312	0.16
Inhalants 30-Day Frequency	57	0.04
Inhalants Age at First Use	584	0.51
Marijuana Recency	428	0.39
Marijuana 12-Month Frequency	1,076	0.71
Marijuana 30-Day Frequency	218	0.16
Marijuana Age at First Use	255	0.28
Pain Relievers Recency	786	0.63
Pain Relievers 12-Month Frequency	713	0.51
Pain Relievers Age at First Use	896	0.99
Cocaine Recency	237	0.33
Cocaine 12-Month Frequency	331	0.36
Cocaine 30-Day Frequency	75	0.12
Cocaine Age at First Use	213	0.33
Heroin Recency	53	0.04
Heroin 12-Month Frequency	35	0.03
Heroin 30-Day Frequency	1	0.00
Heroin Age at First Use	44	0.02

PMN = predictive mean neighborhood.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Finally, to evaluate whether the summary drug measures (i.e., indicator variables that measure the use of any drug) change significantly among imputation methods, the following two summary measures[9] were examined:

- NEWSUMFLAG—the variable that indicates whether a respondent ever used one of the following drugs in his or her lifetime: marijuana, cocaine, inhalants, pain relievers, or heroin; and
- NEWSUMMON—the variable that indicates whether a respondent used one of the drugs listed above in the past month.

For each variable that includes imputed values, indicator variables distinguish imputed from nonimputed values where these indicators have the following three levels: (1) from questionnaire, (2) logically assigned, and (3) statistically imputed. The drug variables have an additional set of imputation indicators that provides a more detailed description of the sources of information used in the imputation of these variables. These detailed drug use imputation indicator variables provide information that is used in setting up the rules used for imputing drug recency and frequency-of-use variables. With these editing procedures, inconsistent responses in the recency-of-use variables were replaced by more general, consistent responses. Subsequently, the specific responses were then imputed. If the response from a recency variable was considered partially known, such as past year use of a given drug with the more specific recency unknown, imputed values had to be limited to what was consistent with this incomplete information. For example, the types of partial information available for the edited recency variable includes a level for the respondent using the drug at some point in the past 12 months (logically assigned value = 8) or using the drug at some point in his or her lifetime (logically assigned value = 9). The detailed imputation indicator variable for recency has the following levels: statistically imputed data with lifetime use imputed (value = 3), statistically imputed data with an edited recency of 9 (value = 4), and statistically imputed data with an edited recency value of 8 (value = 5). In some cases, the skip logic inherent in the questionnaire prevents a respondent from answering certain questions because of his or her responses to previous questions. For the drug use variables, if a respondent had been skipped out of a question, the response in the imputed variables is coded as "never used" or is coded to indicate that the respondent had not used drugs in the relevant time period. For other nondrug-related variables where this occurred, the imputed variable has a level with a label indicating a legitimate skip.

Table 1.3 presents a summary of the complexity of the questionnaire skip patterns or restrictions that are imposed in PMN on the imputed variable to maintain the individual record consistency for the variables selected for this evaluation. Within the hot-deck step of PMN, these restrictions are called constraints, where logical constraints prevent logical inconsistencies between variables and likeness constraints are flexible constraints that assist in finding a suitable match for the missing value. In order to maintain the same amount of internal consistency as provided by PMN, the logical constraints need to be applied to each variable selected for evaluation within each alternative imputation method. Because of the skip patterns and complex relationships between the variables, numerous logical constraints are used in the PMN hot-deck

[9] The two summary measures presented in this report differ from the SUMFLAG and SUMMON variables produced for the annual NSDUH national findings report because they do not include estimates for sedatives, tranquilizers, stimulants, and hallucinogens.

steps. The situation gets especially complex when "child" drugs are involved. For example, crack is a child drug associated with the parent drug cocaine. Logical constraints ensured that no post-imputation record had crack frequency greater than cocaine frequency, crack age at first use less than cocaine age at first use, and so on. However, these relationships between parent and child drugs were not evaluated in this study.

Table 1.3 Number of Logical and Likeness Constraints Implemented in PMN for Variables Selected for Evaluation

	Number of Logical Constraints	Number of Likeness Constraints
Demographic Variable		
Marital Status	0	2
Race	5	9
Hispanic/Latino Origin	0	2
Education Level	0	3
Drug Variable		
Cigarettes Recency and Frequency	8	2
Alcohol Recency and Frequency	26	3
Inhalants Recency and Frequency	26	3
Marijuana Recency and Frequency	26	3
Cocaine Recency and Frequency	40	3
Heroin Recency and Frequency	26	4
Pain Relievers Recency and Frequency	28	5
Cigarettes Age at First Use	9	8
Alcohol Age at First Use	9	8
Inhalants Age at First Use	9	8
Marijuana Age at First Use	9	8
Cocaine Age at First Use	15	8
Heroin Age at First Use	9	13
Pain Relievers Age at First Use	15	13

PMN = predictive mean neighborhood.

2. Predictive Mean Neighborhood Imputation

The predictive mean neighborhood (PMN) imputation method used in the National Survey on Drug Use and Health (NSDUH) is documented annually as part of the NSDUH methodological resource book (MRB). The material in this chapter reproduces some of that information for quick reference and to enable comparisons with the other imputation methods addressed in this evaluation and discussed in Chapters 3, 4, and 5. The PMN imputation method was not replicated for the reduced set of demographic and drug use variables addressed in this evaluation. Estimates based on PMN imputation presented in this report used imputed variables from the 2007 analytic data.

2.1 Overview of PMN Imputation

The PMN imputation method can be described as consisting of three key processes: (1) response propensity adjustment, (2) predictive mean modeling, and (3) hot-deck imputation. The first process, response propensity adjustment, involves adjusting the sampling design weights to make the item respondent sample representative of the entire NSDUH population. These adjusted weights are then used in the predictive mean models. Predicted means are obtained from the models for both item respondents and item nonrespondents, and the means of a particular outcome variable are modeled as a function of predictor variables (or covariates). The predicted means, along with other constraints, are used to define the neighborhoods from which donors are randomly selected for the final assignment of imputed values for the hot-deck imputation step. This assignment is done with either a single predicted mean (univariate matching) or several predicted means all at one time (multivariate matching). The selected donor may supply values to the item respondent for a single variable (univariate assignment) or for more than one variable (multivariate assignment). Sections 2.3 and 2.4 of the 2011 MRB imputation report (Frechtel et al., 2013) describe the steps for performing univariate and multivariate PMN.

Two types of restrictions or constraints are placed on the set of donors: logical constraints and likeness constraints. The constraints are implemented to make imputed values consistent with preexisting, nonmissing values of the item nonrespondents (recipients) and to make candidate donors as much like the recipients as possible. The logical constraints are fixed constraints that prevent logical inconsistencies between variables, and the likeness constraints are flexible constraints that govern the similarity between donors and recipients. Appendix C of the 2011 MRB imputation report presents the model summaries, and Appendix D of the 2011 MRB imputation report presents the hot-deck procedure summaries and the logical and likeness constraints associated with the variables requiring imputation (Frechtel et al., 2013).

In the NSDUH, there are several variables that are highly correlated with age including drug use, income, and health insurance. As a result, PMN is implemented within three or four age groups, and the models are developed separately within these age groups. Occasionally, the aggregation of age groups at the modeling stage is necessary because of a small number of applicable cases. In particular, the models for education level (highest grade completed) were fit within the two age groups of 12 to 17 and 18 or older; the models for employment status were fit

within two age groups of 15 to 25 and 26 or older; and the models for Hispanic/Latino origin, marital status, and immigrant age of entry were aggregated within all age groups.

The PMN method is a combination of two commonly used imputation methods: non-model-based nearest neighbor hot deck (NNHD; Little & Rubin, 1987, p. 65) and a modification of Rubin's model-assisted predictive mean matching (PMM; Rubin, 1986). The PMN method enhances Rubin's PMM method. Specifically, the PMN method can be applied to both discrete and continuous variables, either individually or jointly. The PMN method also enhances the NNHD method so that the distance function used to find neighbors is no longer ad hoc. In PMN, donors and recipients are distinguished by the completeness of their records with regard to the variable(s) of interest (i.e., the donor has complete data and the recipient does not). A donor set deemed "close" to that of the recipient, with respect to a number of predictors, is used to select a donor at random.

PMN is easily applied to both univariate and multivariate imputations. The terms univariate and multivariate matching and univariate and multivariate assignment are used to describe the implementation of PMN. Matching refers to the use of predicted means to match item nonrespondents with potential donors. The matching is univariate if only one predicted mean is used, and it is multivariate if more than one predicted mean is used. Assignment refers to the variables for which values are actually supplied by the donor to the item nonrespondent. The assignment is univariate if values for only one variable are supplied by the donor, and it is multivariate if values for more than one variable are supplied. Exceptions to this general framework are described in the separate sections on imputation of demographic and drug variables. In the multivariate case, PMN can be described as a sequential processing approach, in that each variable in the set is imputed in a particular sequence, where prior imputed variables are used to assist in subsequent imputations.

Six steps in the PMN imputation process are discussed below.

2.1.1 Step 1: Definition of Hierarchy

The first step is to determine the organization of variables requiring imputation into sets. Sets are formed based on the extent of correlation among variables and the level of missingness in the data. Variables with few missing values and no logical relationships with other variables tend to be imputed in a univariate manner. Variables with numerous missing values and logical relationships with other variables tend to be imputed in a multivariate manner.

Once the variable sets are formed, the next step is to determine the order in which variables in each set are modeled so that variables early in the hierarchy can be used for modeling the conditional predicted mean (i.e., they have the potential to be part of the set of predictors for variables later in the hierarchy). Note that usually not all variables in the hierarchy are missing for a particular incomplete record. The hierarchy is determined by considering such factors as the level of missingness in the data (see Appendix A of the 2011 MRB imputation report in Frechtel et al. [2013]) and the degree to which some variables could be used as predictors for others.

2.1.2 Step 2: Response Propensity Adjustment

For each set of variables to be imputed, two groups are created: complete data respondents and incomplete data respondents (item respondents and item nonrespondents, respectively). Complete data respondents have complete data across the variables of interest, and incomplete data respondents encompass the remaining respondents. In the multivariate case, complete data respondents must have complete data across all the variables in the response vector. Predictive mean models are constructed using complete data respondents only.

The weights used in the predictive mean models are adjusted for item nonresponse. The item response propensity model is a special case of the generalized exponential model (GEM),[10] which was developed for NSDUH weighting procedures. In some cases where more than one model was fit and the assignment and matching are multivariate, a shortcut approach is used where a single item response propensity model is applied to the whole set of variables of interest. Item respondents are those with complete data across all variables of interest. In other cases, there is an item response propensity model corresponding to each predictive mean model, but item respondents are still those with complete data across all variables in the response vector. Tables A.1 through A.7 in Appendix A present the response propensity model summaries for the demographic and marijuana usage variables.

2.1.3 Step 3: Predictive Mean Modeling

Each model is built using the complete data respondents only, with weights adjusted for item nonresponse. Logistic regression models are fit for categorical variables, and linear regression models are fit for continuous variables. Tables A.1 through A.7 present the predictors used in the predictive mean model for the demographic variables included in this evaluation and for marijuana drug use variables. The predictive mean models for all other drug usage variables included in this report are presented in Appendix C of the 2011 MRB imputation report (Frechtel et al., 2013).

2.1.4 Step 4: Defining Eligible Donor Sets Based on PMN, Logical Constraints, and Likeness Constraints

Once the model is fit, the predicted means for item respondents and item nonrespondents are calculated using the model coefficients. For each item nonrespondent, a distance is calculated between the predicted mean(s) of the item nonrespondent and the predicted mean(s) of every item respondent. Those item respondents whose predicted means are "close" (within a predetermined value labeled delta) to the item nonrespondent are considered part of the "delta neighborhood" for the item nonrespondent and are potential donors. If the number of item respondents who qualified as donors is greater than some number k, only those item respondents with the smallest k distances are eligible donors.

In practice, the delta is always 5 percent and k is always 30. The delta is always relative; that is, the predicted mean(s) of the item respondent must be within 5 percent of the predicted mean(s) of the item nonrespondent. If the predicted mean is a probability p, then the delta is

[10] The GEM macro, which was written in SAS/IML® software, was developed at RTI International for weighting procedures.

calculated as 5 percent of the minimum of (p, $1 - p$). This is done so that the delta is invariant to whether the probability of success or failure is considered. A looser delta was used for predicted probabilities close to 0.5, and a tighter delta was used for predicted probabilities close to 0 or 1. If the predicted mean was not a probability (e.g., for continuous variables), the delta was simply set to 5 percent.

The pool of donors is further restricted to satisfy both logical and likeness constraints to make imputed values consistent with the preexisting nonmissing values of the item nonrespondent. One example of a logical constraint is where crack age at first use must not be less than cocaine age at first use. Likeness constraints are placed on the pool of donors to make the attributes of the neighborhood as close to that of the recipient as possible. For example, for age at first use, the age of the donor and the age of the recipient are restricted to be the same whenever possible, and the donor and recipient must have come from states with similar usage patterns. A small value of delta also could be considered as a likeness constraint. Whenever insufficient donors are available to meet the likeness constraints, including the preset small value of delta, the constraints are loosened in priority order according to their perceived importance.

If many variables are imputed in a single multivariate assignment, it is advantageous to preserve, as much as possible, correlations between variables in the data. However, the more variables that are included in a multivariate set, the less likely it is that a neighborhood could be used for the imputation within a given delta. Even though there are many advantages to using multivariate matching, one disadvantage, in several instances, is not being able to find a neighborhood within the specified delta.

2.1.5 Step 5: Hot-Deck Imputation Assignment of Provisional Imputed Values

Using a simple random draw from the neighborhood developed in Step 4, a donor is chosen for each item nonrespondent. The missing value is simply replaced by the value of the donor. It is possible, however, that a donated quantity is a function of the final imputed value. For example, for 12-month frequency of drug use, because donors and recipients could potentially have a different maximum possible number of days in the year that they could have used a substance, the observed proportion of the total period is donated rather than the observed 12-month frequency, where the "total period" could range up to a year. In the assignment step, the donor's proportion of total period is multiplied by the recipient's maximum possible number of days in the year that he or she could have used the substance.

In the univariate case, the provisional imputed values are final and the imputation is complete. In the multivariate case, it is necessary to cycle through Steps 2 through 5 for each variable in the set, and then proceed to Step 6 after completing Steps 2 through 4 for the last variable in the set. Step 5 is not completed for the last variable in the set. The only purpose of provisional imputed values is to calculate predicted means for item nonrespondents. Because the last predictive mean model would have already been fit, provisional imputed values for the last variable would not be used in any way.

2.1.6 Step 6: Determination of Final PMN and Assignment of Final Imputed Values

After models are fit for all variables in the set, the neighborhood is defined based on the vector of predicted means. This vector may encompass a subvector of predicted means from a single categorical model (as with a polytomous logit model) in addition to scalar predicted means from any number of models with continuous response variables. For each item nonrespondent, a distance is calculated between the elements of this vector of predicted means where the observed values are missing and the corresponding elements of the vector for every item respondent.

A neighborhood resulting from this vector of distances is constrained by a multivariate preset delta, such that the distance associated with each element of the predictive mean vector has to be less than the preset delta associated with that element. From the donors that satisfied the multivariate delta condition, a single neighborhood is created by first converting the vector of differences into a scalar distance measure, called the Mahalanobis distance.[11] The Mahalanobis distance is used instead of Euclidean distance and is a form of standardizing the distance in terms of the population variances and covariances of vector components.

The Mahalanobis distance is calculated only for those respondents who met the multivariate delta constraint. The neighborhood is determined by selecting the k smallest Mahalanobis distances within this subset of item respondents for a given item nonrespondent. If the number of item respondents who met all constraints is fewer than k but greater than 0, all the item respondents in the resulting subset are selected for the neighborhood.

As with the univariate assignments, a donor is randomly drawn from the neighborhood for each item nonrespondent. For most variables, the observed value of interest is donated directly to the recipient. As in the univariate case, however, it is possible for a donated value to be a function of the final imputed value rather than the imputed value itself. The 12-month frequency example provided in Step 5 applies here as well.

2.2 Imputation of Demographic Variables

The next two sections more specifically address how the PMN method of imputation was applied during normal processing of the 2007 NSDUH data that were used in this evaluation.

Four demographic variables were selected for this evaluation: marital status, race, Hispanic/Latino origin, and education level. In this section, the application of the PMN method to these four variables is discussed in more detail. As part of standard processing, marital status was imputed first, followed by race, then Hispanic/Latino origin, and then education level. All four variables were categorical, each was treated as single set, and all were imputed in a univariate manner using logistic or polytomous logistic regression. Tables A.1 through A.4 show both the response propensity and predictive mean model summaries by age group for the demographic variables.

[11] See Section 2.3.3.3 of the 2011 MRB imputation report (Frechtel et al., 2013) for a definition of Mahalanobis distance. A definition also can be found in Manly (1986).

2.2.1 Marital Status

The application of PMN to marital status was straightforward. The marital status question in the NSDUH has four levels: married, widowed, divorced or separated, and never been married. A polytomous logistic regression model was fit using this four-level variable as the response variable. Respondents aged 12 to 14 were assigned a skip code and were not included in any imputation steps. Model parameters were estimated in combined models for all people aged 15 or older. No logical constraints were used in the hot-deck step. There were two likeness constraints: (1) each of the donor's predicted means was required to be within 5 percent of each of the recipient's predicted means, and (2) the donor was required to have an age within 3 years of the recipient's age. There is usually very little imputation required for this variable. For example, in the 2007 NSDUH, only 18 cases required imputation.

2.2.2 Race

Race, like marital status, involved a single polytomous regression model. However, five logical and nine likeness constraints were applied to limit the final donor set. Additional logical constraints were developed because the final imputation-revised race variables summarized information from several questions in the questionnaire, some of which have write-in responses. There was also a strong correlation between the race variables and the responses to questions on Hispanic/Latino origin and Hispanic/Latino group. Because the race variables were imputed before the Hispanic/Latino origin variables, the latter had missing values at the time the race variables were processed. Thus, the Hispanic/Latino origin variables were not used as predictors in the race models. Instead, likeness constraints were used to exploit the correlation.

The race variable examined in this evaluation has four levels: white, black/African American, American Indian/Alaska Native, and Asian/Other Pacific Islander. The item response rate for this variable is low relative to most other variables in the NSDUH—only 1,860 cases required imputation in 2007. The vast majority of the missing cases involved respondents who answered the Hispanic/Latino origin question affirmatively.

2.2.3 Hispanic/Latino Origin

The imputation of Hispanic/Latino origin is very simple. The outcome variable is dichotomous and maps to a single question in the NSDUH. No logical constraints were involved; two likeness constraints were applied. The item response rate tends to be high—only 109 cases required imputation in 2007. The correlation between race and Hispanic/Latino origin is exploited through the model; that is, the four-level imputation-revised race variable described in the preceding section is used as a predictor.

2.2.4 Education Level

The imputation of the education level variable is also fairly simple. The education variable (EDUCCAT2) has five levels: (1) less than high school and aged 18 or older, (2) high school graduate and aged 18 or older, (3) some college and aged 18 or older, (4) college graduate and aged 18 or older, and (5) 12 to 17 years old. The age group of 12 to 17 is processed separately from the age group of 18 or older, and because EDUCCAT2 has a skip code for respondents aged 12 to 17, no discussion of the processing of the age group of 12 to 17 is

relevant in this report. For the age group of 18 or older, the outcome variable used in logistic regression has four levels: less than high school, high school graduate, some college, and college graduate. These are the same levels of the variable EDUCCAT2, the one examined in this evaluation. No logical constraints and three likeness constraints were employed, and there were only 10 item nonrespondents in 2007.

2.3 Imputation of Drug Variables

Most of the drug variables involved in this evaluation were handled using multivariate PMN as part of the regular NSDUH imputation processing, and five drug measures were involved: lifetime use (yes/no), recency of use (three or four levels depending on the specific drug[12]), 12-month frequency of use (number of days the drug was used in the past 12 months), 30-day frequency of use (number of days the drug was used in the past 30 days), and age at first use. Various skip patterns were involved: those who report never using the drug in their lifetime skip all the other questions; those who do not report past year use in the recency question skip the 12-month frequency and 30-day frequency questions; and those who do not report past month use in the recency question skip the 30-day frequency question.

2.3.1 Ordering of Drugs and Measures and Grouping of Drug Variables into Sets

In the application of PMN to the NSDUH, the order of imputation for drugs was determined by considering such factors as the level of missingness in the data and the degree to which one set of drugs could be used as predictors for other drugs. The sequence of imputation for drug variables follows: cigarettes, smokeless tobacco, cigars, pipes, alcohol, inhalants, marijuana, hallucinogens, pain relievers, tranquilizers, stimulants, sedatives, cocaine, crack, and heroin. The order of drug use measures imputed was determined based on the natural hierarchy of the variables: lifetime usage, recency of use, 12-month frequency of use, 30-day frequency of use, and age at first use. Imputation-revised variables from earlier sets were usually used to inform the imputation of later sets as (1) predictor variables in the regression models, (2) variables that establish logical and likeness constraints in the hot-deck step, or (3) variables that inform skip patterns for eligible donors.

Table 2.1 shows the grouping of drug variables into imputation sets in PMN, as applied to the seven drugs used in this evaluation. Lifetime use for all drugs was treated as an initial set (drug set 1). The term "within drug" defines a dependency within drug measures (recency of use, frequency of use, and age at first use), and the term "across drug" defines a dependency across other drugs. For example, 12-month frequency has within-drug dependency on recency within all drugs, and marijuana recency has an across-drug dependency with cigarette and alcohol recencies because cigarette and alcohol recencies were used as predictor variables for imputing marijuana recency (Tables A.5 through A.7). Note also that some drugs do not have both 12-month and 30-day frequency variables, because the NSDUH questionnaire does not include both frequency questions for all drugs.

[12] There were four recency levels for cigarettes: past month use, past year but not past month use, past 3 years but not past year use, and lifetime but not past year use. There were three recency levels for all other drugs in this evaluation: past month use, past year but not past month use, and lifetime but not past year use.

Table 2.1 Grouping of Drug Variables into Imputation Sets in PMN

Drug Variable	Lifetime Use	Recency	12-Month Frequency	30-Day Frequency	Age at First Use
Cigarettes	Set 1	Set 2 (12-month frequency N/A)			Set 3
Alcohol		Set 4			Set 5
Inhalants		Set 6			Set 7
Marijuana		Set 8			Set 9
Pain Relievers		Set 10 (30-day frequency N/A)			Set 11
Cocaine		Set 12			Set 13
Heroin		Set 14			Set 15

N/A = not applicable; PMN = predictive mean neighborhood.

Steps 2 through 6 presented in Sections 2.1.2 through 2.1.6 are referred to throughout the following sections on the drug measures.

2.3.2 Lifetime Drug Use

For the lifetime drug use variables, Step 2 involved creating an item response indicator and fitting a shortcut response propensity model. A unit respondent was considered an item respondent if he or she gave valid responses to all lifetime drug use questions. A single response propensity model was fit for each of three age groups (12 to 17, 18 to 25, 26 or older). Most of the predictors were imputation-revised demographic variables such as age, gender, race, Hispanic/Latino origin, marital status, education level, and employment status.

Steps 3 through 5 were completed for each drug in turn. A dichotomous logistic regression model was fit for each drug-specific lifetime indicator, within the same three age groups. Only item respondents (with adjusted weights) were used to fit the models, but predicted values were calculated for both item respondents and item nonrespondents. Except for cigarettes, the first drug in the hierarchy, the set of predictors included lifetime indicators for all drugs earlier in the hierarchy. Where necessary, imputed values were used for these earlier lifetime indicators as created in Step 5. These "provisionally" imputed values were necessary for calculation of predicted means for item nonrespondents.

In Step 6, a final hot-deck step assigned a donor to each item nonrespondent. The donor was required to have predicted mean(s) close to the item nonrespondent's predicted mean(s) for all missing lifetime indicators in the response vector. The donor supplied values to the item nonrespondent for all missing lifetime indicators in the response vector. The item response rate was fairly high for all lifetime indicators, partly because survey respondents were considered unit nonrespondents unless they gave valid responses to most of the lifetime use questions.

2.3.3 Recency and Frequency of Drug Use

Because the questions for recency and frequency of drug use for the same drug are logically related, they were imputed in a multivariate manner. Many logical constraints were required to ensure that the post-imputation records were internally consistent (see Table 1.3). This approach was similar to the approach for lifetime drug use in many ways. The discussion in this section highlights the differences.

For the recency and frequency variables, in the first part of Step 2, an item response indicator was created in the same manner as for the lifetime use variables. A unit respondent was considered an item respondent if he or she gave valid responses to all recency and frequency questions for the given drug. However, unlike for lifetime use, separate response propensity models were fit for recency, 12-month frequency (if applicable), and 30-day frequency (if applicable). As for lifetime use, separate response propensity models were fit for each of three age groups (12 to 17, 18 to 25, 26 or older). The item response indicator was the same for each response propensity model.

Within drugs, the first measure to undergo response propensity modeling was recency. The domain of the model included all lifetime users; that is, the model reallocated the weights of the item nonrespondents whose imputation-revised lifetime drug use indicators were positive to the item respondents whose imputation-revised lifetime drug use indicators were positive. Note that respondents with valid values for recency but missing values for one or more frequency variables were considered item nonrespondents, and their weights were reallocated accordingly. After the recency response propensity model was fit, a polytomous regression model was fit using the adjusted weights, and predicted means for both item respondents and item nonrespondents were calculated. Note that these predicted means were conditional on lifetime use of the drug. Finally, a provisional hot-deck step was used to replace missing recency values with valid values where necessary.

For drugs with 12-month frequency variables, the next step was to fit an item response propensity model for 12-month frequency. The domain of this model included all unit respondents who were past year users according to the provisionally imputed recency variable. The provisionally imputed recency was a strong predictor in this model for every drug; that is, past month users tend to have higher 12-month frequency values than past year but not past month users. Once the weights were adjusted for item nonresponse, a linear regression model was fit, and predicted means for both item respondents and item nonrespondents were calculated. Note that these predicted means were conditional on past year use of the drug. Finally, a provisional hot-deck step was used to replace missing 12-month frequency values with valid values where necessary.

For drugs with 30-day frequency variables, the response propensity and predictive modeling steps were completed. Here, the domain included all past month users. The 12-month frequency value, if available, was a strong predictor for 30-day frequency; that is, respondents with large 12-month frequency values also tend to have large 30-day frequency values.

Logical and likeness constraints (see Table 1.3) were then applied to further reduce the set of eligible donors. The last step was a multivariate hot-deck step where a single donor was selected to provide values for recency, if missing, and either or both frequencies, if missing. Lifetime nonusers were automatically assigned skip codes for all variables.

2.3.4 Age at First Drug Use

The imputation procedure for age at first use was relatively straightforward. The imputation was univariate, and the domain included all lifetime users. The final imputed age at first use was bounded in the hot-deck step using logical constraints involving imputation-revised

recency and frequency values, the interview date, and the birth date. These constraints were often complex because of the requirement that an exact date of first use be assigned for each lifetime user of the drug. Likeness constraints were also applied (see Table 1.3).

2.4 Criteria for Comparing PMN with Alternative Methods

This section outlines key features of PMN that were used to compare and contrast the alternative imputation methods. Subsequent discussions of the alternative methods evaluated in this study follow the same structure in this section.

2.4.1 Methodological Steps for PMN

- As discussed before, the PMN method is comprised of three key processes: (1) response propensity adjustment, (2) predictive mean modeling, and (3) hot-deck imputation. During item response propensity modeling, the weights of the item respondents are adjusted to account for the weights of the item nonrespondents. Next, predicted means (or mean vectors) are used to identify a set of near neighbors. Logical and likeness constraints are applied to avoid inconsistencies and to account for additional relationships among variables. The third and final step, hot-deck imputation, randomly selects univariate or multivariate donors who are similar to the item nonrespondent and whose donated values are consistent with other responses.
- The response propensity step ensures that survey weights are adjusted for fitting each predictive mean model. Nearly every predictive mean model developed during imputation uses RTI's SUDAAN® software (RTI International, 2013), which properly accounts for the complex, multistage survey design used in the NSDUH. SUDAAN's more accurate estimates of variance for each parameter estimate ensure that the computation of predicted means is correct based on the survey design.

2.4.2 Complexity of Data Consistency and Order of Imputation for PMN

- PMN ensures consistency in post-imputation records via logical constraints in the hot-deck step. The logical constraints are especially complex for the drug variables, given the often numerous interrelationships across measures within the same drug. In order to maintain consistency, groups of variables are imputed in sets (i.e., multivariately). The procedure to group variables into sets is complicated and somewhat arbitrary, particularly for the drug variables. The set of predictor variables used in each model and the set of logical and likeness constraints involved in each hot-deck step must be developed prior to imputation. Finally, the order in which the likeness constraints are loosened must be decided upon for each hot-deck step. In PMN, it would be possible to reorder the variables to be imputed to accommodate changes in the questionnaire. However, the logical and likeness constraints would need to be redeveloped based on any new reordering.

2.4.3 Issues for Implementation of PMN

- PMN requires two sets of model-fitting exercises for each variable that requires imputation: one for the response propensity model and one for the predictive mean

model. For each variable, model-fitting diagnostics are examined to determine a robust model. This process requires manual intervention to examine model output and determine whether any predictor variables need to be removed from the model in accordance with their significance in the model. Because the models use numerous predictor variables, this diagnostic process is quite time-consuming and constitutes the majority of time spent on PMN implementation each year. One way to reduce the time spent on PMN implementation is to start with shorter lists of predictors. This idea is discussed in detail in the next section.

- In contrast to the model-fitting process, the hot-deck step of PMN does not take a significant amount of time to perform. However, the procedure cycles through each item nonrespondent one at a time and may take additional time to perform if numerous cases require imputation. Furthermore, on occasion, a donor cannot be found using even the least stringent set of likeness constraints. In such cases, an additional attempt to find a donor must be made by further weakening the set of likeness constraints applied. In very rare cases, an item nonrespondent has such an unusual response pattern that no donors meet even the logical constraints, and some sort of random imputation is usually needed. In these situations, additional manual intervention—and therefore additional implementation time—is required.

2.5 Investigation on Reduced Sets of Predictors

During each annual cycle of NSDUH imputation processing, more than 700 models are fit. When a model fails to converge or has some other problem, manual intervention is required where the statistician must decide which predictors (covariates) to remove from the model. Manual intervention requires the statistician to check outputs of the response propensity model program and the predictive mean model program to identify the covariates that cause these models to fail to converge based on a set of criteria.[13] Sometimes, many iterations of the intervention are needed to achieve convergence. These manual intervention steps add to the processing time and to project costs. In any given cycle, approximately one third of the models have required manual intervention.

The purpose of this section is to document an assessment of the feasibility of starting with shorter predictor lists so that fewer response propensity and predictive mean models require manual intervention, and those models that require manual intervention require less of it. An exercise like this was done in 2004, where reduced predictor lists were adopted for 18 of the predictive mean models for frequency of drug use. These reduced lists have been used every year since 2004. Frequency models often have small sample sizes because the domain is limited to past month or past year users of the given drug. These models still often require the removal of predictors, but much fewer predictors need to be removed. The slightly larger-than-needed starting predictor lists used for these models allow some flexibility from year to year, thus

[13] For the response propensity model, the criteria used to determine whether one or more variables should be removed are (1) variables with coefficient estimates equal to 25 or -25, and (2) variables with total nonresponse counts equal to 0. For the predictive mean model program, covariates with large Wald statistics are checked against a list of preferred covariates that are highly correlated with the response variable. However, these preferred covariates often cause singularities in the model. Whenever possible, these covariates are the last to be removed from the model after covariates with Wald p-values greater than 0.05 are removed.

allowing more adaptation to changes in the relationship between the outcome variable and the predictors.

This section of the report is a continuation of an overarching theme to see if the time and money required for NSDUH imputation can be reduced without having a negative impact on quality. It appears that it would be possible to use reduced sets of predictors because of the following observations:

- The experience with the drug frequency models since 2004 has been positive, and the reduction in processing time is substantial. It is reasonable to assume that about an hour of processing time is saved for each of the 18 predictive mean models that use reduced predictor lists. Thus, the reduction in processing time is likely about 15 to 20 hours for each cycle.
- Shortened predictor lists may hamper the ability to adapt to year-to-year changes because similar or identical predictors would be used each year. Nevertheless, the regression coefficients can still vary. When developing the shortened lists for the drug frequency models, the final list of predictors tended to be similar each year, suggesting that similar final predictor lists would result regardless of the starting list length. This was likely due to standardizations in the training of members of the RTI imputation team and in the quality control procedures.
- Under PMN, using fewer predictor variables in imputation models has proven to have a limited impact on quality and an unlikely effect on estimates. For most variables, the item response rate is high: more than 95 percent. This alone makes it unlikely that the estimates are affected by slight differences in imputation methods. Even when the imputation models are different, the same donor or a donor with the same outcome value has a chance of being selected in the hot-deck step.
- Under the modified predictive mean neighborhood multiple imputation (modPMN-MI) method (as described in Chapter 5), the connection between the model and the final imputation results is more direct, especially for categorical variables. Despite this, little impact on the quality of the resulting estimates is expected. The cost-related benefits of less manual intervention are likely to outweigh the impact on quality because of slightly less complicated models.

2.5.1 Methods

In PMN processing, two types of models are fit. The first type, the response propensity model, adjusts the weights for item nonresponse by allocating the weights of item nonrespondents to item respondents. These adjusted weights are then used in the second type of model, the prediction model, where some form of regression is used to obtain predicted means for both item respondents and item nonrespondents. These predicted means are then used to match item nonrespondents with item respondents in the hot-deck step.

For both types of models, it is desirable to use a large set of predictors in the starting list. The longer the starting list, the better the model. Predictors that work well for one cycle may not work well for another, and vice versa. A long starting list allows the statistician to choose the best predictors for that particular cycle. Even without manual intervention by the statistician, the imputation literature supports the idea of long predictor lists. In PMN, the only goal is to achieve

a good prediction. Variables that have predictive power should be included in the model. Model parsimony is not important because no inferences about causality are drawn directly from those models.

The disadvantage to using a large set of predictors is the long processing time. Manual intervention with models takes time, and with more than 700 models to fit, the schedule does not support long predictor lists for each model. The general approach currently taken is to treat the first model that converges as the final model, and then move on. Approximately two thirds of the time, the model requires no intervention using this approach. The choice of a starting predictor list is a balancing act between long lists that require manual intervention and short lists that may not lead to the best model.

In the remainder of this section, good candidate models for reduced sets of predictors (RSOPs) are identified by consulting the model summaries appendix of the 2007-2010 MRB imputation reports (Ault et al., 2009; 2010; 2011; Frechtel et al., 2012). The following criteria are important:

- **Number of Predictors Removed:** Models requiring the removal of numerous predictors tend to take longer to fit, so RSOPs will yield a significant reduction in time.
- **Critical Path:** In any given cycle, the demographic and drug variables are more on the critical path. That is, delays in the imputation processing of demographic and drug variables are most likely to cause delays in the completion of the detailed tables and the national findings reports; time savings are most likely to result in earlier completion of the detailed tables, annual reports, and datasets. RSOPs for any model will reduce the time spent on NSDUH imputation, but time savings for drug models have the greatest impact as they are most likely to affect the overall project schedule.
- **Low Missingness:** If the variable associated with the model has low missingness, then the quality of the model is unlikely to affect the estimates. RSOPs for these models will reduce costs without having a negative impact on quality.

For the models identified as good candidates for RSOPs, the *union* of the four final predictor lists from 2007 to 2010 would appear to be the most optimal for RSOPs. These unions of sets are easy to compile, and it is likely that the impact on the estimates of using such lists will be minimal. The use of these RSOPs might have led to exactly the same final lists that were used from 2007 to 2010. As stated above, the final predictor lists are often similar, even for models that require the removal of numerous predictors.

Only response propensity models are addressed here. This is because an assessment of the prediction models will be performed at a later date upon options resulting from an additional evaluation being performed. An assessment of the impact of disregarding certain SUDAAN warnings while fitting prediction models is ongoing. It is likely that, sometime in the future, at least some of these warnings will be disregarded.[14] That would probably result in final predictor lists with more variables than those seen in 2007 to 2010.

[14] For some variables in the 2015 NSDUH for which trends were disrupted, certain SUDAAN warnings were ignored.

The discussion of response propensity models in the next section is predicated on results from 2007 to 2010. If the SUDAAN warnings are disregarded in future years, the results from 2007 to 2010 are of limited utility in developing RSOPs for prediction models. Therefore, at least for the next few years, the use of the current sets of predictors appears to be better for all prediction models. Some of the prediction models for frequency of drug use already have RSOPs though (see Table 2.4), somewhat mitigating the problem.

2.5.2 Response Propensity Models

Of the 717 models that were fit each cycle, 290 were response propensity models. Of the 290 models, 108 (37 percent) required the removal of predictors. Table 2.2 shows the number of models and the number of models requiring predictor removal by variable group in 2010 processing. There is some variation across the variable groups, but most of the models were fit for the drug variables, and the percentage of drug response propensity models that required predictor removal was about the same as the overall percentage.

Table 2.2 Number of Response Propensity Models Requiring Predictor Removal in 2010, by Variable Group

Variable Group	Number of Response Propensity Models	Number of Response Propensity Models Requiring Removal of Predictors (with Percentage)
Demographics	16	2 (13%)
Drugs	180	67 (37%)
Income	8	4 (50%)
Health Insurance	15	6 (40%)
Roster	32	19 (59%)
Pair	39	10 (26%)
Total	290	108 (37%)

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

In the next sections, the variable groups are examined one at a time to determine good candidate response propensity models. Because the two demographic models requiring the removal of predictors in 2010 did not pose much of a problem, and, historically, few demographic models have required predictor removal, the discussion is limited to the other five variable groups.

2.5.2.1 Drug Variables

The drug variables can be conveniently divided into measures: lifetime, recency, 12-month frequency, 30-day frequency, and age at first use. The recency models can be further divided into model types: polytomous, past year (dichotomous), and past month (dichotomous). The recencies for more common drugs are modeled using polytomous regression, but for rare drugs, two dichotomous recency models are fit instead of a single polytomous one (more than two levels of the outcome variable). See Section 6.5.1.3 of Ault et al. (2010) for details. Table 2.3 reports on the number of models requiring predictor removal by measure.

Table 2.3 Number of Drug Response Propensity Models Requiring Predictor Removal in 2010, by Measure

Drug Measure	Domain	Number of Response Propensity Models	Number of Response Propensity Models Requiring Removal of Predictors (with Percentage)
Lifetime	All respondents	6	1 (17%)
Recency		72	27 (38%)
Polytomous	Lifetime users	30	6 (20%)
Past Year (Dichotomous)	Lifetime users	21	8 (38%)
Past Month (Dichotomous)	Past year users	21	12 (57%)
12-Month Frequency	Past year users	33	15 (45%)
30-Day Frequency	Past month users	30	17 (57%)
Age at First Use	Lifetime users	39	8 (21%)
Total		180	68 (38%)

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

Table 2.4 shows that models with small domains (e.g., past year inhalant users in the age group of 26 or older) require predictor removal more often. For past month recency models and 12-month frequency models, the domain consists of only past year users of the given drug. For 30-day frequency models, only past month users are in the domain. Models with small sample sizes tend to be more difficult to fit. It was mentioned in Section 2.5.1 that, beginning with 2004 processing, RSOPs were adopted for some of the models for frequency of drug use. These RSOPs are used for seven 12-month frequency prediction models and eleven 30-day frequency prediction models. RSOPs are not used for any response propensity models. Table 2.4 shows that all the domains for which RSOPs were used for frequency prediction models required predictor removal in response propensity models with the same domains. The number of predictor variables removed from response propensity models shown in Table 2.4 based on the 2010 NSDUH are representative of the number of variables removed from models in previous NSDUH years.

Table 2.4 Drug Frequency Prediction Models that Already Use RSOPs and History of Predictor Removal for Response Propensity Models with the Same Domain

Frequency Measure	Domain	Drug/Age Group	2007-2010 Predictor Removal History for Recency Response Propensity	2007-2010 Predictor Removal History for Frequency Response Propensity
12-Month	Past year users	Inhalants, 26+	Predictors removed in all years; 31 predictors removed in 2010	Predictors removed in all years; 27 predictors removed in 2010
		Hallucinogens, 26+	Predictors removed in all years; 25 predictors removed in 2010	Predictors removed in all years; 13 predictors removed in 2010
		Stimulants, 26+	Predictors removed in all years; 18 predictors removed in 2010	Predictors removed in all years; 22 predictors removed in 2010

Table 2.4 Drug Frequency Prediction Models that Already Use RSOPs and History of Predictor Removal for Response Propensity Models with the Same Domain (continued)

Frequency Measure	Domain	Drug/Age Group	2007-2010 Predictor Removal History for Recency Response Propensity	2007-2010 Predictor Removal History for Frequency Response Propensity
		Sedatives, 26+	Predictors removed in all years; no model used in 2010 due to no missing values	Predictors removed in all years; 45 predictors removed in 2010
		Heroin, 12-17	Predictors removed in all years; 25 predictors removed in 2010	Predictors removed in all years; 51 predictors removed in 2010
		Heroin, 18-25	Predictors removed in all years; 17 predictors removed in 2010	Predictors removed in all years; 52 predictors removed in 2010
		Heroin, 26+	Predictors removed in all years; 38 predictors removed in 2010	Predictors removed in all years; 17 predictors removed in 2010
30-Day	Past month users	Chewing Tobacco, 12-17	N/A	Predictors removed in 3 out of 4 years; 9 predictors removed in 2010
		Chewing Tobacco, 18-25	N/A	Predictors removed in 3 out of 4 years; 9 predictors removed in 2010
		Chewing Tobacco, 26+	N/A	Predictors removed in all years; 17 predictors removed in 2010
		Inhalants, 18-25	N/A	Predictors removed in all years; 20 predictors removed in 2010
		Inhalants, 26+	N/A	Predictors removed in all years; 35 predictors removed in 2010
		Hallucinogens, 26+	N/A	Predictors removed in all years; 30 predictors removed in 2010
		Cocaine, 12-17	N/A	Predictors removed in all years; 36 predictors removed in 2010
		Cocaine, 26+	N/A	Predictors removed in all years; 37 predictors removed in 2010

Table 2.4 Drug Frequency Prediction Models that Already Use RSOPs and History of Predictor Removal for Response Propensity Models with the Same Domain (continued)

Frequency Measure	Domain	Drug/Age Group	2007-2010 Predictor Removal History for Recency Response Propensity	2007-2010 Predictor Removal History for Frequency Response Propensity
		Heroin, 12-17	N/A	Predictors removed in all years; 52 predictors removed in 2010
		Heroin, 18-25	N/A	Predictors removed in all years; 50 predictors removed in 2010
		Heroin, 26+	N/A	Predictors removed in all years; 53 predictors removed in 2010

N/A = not applicable; RSOPs = reduced sets of predictors.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007-2010.

To summarize, since 2004, shortened predictor lists have been used for 18 frequency prediction models. Each frequency model has a different domain: past year or past month users of the given drug. There are 25 response propensity models associated with these 18 domains—7 recency response propensity models and 18 frequency models (see section 1.3.2 for recency and frequency definitions)—and practically all of them require predictor removal in almost every processing year. These 25 response propensity models appear to be good candidates for shortened predictor lists.

2.5.2.2 Income Variables

The income variable group response propensity modeling involves two phases: the binary variable phase and the finer categories phase. For each phase, imputations are done separately within four age groups (12 to 17, 18 to 25, 26 to 64, 65 or older). Within each age group, there is only one response propensity model for each phase, leading to a total of eight. The binary variable phase response propensity is a shortcut model—it covers item response for 11 variables.[15]

The four income response propensity models that required predictor removal were the four for the two older age groups. However, in 2010, none of these four models required the removal of more than 15 predictors. Because income is not on the critical path, the level of missingness is high, and the number of removed predictors is low, the benefits of using RSOPs for response propensity models for this variable group appear to be minimal.

[15] See Section 2.1.2 for a brief description of the shortcut response propensity model approach. This is the same approach as the one taken for lifetime drug use, as described in Section 2.3.2.

2.5.2.3 Health Insurance Variables

The health insurance variable group involves two methods: the "old method" and the "constituent variables" method. The old method has historically involved eight response propensity models: two for each of four age groups, the same four as used for the income variable group. However, processes were streamlined between the 2010 and 2011 processing cycles after the realization that the results from the first set of response propensity models were not used in later steps. What remained was only one set of four response propensity models. The constituent variables method involves seven response propensity models: again, two for each of four age groups, but two age groups are aggregated for the second one. The old method is still implemented for historical reasons; the more refined constituent variables method has been used only since the 2002 survey. Both methods are associated with the same set of questions, each of which has very low missingness. Usually, the health insurance variables are not on the critical path.

Table 2.5 shows the response propensity models for the old method, and Table 2.6 shows the response propensity models for the constituent variables method. For the old method, the first four response propensity models cover item response for both INSUR and PINSUR, and the last four response propensity models cover item response for both PINSUR and INSUR3.

Table 2.5 Response Propensity Models for Old Method Health Insurance

	Age Group			
	12-17	**18-25**	**26-64**	**65+**
INSUR3	Model O1	Model O2	Model O3	Model O4
PINSUR				

Table 2.6 Response Propensity Models for Constituent Variables Method Health Insurance

	Age Group			
	12-17	**18-25**	**26-64**	**65+**
CAIDCHIP	Model CV1	Model CV2	Model CV3	Model CV4
MEDICARE				
CHAMPUS				
PRVHLTIN				
ANYOTHER	Model CV5	Model CV6	Model CV7	

Of the four response propensity models for the (streamlined) old method, three required predictor removal in 2010. Of the seven response propensity models for the constituent variables method, only one required predictor removal in 2010, and only five predictors had to be removed for that one. None of the response propensity models from the constituent variables method seem to be potential candidates. Of the three old method models that required predictor removal, the one that required removal of the most predictors was the age group of 65 or older (Model O4), which required the removal of 16 predictors. The other two required removal of fewer than 10 predictors. Only Model O4 appears to be suitable as a candidate.

2.5.2.4 Roster Variables

The 19 roster response propensity models that required predictor removal in 2010 required the removal of an average of 20 predictors. Although the roster variable group is typically not on the critical path, extra time is still being spent on these models. Using reduced sets of predictors here would reduce project costs, but this would be unlikely to move up the delivery date of the detailed tables and other important deliverables.

Imputations are done for eight roster variables within the same four age groups, as was done for income (12 to 17, 18 to 25, 26 to 64, 65 or older). Four age groups within each of eight variables lead to 32 response propensity models, as shown in Table 2.7. The most time-consuming age group is 65 or older. There are only about 2,500 unit respondents in this age group each year, and the starting predictor lists are the same for this age group as for the age group of 26 to 64. All eight response propensity models for this age group required predictor removal for all years from 2007 to 2010. In 2010, an average of 28 predictors out of an average of 39 predictors was removed for the eight models. This age group also tends to have very high item response rates. In 2010, only 11 respondents underwent imputation across all eight variables. Because the item response rates are so high, and the number of predictors removed is so large, these eight models for the age group of 65 or older are good candidates for shortened predictor lists.

Table 2.7 Response Propensity Models for Roster

	Age Group			
	12-17	**18-25**	**26-64**	**65+**
TOTPEOP	Model R1	Model R2	Model R3	Model R4
KID17	Model R5	Model R6	Model R7	Model R8
HH65	Model R9	Model R10	Model R11	Model R12
FAMSKIP	Model R13	Model R14	Model R15	Model R16
FMLYSIZE	Model R17	Model R18	Model R19	Model R20
KIDFMLY	Model R21	Model R22	Model R23	Model R24
FAMSIZE	Model R25	Model R26	Model R27	Model R28
KIDFAMSZ	Model R29	Model R30	Model R31	Model R32

Restricting focus to the other three age groups, among the eight variables, the one that most often required predictor removal was the number of people in the household. The response propensity models for this variable (Models R1, R2, and R3 in Table 2.7) required predictor removal for all age groups and for all years from 2007 to 2010. This is the first variable modeled, and its imputation-revised value is used as an auxiliary variable for all the others, so it is perhaps a more important variable than the others. These three response propensity models seem like good candidates, but they are not as good as the eight models for the age group of 65 or older. The conclusion is that it makes most sense to target the eight models for 65 or older.

2.5.2.5 Pair Variables

Ten of the 39 pair models required predictor removal in 2010, and only 3 of these 10 required the removal of more than 10 predictors. For two of these three, no predictor removal

was required for some of the years from 2007 to 2009. The remaining model required the removal of more than 10 predictors in all years from 2007 to 2010. In 2010, another pair model did not require any predictor removal because the model was bypassed due to 100 percent response, but it did require the removal of more than 10 predictors for each of the years from 2007 to 2009.

The pair variable group is on the critical path for the development of pair weights, but the schedule is more flexible because the pair variables are not used to make national estimates for the annual NSDUH reports. The item response rates, at least for the two models mentioned above, are very high. In the first model mentioned above (Stage 1, Group 8), there were three nonrespondents in 2009 and eight in 2010. In the second model mentioned above (Stage 1, Group 7), there were only two nonrespondents in 2009 and none in 2010.[16] Given the low missingness and the number of predictors that need to be removed for these models, these two are recommended as candidates for shortened predictor lists.

This discussion presented candidate models for shortened lists of starting predictors, or RSOPs. However, this analysis was only done for response propensity models because a likely procedural change involving "warning" messages in SUDAAN makes an investigation of prediction models premature at this time. For each response propensity model identified as a candidate for an RSOP, the union of the final predictor lists from 2007 to 2010 is recommended as the actual RSOP. This approach is expected to save time with minimal impact on the quality of the imputation results.

2.6 Summary and Options

This chapter outlined the PMN imputation method for demographic and drug variables and defined criteria that were used to compare alternative imputation methods. An investigation into imputation model-fitting procedures was completed to examine the importance of variables used in response propensity models. The PMN imputation method was reviewed, and potential simplifications were identified for further investigation to determine if they would result in cost or processing time reductions without an impact on data quality. These simplifications include the following:

- If a variable has a low item nonresponse rate (i.e., less than 1 percent), then only perform the hot-deck imputation portion of PMN using logical constraints (i.e., dropping the likeness and delta constraints).
- Similar to the investigation for the response propensity models, simplify the predictive mean models by using a reduced set of predictor variables that normally result in convergent models and could remain static over time.

These simplifications can be performed in different combinations, and some potential optimal combinations are summarized below. After examining all variables requiring imputation, selection criteria would be developed to determine which modification should be applied to each set of variables.

[16] See Section 10.5.2 in Frechtel et al. (2013) for more information on these two models.

- **Modification 1: Response Propensity Adjustment and Hot-Deck Imputation.** Impute variables with few logical constraints and variables with low levels of missingness using the response propensity adjustment and hot-deck imputation.
- **Modification 2: Response Propensity Adjustment and Predictive Mean Modeling.** Impute binary and categorical variables using the response propensity adjustment and the model probabilities. This modification was tested in the modPMN-MI method discussed in Chapter 5.
- **Modification 3: Response Propensity Adjustment and Simplified Predictive Mean Modeling.** Impute variables using PMN but use a simplified predictive mean model that contains a reduced set of predictors that remain static for each variable being imputed. The idea behind this modification was tested with the implementation of the weighted sequential hot-deck methods that are described in Chapter 3.

This page intentionally left blank

3. Weighted Sequential Hot-Deck Imputation

This chapter discusses the steps for implementing the weighted sequential hot-deck (WSHD) imputation for the selected demographic and drug variables. The WSHD method was chosen to identify ways to simplify the imputation process in the National Survey on Drug Use and Health (NSDUH) by considering the following: (1) using fewer predictor variables, (2) using only logical constraints as defined in the predictive mean neighborhood (PMN) method (i.e., no likeness constraints were used), and (3) using the SUDAAN® HOTDECK procedure (RTI International, 2013). Additionally, the WSHD method was used to assess whether imputing the drug variables in a particular sequence changed the final imputed estimates.

3.1 Overview of WSHD Imputation

Sequential hot-deck imputation is a common method used for item nonresponse. This method uses the respondent survey data (donors) to provide imputed values for records with missing values by defining imputation classes, which generally consist of a cross-classification of covariates that are related to the variable needing imputation and then replacement of missing values sequentially within the imputation classes. In a sequential hot-deck procedure, data are sorted using specific criteria within each imputation class. Each time an item respondent is encountered, the variable response is stored and then used as the donor value for subsequent nonrespondents, which results in a statistically imputed response. Because the data are sorted by relevant auxiliary variables, the item respondent (donor) closely matches the item nonrespondent (recipient) with respect to the auxiliary variables.

When sequential hot-deck imputation is performed using the sampling weights of the item respondents and nonrespondents, the method is called *weighted* sequential hot-deck imputation. This method takes into account the unequal probabilities of selection in the original sample by using the sampling weight to specify the expected number of times a particular respondent's answer is used to replace a missing item. Selection frequencies are specified so that, over repeated applications of the algorithm, the expected value of the weighted distribution of the imputed values will equal the weighted distribution of the reported answers within imputation class. An advantage of WSHD imputation is that it controls the number of times a respondent record can be used for imputation and gives each respondent record a chance to be selected for use as a hot-deck donor.

Release 11.0.1 of SUDAAN includes a new procedure called PROC IMPUTE for the WSHD method. This procedure allows for multivariate (several variables imputed at the same time) and multiple (several imputed versions of the same variable) imputations. As part of this study, the PROC IMPUTE procedure's ability to do multivariate imputations was evaluated. However, its ability to perform multiple imputations was not tested.

There are three key processes for implementing WSHD: (1) forming imputation classes, (2) sorting the data file, and (3) assigning imputed values for missing values. The following sections describe these processes in detail.

3.1.1 Using Chi-square Automatic Interaction Detection Analysis for Imputation Class Development

To implement WSHD, the first step is developing imputation classes (or donor sets) for each variable requiring imputation. Imputation classes are formed by a cross-classification of variables that are correlated to the variable needing imputation. Within each imputation class, imputation procedures are implemented independently. The selected predictor variables must have a strong association with the variable requiring imputation. To perform cross-classification of variables, a Chi-square Automatic Interaction Detection (CHAID) analysis is typically used, as described in Kass (1980). The CHAID analysis is a decision tree method that first determines the optimal partitioning of the data that maximizes the between-group variance for the variable being imputed (or dependent variable). The CHAID analysis divides the data into groups based on the most significant predictor variable of the item being imputed (target or dependent variable). Subsequently, this procedure is repeated using the remaining predictor variables to split each of the emerging groups into smaller subgroups. This continues until stopping rules, based on group sizes or variance reduction thresholds, cause the process to terminate. In this process, a number of subgroups created during a previous iteration might be merged back to form new subgroups. This splitting and merging process continues until no more statistically significant predictors are found, at which point imputation classes are defined from the resulting segments. The CHAID analysis is performed as follows:

- **Step 1:** For each predictor variable, find the categories that are least significantly different (the largest p-value), with respect to the target (or dependent) variable. If the target variable is continuous, the p-value is based on the F-statistic. If the target variable is nominal, the p-value is based on the Pearson chi-square test.
- **Step 2:** For the categories of the predictor variable with the largest p-values, compare the p-value with a specified significance level, α_{merge}. If the p-value is greater than α_{merge}, merge this pair into a single category. As a result, a new set of categories of the predictor variable is formed, and the algorithm reexamines this predictor, proceeding from Step 1 again. If the p-value is less than α_{merge}, then proceed to Step 3.
- **Step 3:** Select the predictor variable with the most significant p-value. Compare this value with a specified split level, α_{split}. If the p-value is less than or equal to α_{split}, then split the group based upon the set of categories of the predictor variable. If the p-value is greater than α_{split}, then this is a final group and no more splitting will occur. The values of α_{split} and α_{merge} were both set at 0.05 (or 5 percent), resulting in a 5 percent significance level of the splits and the merged categories.
- **Step 4:** Continue the tree growing process until all stopping rules have been met.

The CHAID analysis was performed by using a SAS® macro that used the SAS Enterprise Miner 4.3 Decision Tree Node application (SAS Institute Inc., 2002).[17] When performing CHAID analysis, the key components include setting the merging and splitting significance criteria and tree growth limits (i.e., number of branches in the decision tree). Table 3.1 presents a brief description of the Decision Tree Node options and the starting criteria

[17] The CHAID analysis can be performed with other statistical software packages such as Statistical Package for the Social Sciences (SPSS) Answer Tree and Classification and Regression Tree (CART).

settings that were used for developing the imputation classes. The Decision Tree Node options for performing CHAID analysis are described in detail in SAS documentation (SAS Institute Inc., 2002).

Table 3.1 Starting Criteria for Decision Tree Node Options for CHAID Analysis

Decision Tree Node Option	Description	Starting Criteria
Significance Level	Threshold p-value for the splitting criterion	0.05
Leaf Size	Minimum number of observations necessary before a split can occur	50
Split Size	Minimum number of observations required for a split	100
Maximum Depth	Maximum number of groups that can be generated (tree pruning)	6
Subtree	Construction of a subtree where the smallest subtree with the best assessment value is chosen	Assessment

CHAID = Chi-square Automatic Interaction Detection.

3.1.2 Sorting the File

Within each imputation class, the data were sorted by auxiliary variables relevant to the variable being imputed. The sort order of the auxiliary variables was chosen to reflect the degree of importance of their relation to the variable being imputed (i.e., those variables that were better predictors for the variable being imputed were used as the first sorting variables).

3.1.3 Assigning Imputed Values

Once the imputation classes were formed and the data were sorted, the data were divided into two datasets: one for respondents and one for nonrespondents. Scaled weights v_j were then derived for all nonrespondents using the following formula:

$$v_j = w_j s_+ / w_+; j = 1,2,\ldots, n,$$

where n is the number of nonrespondents, w_j is the sample weight for the j^{th} nonrespondent, w_+ is the sum of the sample weights for all the nonrespondents, and s_+ is the sum of the sample weights for all the respondents (Cox, 1980). The respondent data file is partitioned into zones of width v_j, where the imputed value for the j^{th} nonrespondent is selected from a respondent in the corresponding zone of the respondent data file.

3.2 Imputation of Demographic Variables

Demographic variables were imputed univariately[18] in a sequential order where prior imputed variables were used as predictor variables for subsequent imputations. The order in which the demographic variables were imputed was in the same order as in the standard main

[18] "Univariate imputation" is defined as imputing one variable at a time. For hot-deck methods, the imputation is univariate if the donor supplies values to the recipient for only one variable.

study processing: marital status, race, Hispanic/Latino origin, and education level. The steps below summarize this approach for imputing the demographic variables.

- **Step 1:** Perform a CHAID analysis for marital status using a starting list of predictor variables, and then impute marital status.
- **Step 2:** Perform a CHAID analysis for race using a starting list of predictor variables plus imputed marital status as a predictor, and then impute race.
- **Step 3:** Perform a CHAID analysis for Hispanic/Latino origin using a starting list of predictor variables plus imputed marital status and race as predictors, and then impute Hispanic/Latino origin.
- **Step 4:** Perform a CHAID analysis for education level using a starting list of predictor variables plus imputed marital status, race, and Hispanic/Latino origin as predictors, and then impute education level.

The starting list of predictor variables for the demographic variables is presented in Tables A.8 through A.12 in Appendix A. This starting list is similar to the predictor variables used when performing the response propensity adjustment and the predictive mean modeling process in PMN.

When performing the CHAID analysis for the demographic variables, none of the starting criteria options (Table 3.1) needed to be modified. For each demographic variable, the cases were first sorted by imputation class and then by age. For the marital status variable, imputations were conducted separately within each of three age groups (12 to 17, 18 to 25, and 26 or older), though only a single CHAID analysis was performed for all age groups because all respondents younger than 15 years were assigned a code of "not applicable." Similarly, the imputations for education level were conducted separately within the three age groups, and a single CHAID analysis was performed for all age groups because the education level variable did not distinguish education levels for those younger than 18 years. The CHAID analyses were performed within each of the three age groups listed above for the race variable and only for those aged 12 or older for Hispanic/Latino origin, similar to PMN. Also shown in Tables A.8 through A.12 are the final sets of predictor variables that form the imputation classes resulting from the CHAID analyses. In subsequent chapters that discuss comparison of estimates between imputation methods, the imputation of these demographic variables is referred to as *simple* WSHD.

3.3 Imputation of Drug Variables

To address the research question of whether a simpler imputation procedure could be implemented for the drug variables that would allow for cost and time savings but not degrade the quality of the national estimates, two options were developed using the WSHD imputation method: *simple* WSHD and *complex* WSHD.

3.3.1 Simple WSHD

The first option, simple WSHD, examined (1) whether using only a small set of predictor variables (e.g., demographics and one key lifetime drug use variable—cigarettes use) would be sufficient for imputing all drug variables, and (2) whether imputing the drug variables in a

particular sequence would make a difference. The first approach for developing imputation classes for the drug variables was labeled as simple WSHD because the number of predictor variables selected for the CHAID analysis was limited. Furthermore, this approach would allow for simultaneous processing for imputing drug variables because there are fewer dependencies between the variable being imputed and predictor variables (i.e., the variables being imputed are not subsequently used as predictor variables for other variables being imputed). For this option, only imputed demographic variables and cigarettes lifetime use were selected as predictor variables for the CHAID analysis. Each CHAID analysis was performed within the three age groups that are typically used in PMN. However, because of smaller domain sizes for inhalants, pain relievers, cocaine, and heroin, it was necessary to combine all age groups for the CHAID analysis. The steps below summarize this approach for imputing the drug variables.

- **Step 1:** Perform a CHAID analysis for each lifetime drug use variable (cigarettes, alcohol, inhalants, marijuana, pain relievers, cocaine, and heroin) and develop imputation classes using imputed demographic and cigarettes lifetime use variables as predictors.[19]
- **Step 2:** Impute each lifetime drug use variable (with the exception of cigarettes) univariately.
- **Step 3:** For each drug, impute multivariately[20] the recency and frequency variables by utilizing the imputation classes based on the CHAID analysis for each lifetime drug use variable (from Step 1).
- **Step 4:** For each drug, impute the age-at-first-use variable univariately, again using the imputation classes from the lifetime drug use variables (from Step 1).

For each of these steps, the data were divided into groups based on eligible donor sets (or missingness patterns; see Table 3.5). Section 3.4.2 describes the differences and similarities between the simple and complex WSHD method and the PMN method.

3.3.2 Complex WSHD

After developing the simple WSHD procedure, several options were explored that would allow for a more complex approach to be adopted that would assess whether additional variables were needed as predictors, yet would still allow for independent processing of sets of drug variables (lifetime, recency, frequency, and age at first use). The addition of predictor variables is assessed by whether there are differences in the estimates based on using a small (simple) set or a larger (complex) set. Chapter 6 examines the differences in the estimates. Consequently, a second option, complex WSHD, was developed that expanded the list of predictor variables for each drug to include imputed demographic variables, lifetime use of any and all drugs, and the respective drug use measures. For example, cigarettes recency of use would be imputed using imputed demographic variables and all imputed lifetime drug use variables as predictor variables. After the imputation of cigarettes recency is completed, cigarettes 30-day frequency would be imputed using imputed demographic variables, all imputed lifetime drug use variables,

[19] Since imputation classes were needed for cigarette recency and frequency variables, a CHAID analysis was performed for cigarettes lifetime use. For the cigarettes lifetime CHAID model, only the imputed demographic variables were used as predictors.

[20] "Multivariate imputation" is defined as imputing more than one variable at a time. For hot-deck methods, the imputation is multivariate if the donor supplies values to the recipient for more than one variable.

and imputed cigarettes recency as predictor variables. Finally, cigarettes age at first use would be imputed using imputed demographic variables, all imputed lifetime drug use variables, imputed cigarettes recency, and imputed cigarettes 30-day frequency as predictor variables. This sequence was completed for each drug and its respective variables.

Complex WSHD starts by performing a CHAID analysis for each lifetime drug use variable and uses imputed demographic and other lifetime drug use variables as predictors. Next, a CHAID analysis is performed for each set of drug variables using the same hierarchy (as discussed in Chapter 2) that is followed in PMN. This sequence is completed for each drug and its respective variables. The steps below summarize complex WSHD.

- **Step 1:** Perform CHAID analysis for alcohol lifetime using cigarettes lifetime and imputed demographics as predictors, and then impute alcohol lifetime use univariately.
- **Step 2:** Perform CHAID analysis for inhalants lifetime using cigarettes and alcohol lifetime and imputed demographics as predictors, and then impute inhalants lifetime use univariately.
- **Step 3:** Perform CHAID analysis for marijuana lifetime using cigarettes, alcohol, and inhalants lifetime and imputed demographics as predictors, and then impute marijuana lifetime use univariately.
- **Step 4:** Perform CHAID analysis for pain relievers lifetime using cigarettes, alcohol, inhalants, and marijuana lifetime and imputed demographics as predictors, and then impute pain relievers lifetime use univariately.
- **Step 5:** Perform CHAID analysis for cocaine lifetime using cigarettes, alcohol, inhalants, marijuana, and pain relievers lifetime and imputed demographics as predictors, and then impute cocaine lifetime use univariately.
- **Step 6:** Perform CHAID analysis for heroin lifetime using cigarettes, alcohol, inhalants, marijuana, pain relievers, and cocaine lifetime and imputed demographics as predictors, and then impute heroin lifetime use univariately.

Steps 7 through 10 are applied to each drug set.

- **Step 7:** Perform CHAID analysis for recency using imputed demographic variables and all imputed lifetime drug use variables as predictors, and then impute recency univariately.
- **Step 8:** Perform CHAID analysis for 12-month frequency (where necessary) using imputed demographic variables, all imputed lifetime drug use variables, and imputed drug-specific recency as predictors, and then impute 12-month frequency univariately.
- **Step 9:** Perform CHAID analysis for 30-day frequency (where necessary) using imputed demographic variables, all imputed lifetime drug use variables, and imputed drug-specific recency and 12-month frequency as predictors, and then impute 30-day frequency univariately.
- **Step 10:** Perform CHAID analysis for age at first use using imputed demographic variables, all imputed lifetime drug use variables, and imputed drug-specific recency, 12-month frequency, and 30-day frequency as predictors, and then impute age at first use univariately.

Table 3.2 shows the grouping of drug variables by the steps described above. Steps 1 to 6 are labeled as "Set 1" where each lifetime use variable is imputed in the sequence shown in the table. Steps 7 to 10 are labeled as "Set 2" to "Set 7" because each drug measure is imputed within each set of drugs.

Table 3.2 Grouping of Drug Variables for Complex WSHD

Drug Variable	Lifetime Use	Recency	12-Month Frequency	30-Day Frequency	Age at First Use
Cigarettes	N/A	(Set 2) Steps 7-10 (12-Month Frequency N/A)			
Alcohol	(Set 1) Steps 1-6	(Set 3) Steps 7-10			
Inhalants		(Set 4) Steps 7-10			
Marijuana		(Set 5) Steps 7-10			
Pain Relievers		(Set 6) Steps 7-10 (30-Day Frequency N/A)			
Cocaine		(Set 7) Steps 7-10			
Heroin		(Set 8) Steps 7-10			

N/A = not applicable; WSHD = weighted sequential hot deck.

Complex WSHD can be described as (1) a sequential process within each set of drug variables because there is a dependency among drug measures (recency of use, frequency of use, and age at first use) where each measure is used as a predictor variable for imputation of subsequent measures, and (2) a simultaneous process across drugs because no additional drug sets were used as predictor variables for each drug set (except for lifetime use). Because complex WSHD performs a cyclic process of CHAID analysis and imputation, it is a sequential process similar to PMN where prior imputed data are used to assist with subsequent imputations. Because some of the same predictors were used in each subsequent imputation, this option allows for a better comparison between the WSHD and PMN methods. It also enables evaluation of whether the addition of other drug variables would improve the final imputed estimates.

3.3.3 Imputation Class Development for Drug Variables

The CHAID analysis criteria for each drug variable for simple WSHD are presented in Table 3.3. For this method, the CHAID analyses for each drug were first tested within each of the three age groups. Based on the small domain size of some of the drug variables (e.g., inhalants, cocaine, and heroin use), some options such as the Leaf Size and Split Size needed to be adjusted to accommodate the smaller domain sizes. If a decision tree did not result in the selection of at least two predictors, then the CHAID analysis criteria were adjusted to lower thresholds for splitting levels. For some variables, the Maximum Depth and Subtree options were modified to ensure that a tree with at least one branch was created.

No decision trees were successfully produced by age group for inhalants, pain relievers, cocaine, and heroin when using the starting criteria nor during subsequent tries when the splitting criteria were reduced to the lowest possible levels (Split Size = 20 and Leaf Size = 10). Therefore, the CHAID analyses for these four drugs were not performed individually for each of the three age groups, but they were performed for the entire sample (12 or older). The starting criteria options for the other three drugs—cigarettes, alcohol, and marijuana—were used for each age group.

Table 3.3 Decision Tree Node Options Summary for CHAID Analysis for Simple WSHD

Drug Variable	Significance Level	Split Size	Leaf Size	Maximum Depth	Subtree
Cigarettes Lifetime Use	0.05	100	50	6	Assessment
Alcohol Lifetime Use	0.05	100	50	6	Assessment
Inhalants Lifetime Use*	0.05	100	50	4	Largest
Marijuana Lifetime Use	0.05	100	50	6	Assessment
Pain Relievers Lifetime Use*	0.05	20	10	6	Assessment
Cocaine Lifetime Use*	0.05	100	50	4	Largest
Heroin Lifetime Use*	0.05	100	50	4	Largest

CHAID = Chi-square Automatic Interaction Detection; WSHD = weighted sequential hot deck.

*The CHAID analyses for these drugs were performed for those aged 12 or older.

Table 3.4 summarizes the options that were changed from the starting criteria (Table 3.1) for complex WSHD. For cigarettes, alcohol, and marijuana, all of the starting criteria options were used for each age group with one exception. For cigarettes recency among those aged 26 or older, the Split Size (50) and Leaf Size (25) starting criteria options had to be reduced. Similarly, for pain relievers recency, the same age group required a reduction in the Split Size and Leaf Size criteria. However, the starting CHAID criteria did not need to be modified for the other two age groups. For the remaining drug variables, most of the starting criteria options needed to be modified in order to produce a decision tree. There is no pattern related to the number of modifications that needed to be made among each age group. Tables A.13 through A.18 present the starting list of predictors and the final set of predictor variables chosen from the CHAID analysis for imputation classes for both simple and complex WSHD for the marijuana drug use variables.

Table 3.4 Decision Tree Node Options Changed for CHAID Analysis for Complex WSHD

Drug Variable	Age Group		
	12-17	18-25	26+
Inhalants Lifetime	Split Size = 20 Leaf Size = 10	Split Size = 20 Leaf Size = 10	Subtree = Largest Maximum Depth = 4
Cocaine Lifetime	Subtree = Largest Maximum Depth = 4		Subtree = Largest Maximum Depth = 4
Heroin Lifetime	Subtree = Largest Maximum Depth = 4	Subtree = Largest Maximum Depth = 4	Subtree = Largest Maximum Depth = 4
Cigarettes Recency			Split Size = 50 Leaf Size = 25
Inhalants Recency	Split Size = 50 Leaf Size = 25	Split Size = 20 Leaf Size = 10	Subtree = Largest Maximum Depth = 4
Cocaine Recency	Subtree = Largest Maximum Depth = 4		
Pain Relievers Recency			Split Size = 50 Leaf Size = 25
Heroin Recency	Split Size = 20 Leaf Size = 10	Split Size = 20 Leaf Size = 10	Split Size = 20 Leaf Size = 10

Table 3.4 Decision Tree Node Options Changed for CHAID Analysis for Complex WSHD (continued)

Drug Variable	Age Group		
	12-17	18-25	26+
Inhalants 12-Month Frequency			Split Size = 50 Leaf Size = 25
Heroin 12-Month Frequency	Split Size = 20 Leaf Size = 10	Split Size = 20 Leaf Size = 10	Split Size = 20 Leaf Size = 10
Inhalants 30-Day Frequency		Split Size = 20 Leaf Size = 10	Split Size = 20 Leaf Size = 10
Cocaine 30-Day Frequency	Split Size = 50 Leaf Size = 25	Split Size = 50 Leaf Size = 25	Split Size = 50 Leaf Size = 25
Heroin 30-Day Frequency	Split Size = 20 Leaf Size = 10	Split Size = 20 Leaf Size = 10	Split Size = 20 Leaf Size = 10

CHAID = Chi-square Automatic Interaction Detection; WSHD = weighted sequential hot deck.

3.3.4 Definitions of Eligible Donors for Drug Use Variables

Once the CHAID analyses were completed and the imputation classes were created, the next step was to impute the missing data using the SUDAAN HOTDECK procedure. For both simple and complex WSHD, data were divided into groups based on eligible donor sets (or missingness patterns) as shown in Table 3.5. The data were separated based on edited recency values and either the 12-month frequency or the 30-day frequency values to help maintain the logical constraints for these variables. Respondents could be assigned into multiple missingness patterns, but nonrespondents were assigned to only one missingness pattern. By using these missingness patterns, most of the recency-related logical constraints from PMN were applied. However, the logical constraints for age at first use that required the recipient and donor to have similar ages based on recency and frequency-of-use values (e.g., age and birth dates within a certain range specified by 30-day and 12-month frequencies) were not guaranteed to be maintained because of the limitations for developing imputation classes based on the CHAID algorithm.

For all drug variables, the first variable of the drug set that required imputation was the lifetime drug use indicator. For both simple and complex WSHD, these lifetime indicator variables were imputed univariately. For simple WSHD, the recency and frequency imputations were performed multivariately. For complex WSHD, the recency and frequency variables were imputed univariately. For both simple and complex WSHD, the cases were sorted by imputation class and recency and frequency variables. However, in situations where all the age groups were combined during the CHAID analysis, the data were sorted by categorical age and then imputation class. For the age-at-first-use imputations, the data were sorted by imputation class and descending age. Age at first use was imputed last in both approaches.

Table 3.5 Missing Data Patterns and Definitions of Eligible Donors for Recency and Frequency Variables

Drug Variable	Recency	12-Month Frequency	30-Day Frequency	Eligible Donors
Cigarettes	Past Month or Past Year	N/A	Missing	Past Month or Past Year with Nonmissing Frequency
	At Least Lifetime Use	N/A	Missing	Nonmissing Recency and Frequency
	Past Month	N/A	Missing	Past Month with Nonmissing Frequency
	Not Past Year	N/A	Did Not Use in Past Month	Past 3 Years or Lifetime Use
	Not Past Month	N/A	Did Not Use in Past Month	Past Year, Past 3 Years, or Lifetime Use
	Past Year or Past 3 Years	N/A	Did Not Use in Past Month	Past Year or Past 3 Years
	At Least Past 3 Years	N/A	Missing	Past Month, Past Year, or Past 3 Years with Nonmissing Frequency
Alcohol, Inhalants, Marijuana, Cocaine, and Heroin	Past Month	May Be Missing	May Be Missing	Past Month with Nonmissing Frequencies
	Past Year Not Past Month	Missing	Did Not Use in Past Month	Past Year with Nonmissing Past Year Frequencies
	Past Month or Past Year			Past Month or Past Year with Nonmissing Frequencies
	Missing	Missing	Missing	Nonmissing Recency and Frequencies

Table 3.5 Missing Data Patterns and Definitions of Eligible Donors for Recency and Frequency Variables (continued)

Drug Variable	Recency	12-Month Frequency	30-Day Frequency	Eligible Donors
Pain Relievers	Past Month	May Be Missing	N/A	Past Month with Nonmissing Frequencies
	Past Year Not Past Month	Missing	N/A	Past Year with Nonmissing Past Year Frequencies
	Past Month or Past Year		N/A	Past Month or Past Year with Nonmissing Frequencies
	Missing	Missing	N/A	Nonmissing Recency and Frequencies

N/A = not applicable.

One data quality issue that was encountered in simple WSHD dealt with cases that remained missing after a first pass of the WSHD method for each missingness pattern. There were two cases (for cocaine and heroin) where donors were not found. After going back and reviewing the data, it was observed that there was a lack of donors within the imputation classes. In one instance, the two neighboring imputation classes were collapsed into one class so that a suitable donor could be found. In the other instance, the same imputation class across two age groups was collapsed.

3.3.5 Inconsistencies after Imputation among Drug Use Variables

This section discusses the issue of data quality as it relates to the amount of consistency that is required for each individual respondent. For example, a respondent could provide consistent data on age at first use and most recent use but could have missing data for frequency of use. Under the current PMN approach, logical constraints are applied to ensure that the imputed frequency-of-use data are consistent with other survey responses and other imputed data. When imputing missing data under simple and complex WSHD, most of these logical constraints were implemented by restricting the imputation of data based on their missingness patterns. However, some inconsistencies in the imputed data still occurred. The types of inconsistency, frequency, and percentages of occurrence are summarized for both WSHD methods in Table 3.6. The percentage of inconsistent data was computed by dividing the number of inconsistent cases by the number of cases that were imputed.

Many of the inconsistencies occurred very infrequently during either simple or complex WSHD. Alcohol and marijuana had the highest percentages of inconsistency. Of the inconsistencies that were checked, age at first use was involved in three of them. Across the different methods, the inconsistency "30-day frequency greater than 12-month frequency" occurred most often. For the two WSHD methods, some of these inconsistencies may have been prevented by using additional predictor variables and also by sorting by additional variables. The inconsistencies between 30-day frequency of use and 12-month frequency of use for simple WSHD were mainly due to missingness patterns where recent use was unknown but was

restricted to either past month or past year recency and the 30-day frequency values were missing. However, for complex WSHD, the inconsistencies between 30-day frequency and 12-month frequency were due to missing 12-month frequency values. To assist with some of the inconsistencies for the age-at-first-use variable, interview date and birth date could have been included as sorting variables. Because likeness constraints were not applied for the WSHD methods at the individual respondent level, the same level of stringency as compared with the constraints of PMN could not be accomplished. Nonetheless, it was possible to account for some likeness constraints in the WSHD methods by using variables included in the PMN likeness constraints to help define imputation classes. Further research is needed to determine how the imputation classes could be restricted to maintain more consistency.

Table 3.6 Inconsistent Imputed Values for Simple and Complex WSHD

	Simple WSHD		Complex WSHD	
Type of Inconsistency	**Frequency Counts**	**Unweighted Percentage of Imputed Cases**	**Frequency Counts**	**Unweighted Percentage of Imputed Cases**
30-Day Frequency Greater than 12-Month Frequency				
Alcohol	68	6.81	77	7.53
Inhalants	3	4.17	9	10.11
Marijuana	15	5.84	31	11.52
Cocaine	2	2.70	4	5.13
Heroin	0	0.00	0	0.00
Past Month User and 12-Month Frequency Greater than 30-Day Frequency + 335 Days				
Alcohol	8	0.80	7	0.68
Inhalants	1	1.39	2	2.25
Marijuana	6	2.33	1	0.37
Cocaine	0	0.00	0	0.00
Heroin	0	0.00	1	33.33
Age at First Use Greater than Age				
Cigarettes	0	0.00	1	0.19
Alcohol	1	0.13	0	0.00
Inhalants	0	0.00	0	0.00
Marijuana	4	1.78	2	0.88
Pain Relievers	0	0.00	1	0.16
Cocaine	0	0.00	0	0.00
Heroin	1	12.50	1	12.50
Recency Not in the Past Year and Age at First Use Not Equal to Age				
Cigarettes	24	5.30	14	3.02
Alcohol	3	1.06	2	0.82
Inhalants	10	1.93	3	0.80
Marijuana	9	4.55	0	0.00
Pain Relievers	5	0.79	5	1.27
Cocaine	1	0.92	1	1.41
Heroin	0	0.00	0	0.00

Table 3.6 Inconsistent Imputed Values for Simple and Complex WSHD (continued)

Type of Inconsistency	Simple WSHD		Complex WSHD	
	Frequency Counts	Unweighted Percentage of Imputed Cases	Frequency Counts	Unweighted Percentage of Imputed Cases
Past Month User and 12-Month Frequency Greater than Interview Date − (Birth Date + Age at First Use) + 1				
Alcohol	1	0.08	6	0.61
Inhalants	1	1.08	2	1.89
Marijuana	6	2.21	5	1.84
Pain Relievers	4	1.94	0	0.00
Cocaine	1	1.33	0	0.00
Heroin	0	0.00	1	50.00
Past Year Not Past Month User and 12-Month Frequency Greater than Interview Date − (Birth Date + Age at First Use) − 29				
Alcohol	1	0.18	5	0.93
Inhalants	4	2.82	2	1.33
Marijuana	3	1.49	4	2.04
Pain Relievers	2	0.66	2	0.69
Cocaine	0	0.00	0	0.00
Heroin	1	16.67	0	0.00
Total Inconsistencies	**185**	**1.54**	**189**	**1.67**

WSHD = weighted sequential hot deck.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

3.4 Comparison of Simple and Complex WSHD with PMN

This section presents a comparison of differences in final estimates based on the 2007 NSDUH data imputed with PMN and both the simple and complex WSHD methods. Additionally, methodological differences between the PMN and simple and complex WSHD methods are outlined, and the trade-offs for implementing these procedures are discussed.

3.4.1 Summary of Statistical Tests Comparing Estimates Based on PMN with Simple and Complex WSHD

A statistical analysis was performed to examine whether there were significant differences in final imputed estimates using data imputed with PMN versus the simple and complex WSHD methods. Tables 3.7 through 3.9 present the results of these comparisons along with the weighted percentages of imputed data. The results for the demographic variables are presented in Table 3.7, and Table 3.8 presents the results for the recency variables. Table 3.9 presents the results of the continuous variables (30-day and 12-month frequency of use and age at first use). Appendix D describes the methodology used for comparing the estimates from the different imputation methods, and Appendix F presents the estimates that relate to these comparisons. Although PMN and both WSHD methods differed greatly in operation, statistically there were relatively few significant differences between the estimates.

For the demographic variables, race was the only significant variable (Table 3.7). In comparison, the other demographic variables had much lower percentages imputed than race. Thus, the lack of significance is not surprising. The percentage imputed was 2.52 percent for the race variable compared with less than 0.05 percent for the other demographic variables.

Among the drug variables, comparisons were made between PMN and simple WSHD as well as PMN and complex WSHD. For these comparisons, a few significant differences were found. As shown in Tables 3.8 and 3.9, there were no significant differences found between the simple WSHD and PMN estimates for all of the recency, 12-month frequency, 30-day frequency, and age-of-first use variables.

When the complex WSHD estimates were compared with the PMN estimates, three significant differences were found for alcohol and inhalants drug recency (Table 3.8) and for alcohol 12-month frequency (Table 3.9). The percentages imputed for two of these variables (0.90 for alcohol and 0.30 for inhalants) were relatively low. Notable was the higher percentage imputed for alcohol 12-month frequency (2.12 percent) as compared with the other variables. Table F.9 in Appendix F presents the comparable recency and frequency estimates based on complex WSHD. The estimated number of days using alcohol in the past 12 months is 86.9 for PMN as compared with 86.7 days for complex WSHD. The difference in the means is 0.2 days. Even though this difference is noted as statistically significant, when the average number of days is rounded, this difference is negligible.

Tables F.1 through F.8 present the comparisons for each level of the demographic variables and note the significant differences. Tables F.9 through F.16 present the comparisons for the drug variables and note the significant differences. Chapter 6 provides a detailed discussion on the importance of these significant differences and presents estimates for all methods tested in this evaluation.

Table 3.7 Comparisons of PMN and Simple WSHD Imputed Estimates for Demographic Variables

Demographic Variable	Number of Categories	Weighted Percentage Imputed	P-Value for Chi-square Test of Interaction with Method
Marital Status	4	0.03	0.3972
Hispanic/Latino Origin	2	0.05	0.2502
Race	4	2.52	< 0.0001
Education Level	4	0.05	0.4319

PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

Note: The weighted percentage imputed is based on the 2007 imputation indicators for cases noted as logically assigned or statistically imputed.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table 3.8 Comparisons of PMN and Simple and Complex WSHD Imputed Estimates for Recency Variables

Drug Variable	Number of Categories	Weighted Percentage Imputed	P-Value for Chi-square Test of Interaction with Method	
			PMN vs. Simple WSHD	PMN vs. Complex WSHD
Cigarettes Recency	5	0.27	0.3389	0.2568
Alcohol Recency	4	0.90	0.1658	0.0029
Inhalants Recency	4	0.30	0.5202	0.0118
Marijuana Recency	4	0.39	0.4813	0.3978
Pain Relievers Recency	4	0.63	0.8083	0.5328
Cocaine Recency	4	0.33	0.2749	0.1961
Heroin Recency	4	0.04	0.6101	0.4222

PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

Note: The weighted percentage imputed is based on the 2007 imputation indicators for cases noted as logically assigned or statistically imputed.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table 3.9 Comparisons of PMN and Simple and Complex WSHD Imputed Estimates for Frequency and Age-at-First-Use Variables

Drug Variable	Weighted Percentage Imputed	2007 PMN Estimate[1]	Difference of Means	
			PMN vs. Simple WSHD[1]	PMN vs. Complex WSHD[1]
30-Day Frequency for Past Month Users				
Cigarettes	0.20	22.6	0.0	0.0
Alcohol	1.06	8.4	0.0	0.0
Inhalants	0.04	4.0	0.3	0.1
Marijuana	0.16	12.9	0.0	-0.1
Cocaine	0.12	6.0	0.3	0.2
Heroin	0.00	15.5	-0.1	0.0
12-Month Frequency for Past Year Users				
Alcohol	2.12	86.9	0.6	0.2[a]
Inhalants	0.16	28.6	0.0	-0.9
Marijuana	0.71	101.9	-0.5	-0.7
Pain Relievers	0.51	46.2	0.9	1.0
Cocaine	0.36	43.3	1.6	1.7
Heroin	0.03	92.8	1.2	-2.7
Age at First Use for Lifetime Users				
Cigarettes	0.66	15.7	0.0	0.0
Alcohol	1.21	17.0	0.0	0.0
Inhalants	0.51	17.3	0.0	0.0
Marijuana	0.28	18.0	0.0	0.0
Pain Relievers	0.99	22.1	0.1	0.0
Cocaine	0.33	21.9	0.0	0.0
Heroin	0.02	22.9	0.0	0.0

PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

Note: The weighted percentage imputed is based on the 2007 imputation indicators for cases noted as logically assigned or statistically imputed.

[1] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

[a] Difference is statistically significant at the 0.01 level.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

It should be noted that the observed statistical differences in the estimates between methods could be attributed to using fewer consistency rules for finding donor values as well as a change in method. Additional testing needs to be performed to determine whether additional constraints could be incorporated into the WSHD methods. However, any advantage in simplicity, cost, or time required to execute could be lost if these additional modifications are built into the imputation procedure.

3.4.2 Differences and Similarities between PMN and Simple and Complex WSHD

This section attempts to quantify the similarities and differences between PMN and the two WSHD methods. Using the criteria outlined in Chapter 2, features of PMN are compared and contrasted as they relate to simple and complex WSHD.

3.4.2.1 Methodological Steps for WSHD

- Both simple and complex WSHD have two main steps: (1) imputation class development or CHAID analysis, and (2) hot-deck imputation. Unlike PMN, an item response propensity adjustment (i.e., adjustment of sampling weights for item nonresponse) was not performed for simple or complex WSHD.
- In PMN, the relative importance of predictor variables is determined by standard estimating equation techniques. In other words, there are objective criteria based on methodology, such as regression, which quantify the relationship between a given predictor variable and the response variable in the presence of other predictors. In contrast, the CHAID analysis measures the associations between variables and identifies patterns in the data to help with donor selection, but it does not measure the relationship of the imputation variable and the predictors and does not incorporate the survey weights in the association tests.
- During the predictive modeling step of PMN, the survey design is accounted for by the use of the sampling stratification variables and the use of the response propensity adjusted weights for item respondents to ensure unbiased estimates. In contrast, the sample weights are used in the hot-deck imputation process so that the expected value of the weighted distribution of the imputed values is preserved when compared with the weighted distribution of the item respondent data within imputation class. The WSHD methods do not use the stratification variables as part of the imputation process, because a variance component is not required (as compared with the PMN regression models) for implementation of WSHD.

3.4.2.2 Complexity of Data Consistency and Order of Imputation for WSHD

- One way PMN varies from the WSHD approach is in its use of many likeness constraints at the individual case level. Likeness constraints as defined in PMN are a restriction imposed to limit the pool of potential donors for a given recipient so that donors and recipients are as alike as possible. For example, in the imputation of education level, one of the likeness constraints required the donor's age to be the same as that of the recipient. These types of restrictions or constraints are more difficult to implement in simple and complex WSHD because of the limitations for developing imputation classes based on the CHAID algorithm. The constraints imposed in these

methods were based on defining eligible donors using missingness patterns in the data and did not use the stringent likeness constraints of PMN. Therefore, the two WSHD methods as implemented do not ensure internal consistency as well as PMN does, because there is no convenient way to deal with records on a case-by-case basis.

- In terms of changing the order in which variables are imputed, it is easier to make changes in both simple and complex WSHD as compared with PMN because fewer constraints are applied at the individual case level. For simple WSHD, the reordering or insertion of variables requiring imputation would not be problematic because of the limited set of predictor variables used for each imputation. Similarly, for complex WSHD, a new set of drug variables could be imputed without any issues related to the ordering of questions.

3.4.2.3 Issues for Implementation of WSHD

- Unlike the PMN process, only one modeling step is required for both WSHD methods. The development of the imputation classes or CHAID analysis can be compared with the predictive mean modeling step in PMN. Similar to the model-fitting exercises in PMN, the CHAID analysis requires manual intervention, though the level of intervention is quite different from PMN. For the WSHD approaches, the manual intervention involves setting initial significance and splitting criteria based on sample sizes, whereas the model-fitting exercises for PMN involve determining which predictor variables are not significant and evaluating many statistical tests for variable importance in regression models. The CHAID analysis options summarized in Tables 3.3 and 3.4 outline the amount of model-fitting work that would be required for each of the WSHD methods. For simple WSHD, drugs with smaller domain sizes (inhalants, pain relievers, cocaine, and heroin) required collapsing of age groups and required adjustments to the starting CHAID criteria. For complex WSHD, more modifications to the starting CHAID criteria were required for the same set of drug variables. For both simple and complex WSHD, cigarettes, alcohol, and marijuana did not require modifications and collapsing by age groups to complete the CHAID analyses.
- The implementation of both simple and complex WSHD requires SAS and a special component of SAS called Enterprise Miner for CHAID analysis (or some other software that performs CHAID) and SUDAAN for hot-deck imputation. PMN requires only base SAS and SUDAAN.
- As compared with PMN, both WSHD methods require less time to develop because of the simplicity of the approach that was chosen to be implemented. However, if additional constraints were built into the WSHD methods to match the complexity of PMN, then the level of effort would increase, but based on experience, it appears that it would not approach the same level required by PMN.
- Both simple and complex WSHD require fewer model-fitting (e.g., CHAID analysis) exercises, and the number of specifications that need to be modified are small as compared with the number of models and predictor variables that need to be evaluated in PMN. Because both WSHD methods have less dependency built into them (i.e., use less prior imputed data), these programs could be implemented in a shorter time frame. Also, because of the simultaneous process built into both WSHD methods, many drug variables could be imputed simultaneously and thus save additional time in the annual data processing schedule.

3.5 Summary and Options

The two WSHD methods that were evaluated were fairly easy to implement and may be viable alternatives for imputing the NSDUH data. The goal of identifying a procedure simpler to implement than PMN was accomplished, but the simpler procedure has a few disadvantages and needs additional refinements. The main disadvantage is that the level of data consistency that is currently required for the NSDUH is not easily implemented using WSHD methods. Additional restrictions would need to be added to the WSHD imputation class development process to help maintain more internal consistency among variables and to refine the definitions for eligible donor sets for drug use recency and frequency and age-at-first-use variables. Moreover, because of small domain sizes for some of the rarer drugs, many of the CHAID modeling options would need to be modified as well. One advantage is that once these modifications are made in one annual survey cycle, the options may not need to be changed unless the sample sizes for those drugs change significantly.

4. Sequential Regression Multivariate Imputation Using IVEware

This chapter describes the steps that were performed to evaluate the use of IVEware (Imputation and Variance Estimation Software)[21] (Raghunathan, Solenberger, & Van Hoewyk, 2002) in imputing selected core demographic and drug variables. In particular, IVEware was tested on a subset of these variables to determine how much computational time would be needed to impute all of the variables selected for the evaluation. The ability to use IVEware to develop estimates based on the multiply imputed datasets was also evaluated. Because of its ease of implementation, IVEware was the first of the imputation methods tested, and as a result, the set of predictor variables used in this method differ from those used in the other methods. To see whether the software could handle all drug variables that would require imputation, all lifetime drug variables that are imputed in the annual National Survey on Drug Use and Health (NSDUH) were tested.

4.1 Overview of Sequential Regression Multivariate Imputation Using IVEware

IVEware is a SAS®-callable software application that includes a general-purpose, multivariate imputation procedure that can handle relatively complex data structures when the data are missing at random. The IMPUTE module in IVEware performs multiple imputations by using a set of sequential univariate imputations in a cyclic fashion; that is, after completing the sequence of univariate imputations, it cycles back to the beginning and repeats the sequence using the latter imputations as predictors. Ultimately, all the variables are conditioned on all the others. This is a useful property to have when there are complex patterns of nonresponse among correlated items.

The procedure can be referred to as a sequential regression procedure, in that a separate regression model is developed for each variable with missing data (as the dependent variable), and each model is then used to generate one set of random imputations. The sequential regression procedure starts by imputing the variable with the least amount of missing data. During this first imputation, only the variables that have complete data are used as covariates in the regression model. After the first variable has been imputed, this procedure continues through the remaining variables with the order of imputation being determined by the amount of missing data. For each subsequent regression model, both the observed and previously imputed variables are used as predictors. Sequential regression multiple imputation (MI) makes use of the intuitively appealing idea to use univariate regression models for imputation purposes, but it is more powerful than using single imputations performed with univariate regression models because it takes into account the multivariate covariance structure of the data.

[21] IVEware can be downloaded free of charge at http://www.src.isr.umich.edu/software/. Version 0.1 of IVEware was tested for this report. In December 2010, Version 0.2 was released and this version was not tested for this report.

In addition, IVEware allows restrictions and bounds to be placed on the variables being imputed to allow for the appropriate subpopulations to be used in the regression models. As an example of a restriction, the imputation of recency of drug use can be restricted to only lifetime users. As an example of bounds, the frequency of drug use can be bounded by the appropriate number of days (i.e., 30 days for past month use or 365 days for past year use). This software has been used to produce national estimates for family and personal income in the National Health Interview Survey (Schenker et al., 2006; 2007).

4.1.1 Setup for Imputation

The first step in the process is to prepare the data for imputation. During this phase, logical constraints are programmed into the data to prevent erroneous imputations. For instance, if an individual is 14 years or younger, then he or she would be assigned a marital status of "not applicable." Also, because the drug use measures were imputed in a hierarchical manner similar to the predictive mean neighborhood (PMN) method (first recency, followed by 12-month frequency, then 30-day frequency, and then age at first use), certain indicator variables were created to assist with the appropriate subsetting of the data. For example, if an individual is imputed to be a lifetime user but not a past year user of alcohol, then in the setup stage for the 12-month frequency variable, the indicator variable would be assigned a value to indicate this level of information about the individual. These steps were performed to prevent erroneous values from being used in the model creation steps as IVEware cycles through the variables requiring imputation.

4.1.2 Invocation of the IMPUTE Module

Once the data have been prepared for imputation, the next step is to invoke the IMPUTE module. As described in the next section, several options are available and must be defined by the user. In addition, each variable on the input dataset must be identified during this step as one of the following types or requiring a specific action: continuous, categorical, count, mixed, transfer, and drop. The variables used as predictors in the model and the variable(s) requiring imputation are identified as one of the first four types. Variables listed after the TRANSFER statement are carried over to the imputed dataset but are not imputed or used as predictors in the imputation model. Variables listed after the DROP statement are excluded from the imputation procedure and do not appear in the imputed dataset.

After each option and variable type has been defined, the IMPUTE module is then invoked to begin the imputation process. IVEware then creates a sequence of multiple regression models based on the type of variable being imputed. For example, when imputing drug recency of use, a polytomous logistic regression model is constructed. Once the regression model has been fit, the model coefficients are perturbed by the addition of a random error term using the PERTURB option as described in Section 4.1.3. Based on the perturbed model, the imputed values for the variable undergoing imputation are assigned. The sequence of imputing missing values can be continued in a cyclical manner, each time overwriting previously drawn values, building interdependence among imputed values, and exploiting the correlational structure among covariates. When the last cycle is complete or until the convergence criteria is met, the final imputed values are output along with all model covariates and variables listed on the transfer statement.

4.1.3 Model-Building Options in the IMPUTE Module

The IMPUTE module in IVEware contains many options for specifying the imputation model. If the user wishes to use a stepwise regression approach, two options are available to limit the size of the final model: MAXPRED and MINRSQD. The first option specifies the maximum number of predictors to include in the regression model. If the MAXPRED option is used, then the software selects the "best" set of *N* predictors. In this context, "best" is defined as those variables that contribute the most to the R-square value. Also, because the software is able to impute multiple variables at one time, the user may specify different model sizes for different variables. An alternative method for limiting the size of the model is the MINRSQD option. Specifying a value for this option requires a minimum change in R-square to be observed before a predictor is added to the model. A small number such as 0.05 results in a regression model with a higher number of predictors, whereas a larger number such as 0.25 tightens the restriction and leads to a reduced number of predictors in the model. If neither option is specified, then no stepwise regression is performed.

Both the MAXPRED and MINRSQD options are useful tools for limiting the size of the model and reducing computation time. However, early tests of the software indicated that computation time would not be an issue for the NSDUH. Therefore, these options were omitted during the analysis. This decision was also motivated by a review of the relevant literature, which recommended using all available information about a variable when performing imputation (Little & Raghunathan, 1997; Khare, Little, Rubin, & Schafer, 1993). Larger models are typically used because the goal is to predict a missing value rather than identify the exact relationship between variables. Including more predictor variables results in a higher correlation between the observed and predicted values and preserves important statistical relationships in the dataset. Maintaining such relationships also helps preserve the validity of analyses by secondary users.

Another option available to users of IVEware is MAXLOGI. This statement identifies the maximum number of iterations used in the Newton-Raphson algorithm for producing maximum likelihood estimates. If the algorithm fails to converge before *N* iterations, then a warning is printed to the SAS log file. The default value for this option is 50, and this value was found to be sufficient during the evaluation. IVEware also allows the user to control perturbations (i.e., the addition of a random error value) to the imputed values with the PERTURB option. The addition of a random error value to the imputed value is used to account for variance from the model-fitting step. The default setting for this statement causes the software to perturb model coefficients using a multivariate normal approximation of the posterior distribution. The user may also request that the software use the Sampling-Importance-Resampling algorithm to draw coefficients from the actual posterior distribution of parameters in the logistic, polytomous, and Poisson regression models. This alternative setting may be useful when the range of imputed responses is restricted (i.e., alcohol age at first use is less than or equal to age of respondent) because it is difficult to draw values of parameters directly from the posterior distribution with truncated normal likelihoods (Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001). However, because multiple variables were being imputed at one time (recency, frequency, and age at first use), the decision was made to use the default setting.

Two other options available in IVEware can be used to restrict imputed values. The RESTRICT statement instructs the software to only perform imputations for observations that satisfy the logical expression. This option was used extensively during the evaluation in an attempt to maintain consistency between imputed and observed values. An example of this constraint was trying to restrict the imputation of monthly frequency to only those individuals identified as past month users. The second option available to restrict imputation is the BOUND statement. Rather than restricting which values are imputed, this statement is used to limit the range of possible values for the imputed variable. For instance, past month frequency should not exceed 30 days. This method was also used during the testing to ensure that imputed values represented valid responses.

In addition, the user of IVEware may specify the number of cycles used in the sequential regression algorithm. In the first cycle, the variable with the least amount of missingness is imputed first. The predictors used for this imputation include all other variables specified that contain no missing values. Once the imputation of this variable is complete, the variable with the second least amount of missingness is imputed. The predictors for this model include all other variables with no missing values as well as the variable that was just imputed. This first cycle continues until all missing values have been imputed for all variables. In the second cycle, the order of imputation is the same as in the first cycle, but now the model includes all other variables and prior imputed values. This process continues until the N^{th} cycle is completed as specified by the user. Except for when trying to diagnose problems, five iterations were used. This number of iterations was found to be sufficient for obtaining stable regression coefficient estimates (Yulei, Zaslavsky, Harrington, Catalano, & Landrum, 2007).

4.2 Imputation of Demographic Variables

The demographic variables (race, Hispanic/Latino origin, education level, and marital status) were the first set of variables tested using the IMPUTE module in IVEware. In contrast to PMN, all age groups were imputed concurrently. This approach was implemented with the intent of saving time in program development and maximizing the amount of information used to construct the regression models. IVEware performed well during these initial tests, and no problems were found with the IMPUTE module in handling the large dataset. Flag variables were constructed to restrict the imputation of marital status to individuals aged 15 or older and education level to those aged 18 or older. The software correctly handled these restrictions and set these observations to a "null" value. However, the IVEware user guide (Raghunathan et al., 2002) did not clarify whether these cases, which were noted as legitimate skips, would be used in subsequent iterations of the sequential regression algorithm for model building. This was one of the first drawbacks found with using IVEware because the amount of control over the process was very limited. The next section describes the final approach used for imputing the demographic variables.

4.3 Imputation of Drug Variables

4.3.1 Lifetime Drug Use Variables

One of the key features that prompted the evaluation of IVEware is its ability to perform multivariate imputations; thus, the demographic variables and lifetime drug use variables were

imputed simultaneously. The same demographics previously mentioned were imputed, along with the lifetime drug use variables for the following drugs: cigarettes, cigars, pipes, chewing tobacco, snuff, alcohol, marijuana, smokeless tobacco, inhalants, sedatives, cocaine, crack, heroin, LSD, PCP, Ecstasy, methamphetamine, OxyContin, hallucinogens, pain relievers, and stimulants. Although the software was able to impute all values without program errors or warnings, inconsistencies were found in the resulting output. For example, IVEware could impute an individual to be a lifetime user of crack and a lifetime nonuser of cocaine, which violates established relationships for drug use measures for parent and child drugs.

In an attempt to remedy this situation, the BOUND and RESTRICT statements were incorporated into the program to test the ability to control the inconsistencies. However, when both the parent and child drug values were missing and required imputation, the RESTRICT statement failed to work. Instead, the RESTRICT statement only restricted the respondent cases used for modeling. An attempt was also made to ensure that the parent drug value was greater than or equal to the child drug value, but this also failed because a categorical variable was not allowed in the BOUND statement. Specifying these variables as categorical was necessary to force the software to use a logistic regression model rather than a linear regression model. Another option would be to recode these variables once the imputed datasets are created so that they are consistent, but this option does not seem feasible because of the algorithm being used. Only the first variable in the first cycle is conditional on observed values. All subsequent iterations/cycles are conditional on both observed and imputed values. Consequently, a post-imputation recode would not prevent erroneous values from being used during the modeling process. The only solution for preventing such inconsistencies appears to be imputing the lifetime drug use variables one at a time and using the RESTRICT statement to identify legitimate skips. For example, once lifetime cocaine use has been imputed, the RESTRICT statement could be used to only impute missing crack values for lifetime cocaine users. However, this approach nullifies one of the main attractions of using IVEware, namely, the ability to impute multiple variables in a single program.

4.3.2 Recency and Frequency of Drug Use

This section describes two scenarios for testing the IVEware software for imputing the recency and frequency of drug use. The scenarios tested the use of RESTRICT and BOUND statements with the IMPUTE statement for imputing these drug use variables.

Approach 1: Despite the inconsistencies observed with the lifetime drug use variables, the purpose of the evaluation was to determine which set of variables could be effectively and efficiently imputed with IVEware. Therefore, the evaluation continued by next imputing the recency, frequency, and age-at-first-use variables. The first approach attempted to impute all measures for all drugs simultaneously in addition to the demographic variables and lifetime drug use variables, but this approach created many inconsistencies between the recency and frequency variables. The RESTRICT and BOUND statements were added to the program to ensure valid responses and to try to maintain consistency between variables, but, as with the lifetime drug use variables, this attempt was not always successful. If individuals were missing variables for both recency and frequency of use, then they could have been given responses that were contradictory. For example, respondents may be imputed as past year but not past month users and then be given 30-day frequencies that indicate they are past month users. In addition to the

inconsistencies observed, many software errors were encountered during this approach. Multiple attempts were made to resolve these issues including contacting the software developer, but these efforts were generally unsuccessful.

Approach 2: Given the problems encountered with the first approach, a second approach was tried that only focused on the imputation of three sets of drug variables: cigarettes, alcohol, and marijuana. In addition to concentrating on a smaller set of the drugs, the decision was made to impute the drug variables in a similar sequence as in PMN to reconcile some of the inconsistencies through subsetting the data before using the IMPUTE module. Another change in this approach was the way in which the multiply imputed datasets were created. The earlier tries used a single program to create five output datasets and five imputed values for each variable requiring imputation. The new method separated programs and different random starting seeds to create each output dataset. There was a separate program for each stage of the imputation process under this approach. The first stage was imputing the demographic and lifetime drug use variables, as described previously, with each of five programs producing a single set of imputed values and a final dataset. Based on the results of earlier testing, it was also determined that cigarettes recency could be added to this initial program without problems, so it was imputed along with the demographic and lifetime drug use variables.

After this stage was complete, five programs were then created to impute cigarettes 30-day frequency. Each program would read in a single dataset from the first stage and use a different random starting seed to impute 30-day frequency. The RESTRICT statement was added to only impute 30-day frequency for those respondents with an appropriate recency of drug use, and the BOUND statement was added to ensure only valid responses. In a similar manner, the five output datasets from this stage were then used to impute cigarettes age at first use, with appropriate steps taken before imputation to ensure only the appropriate individuals were imputed. This process then continued with alcohol recency, 12-month frequency, 30-day frequency, and age at first use and was followed by marijuana recency, 12-month frequency, 30-day frequency, and age at first use. This method seemed to do a better job of preventing the blatant inconsistencies between recency and frequency that were observed with previous attempts. Also, the software errors previously encountered when trying to impute all variables at one time did not occur. However, using such an approach negates many of the purported benefits of IVEware and requires a great deal of programming time to set up multiple programs for each step in the process.

After completion of the second approach (imputing the cigarettes, alcohol, and marijuana drug variables), the decision was made to not attempt any additional imputations using IVEware. The decision was based on the performance issues of the software and the amount of time required for the IMPUTE module to run. Most of the programs that were developed took several hours of computer run time to complete the imputations. Tables A.19 through A.21 in Appendix A present the model summaries for the demographic variables and the marijuana drug use variables.

4.3.3 Inconsistencies after Imputation among Drug Use Variables

One main objective of this evaluation was to determine how well imputation methods maintained consistency among variables. If a method was considered efficient and easy to

implement and it would allow estimation of the variance due to imputation, but it failed to maintain consistency among the variables, then that method might not be acceptable. As previously mentioned, multiple attempts were made to ensure the congruency between the lifetime drug use measures for parent and child drugs. However, because of the limited amount of control granted by IVEware to the user, these attempts were ultimately unsuccessful. For example, 66 individuals were missing lifetime drug use values for both stimulants and methamphetamine. After imputation, 23 inconsistent values resulted such that lifetime users of methamphetamine were imputed to be lifetime nonusers of stimulants. This inconsistency should not occur. If an individual never used stimulants, then he or she also should not have ever used methamphetamine.

The first approach where recency, frequency, and age at first use were imputed simultaneously led to several inconsistencies between recency and frequency. For example, an individual who was missing both alcohol recency and 30-day frequency could have been imputed as not a past month user and still be given a 30-day frequency. Of the 863 individuals who were missing both recency and 30-day frequency, 838 were given inconsistent values after imputation because their imputed recency values indicated no past month use (i.e., more than 30 days ago) and their imputed 30-day frequency values indicated use in the past 30 days. Although it would be quite easy to "hard-code" the 30-day frequency to a value that was consistent, these values have already been used by IVEware to model other variables in subsequent iterations.

The second approach, where a separate program for each drug use measure was used, eliminated the problems with inconsistencies between recency and frequency by using the hierarchical structure and editing before the imputation. Because recency was done first, the input dataset for the frequencies did not have any missing values for recency. This allowed editing those who were not past year or past month users to the appropriate skip code before imputing 12-month or 30-day frequency. However, there were some instances where 30-day frequency was greater than 12-month frequency. These problems could have been prevented by using the BOUND statement and some additional editing, but this type of error was not anticipated. Rather than incurring the expense of editing all of the programs and resubmitting them, a test was performed that resulted in the 30-day frequency not being greater than the 12-month frequency when the BOUND statement was used.

4.4 Comparison of IVEware with PMN

This section presents a comparison of differences in final estimates based on the 2007 NSDUH data imputed with PMN and IVEware. Additionally, methodological differences between PMN and IVEware are outlined and the trade-offs for implementing these procedures are discussed.

4.4.1 Summary of Statistical Tests Comparing PMN Estimates with IVEware Estimates

Although IVEware is not recommended for use in the NSDUH because of performance issues, the imputed data were available to use for comparing estimates. Similar to the analysis presented in Chapter 3, the IVEware estimates for demographics, cigarettes, alcohol, and marijuana were compared with PMN estimates. Tables 4.1 and 4.2 present the results of these comparisons along with the weighted percentages of imputed data.

Table 4.1 Comparisons of PMN and IVEware Imputed Estimates for Demographic and Drug Recency Variables

	Number of Categories	Weighted Percentage Imputed	P-Value for Chi-square Test of Interaction with Method
Demographic Variable			
Marital Status	4	0.03	0.3750
Hispanic/Latino Origin	2	0.05	0.1620
Race	4	2.52	< 0.0001
Education Level	4	0.05	0.4062
Drug Variable			
Cigarettes Recency	5	0.27	< 0.0001
Alcohol Recency	4	0.90	0.0036
Marijuana Recency	4	0.39	< 0.0001

PMN = predictive mean neighborhood.

Note: The weighted percentage imputed is based on the 2007 imputation indicators for cases noted as logically assigned or statistically imputed. Imputations were not performed for inhalants, pain relievers, cocaine, and heroin for IVEware.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table 4.2 Comparisons of PMN and IVEware Imputed Estimates for Frequency and Age-at-First-Use Variables

Drug Variable	Weighted Percentage Imputed	2007 PMN Estimate[1]	Difference of Means: PMN vs. IVEware
30-Day Frequency for Past Month Users			
Cigarettes	0.20	22.6	0.0[a]
Alcohol	1.06	8.4	0.0[a]
Marijuana	0.16	12.9	0.0
12-Month Frequency for Past Year Users			
Alcohol	2.12	86.9	0.1
Marijuana	0.71	101.9	-1.0[a]
Age at First Use for Lifetime Users			
Cigarettes	0.66	15.7	0.0[b]
Alcohol	1.21	17.0	0.0
Marijuana	0.28	18.0	0.0

PMN = predictive mean neighborhood.

Note: The weighted percentage imputed is based on the 2007 imputation indicators for cases noted as logically assigned or statistically imputed. Imputations were not performed for inhalants, pain relievers, cocaine, and heroin for IVEware.

[1] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

[a] Difference is statistically significant at the 0.01 level.

[b] Difference is statistically significant at the 0.05 level.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Similar to the weighted sequential hot-deck method, the only demographic variable that showed significant differences between the IVEware estimate and the PMN estimate was race. One possible explanation for this difference is the use of additional predictor variables for the

imputation of race. Unlike PMN, where the order of variables and predictors are controlled, IVEware imputes the variables with the least amount of missingness first and then cycles through the remaining variables using all imputed variables as predictors for the next variable needing imputation. Because the race variable has the largest amount of missing data, it was the last variable imputed for the set of demographic variables. As a result, the Hispanic/Latino origin variable was used as a predictor for race, which differs from PMN. Tables F.1 through F.8 in Appendix F present estimates for each level of the race variable and denote the significant differences for each level. Chapter 6 presents a detailed discussion of the significant differences among all methods tested in this evaluation. Figure G.1 in Appendix G presents graphical results of the significant differences for the race variable.

As shown in Table 4.1, each of the recency measures for cigarettes, alcohol, and marijuana showed significant differences between the PMN and IVEware estimates. Tables F.9 through F.16 present the estimates of each level of the drug variables and denote significant differences. Although these three recency variables denote significant differences, the magnitude of the differences is quite small. The differences in the estimated proportions for drug recency for cigarettes, alcohol, and marijuana between PMN and IVEware (Table F.9) do not differ more than 0.2 percent. Additional significant differences were found for 30-day frequency for cigarettes and alcohol, 12-month frequency for marijuana, and age at first use for cigarettes (Table 4.2). Many of these differences can be attributed to the cyclic nature of the IVEware procedures and thus the use of different predictor variables between PMN and IVEware.

4.4.2 Differences and Similarities between PMN and IVEware

This section attempts to quantify the similarities and differences between PMN and the sequential regression method used in IVEware. Using the criteria outlined in Chapter 2, features of PMN are compared and contrasted as they relate to IVEware.

4.4.2.1 Methodological Steps for IVEware

- IVEware has only one key step for implementation, performing a sequence of regression models to determine the final imputed values, as compared with PMN's three key processes (response propensity adjustment, predictive mean modeling, and hot-deck imputation).
- One of the primary differences between IVEware and PMN is the ability to incorporate features of the sample design in the analysis. Under both methods, the means of a particular outcome variable are modeled as a function of the predictor variables, where these means provide a summary of the effects of predictors on the outcome variable. IVEware uses unweighted regression models in comparison with PMN, which performs weighted regression where the sample design weights are used to ensure the proper variance-covariance is computed.
- Both IVEware and PMN fit regression models based on the respondents for whom the outcome is observed, resulting in a vector of beta coefficients and a variance-covariance matrix. However, the methods differ in the usage of this vector and its corresponding matrix. In IVEware, this vector is perturbed to obtain predicted values for both the item respondents and item nonrespondents. The PMN method uses the unperturbed model coefficients to produce predicted values for both item respondents

and item nonrespondents. With IVEware, the final imputed values for nonrespondents are the predicted values obtained from the regression; that is, no respondent acts as a "donor," as compared with PMN where the calculation of predicted means is followed by a hot-deck step to identify a donor value for the final imputed value.
- Because of its cyclic nature, IVEware has the ability to use more predictor variables to determine final imputed values than PMN. The set of predictor variables for PMN is restricted to a predetermined list. In IVEware, if a predictor variable needed to be removed from a model because of convergence issues, then this variable had to be removed from the possible set of predictor variables and then would not be used as a predictor for other imputations. This is unlike PMN, where the model convergence is controlled by the user assessing which variables are problematic.

4.4.2.2 Complexity of Data Consistency and Order of Imputation for IVEware

- PMN preserves the complex relationships by using conditional probabilities and logical and likeness constraints to restrict the neighborhood of potential donors. In IVEware, it was possible to limit the possible range of imputed values through the use of the RESTRICT and BOUND statements. However, these options are not all-inclusive of the number of constraints that need to be applied. Additional subsetting of the data by developing a sequence of programs (similar to PMN) would need to be developed to maintain all of the consistency issues.
- In IVEware, the ordering of imputations is defined by the rate of nonresponse in the data. Variables with the least amount of missingness are imputed first, and this order continues until finally the variable with the greatest amount of missingness is imputed. This procedure would accommodate reordering of variables requiring imputation better than PMN because the procedure controls the order based on the amount of missing data.

4.4.2.3 Issues for Implementation of IVEware

- The IVEware software package can be downloaded for free and must be used in conjunction with SAS. The testing of the DESCRIBE module of IVEware resulted in unresolved performance issues and thus required the use of SUDAAN® software (RTI International, 2013) to analyze the multiply imputed data. The use of IVEware has several major drawbacks including the lack of documentation for diagnosing programming errors and technical support, performance issues such as long program run times, and uncertainty of whether the software will be upgraded to be compatible with newer versions of SAS.
- One of the advantages of IVEware over the PMN method is its ability to cycle through multiple variables quickly. For some demographics and lifetime drug use variables, this could be done in a single program. Completing this same task with the PMN method requires many additional steps and SAS programs. IVEware would require less time to develop SAS programs because only the set of predictors and type of variable (e.g., categorical, continuous) needs to be defined. Based on the variable type, the software determines the appropriate regression model and then performs the imputations all in one single SAS step, as compared with PMN where there are a series of SAS programs that must run for each of the three key processes.

4.4.3 Estimating Variance Due to Imputation

One of the goals of this evaluation was to estimate the variance introduced to the NSDUH estimates due to imputation. To address the objective of quantifying the variance due to imputation, a heuristic approach for assessing some of the variance inflation that occurs within the IVEware imputation method is presented in this section. The DESCRIBE module of IVEware performs MI analyses using the combining rules described in Rubin and Schenker (1986) and provides descriptive analyses such as the estimation of means, proportions, and contrasts. Available options in the DESCRIBE module include the ability to specify a stratum variable, a cluster or primary sampling unit variable, and a weight to use in the analysis. Once the five complete datasets were created, the DESCRIBE module was tested. However, fatal errors occurred that caused the software to repeatedly stop working and the software manual did not provide information regarding the source of the errors. Discussions with the software developer were also unsuccessful at resolving the issue. Therefore, the decision was made to use SUDAAN to analyze the multiply imputed data.

When data are multiply imputed, the between-imputation variance estimate is the component that quantifies the variation due to differences across the *m* sets of imputations. Smaller values of the between-imputation variance estimate indicate a more stable imputation process, whereas larger values indicate a more unstable process. The within-imputation variance estimate is the component that is obtained by assuming that the imputed data are the actual missing values. The relative increase in variance due to imputation represents the increase relative to the naive within-imputation variance contribution that results when the imputation variance contribution is ignored.

Table B.1 in Appendix B presents the percentages of imputed data, the multiply imputed estimates, and the estimated variance components (between, within, and total) generated using IVEware. The (estimated) within-imputation variance[22] is simply the average of variance estimates for the given estimate obtained from each of the five sets of imputed data, whereas the (estimated) between-imputation variance is the sample variance of the resulting five estimates. Additionally, the table presents measures of relative increase in variance due to imputation and 95 percent confidence intervals[23] for the estimates. The methodology described in Appendix B was used to compute the two components of variance for the five MI sets (where $m = 5$ denotes the number of times the imputations were performed) generated from the second approach described in Section 4.5.

Relative to the size of the estimates, the between-imputation variance estimates presented in Table B.1 are small with the exception of American Indian/Alaska Native (0.0067). However, the percentage imputed for this racial category (24.56 percent) is larger than any other racial category, and thus the relative increase in variance due to imputation (63.52 percent) is large. For the other variables, the relative increase in variance due to imputation is considerably small.

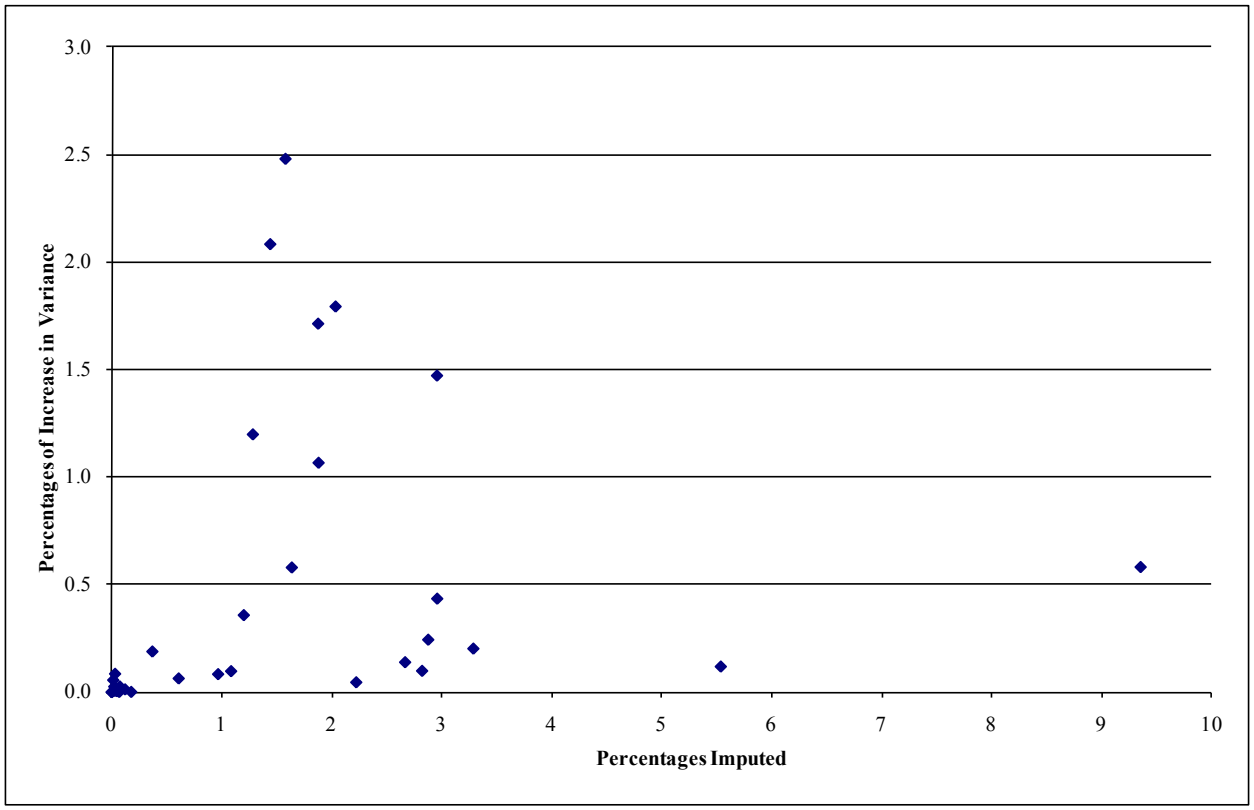
A comparison of relative increase in variance versus the percentage of imputed data can be a useful tool for evaluating variance inflation due to imputation. If a positive linear

[22] This means that the word "estimated" is hereafter implied whenever "within-imputation variance" is mentioned.

[23] Barnard and Rubin (1999) discuss the computation of confidence intervals for MI data.

relationship exists between the percentage of missing data and the relative increase in variance, then there is evidence to support the hypothesis that an increase in variance is a function of the amount of missing data. Figure 4.1 is a pictorial representation of the relative increases in variance from IVEware (Table B.1) that illustrates the nature of the relationship between the percentage of imputed data and the variance inflation due to imputation. As seen in the figure below, the relative increase in variance is generally low (less than 3 percent) for IVEware.

Figure 4.1 Relative Percentages of Increase in Variance as a Function of the Percentages of Imputed Data for IVEware



Note: This figure excludes one extreme relative percentage of increase in variance for American Indian/Alaska Native (63.52 percent) shown in Table B.1 in Appendix B.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Using this measure of relative increase in variance due to imputation, there appears to be some linear relation, which indicates that as the percentage imputed increases, the variance increases for some but not all of the estimates. Although the number of estimates involved in this analysis does not include all of the variables imputed in the NSDUH, there appears to be a positive linear relation between the weighted percentage of missing data and the relative increases in variance.

The variance due to imputation has not been accounted for in national estimates for the NSDUH because there is no simple way to estimate this variance and include it in all released estimates and data. By examining the amount of variance inflation using the data from IVEware, it can be concluded that the relative increase in variance due to imputation is typically small (with a majority of the relative increases being lower than 3 percent), thus supporting the

assumption that this variance is ignorable. In other words, for variables that undergo imputation and have low item nonresponse rates, the variance inflation is likely not a serious problem.

4.5 Summary and Options

The IMPUTE module in IVEware seemed promising on paper, much like many other off-the-shelf software procedures that claim to provide an easy and quick way to impute data. One main reason for evaluating this software was its claim to handle complex data. The main decision for not investigating these software procedures further was based on the issue of technical support. Although the software developers were willing to help with the problems, they were not able to determine a solution for the problem of the software just shutting down and not giving any warning messages in the program logs. The software manual does not provide sufficient instructions on how to use the procedures, and no information was provided for troubleshooting problems. Another important factor in determining whether to continue the evaluation of this software was the uncertainty that the software will continue to be developed or enhanced and whether it will be compatible with future versions of SAS. Finally, the inability of the user to decompose the process into its component parts to find out exactly what happened at each step of the way was a major disadvantage.

Among the variables chosen for testing, it was determined that the demographic and lifetime drug use variables could be imputed without consistency problems, with the exception of inconsistencies between parent and child lifetime drug use variables indicators. The drug variables that had complex relationships and bounds were not imputed without consistency problems. One of the main selling features of the software was the ability to use the RESTRICT and BOUND options to control the data being imputed. However, these options did not work in the simplest of tests for the drug use variables. This software is best suited for small imputation problems where only simple consistency relationships between variables are needed. For example, this method could be tested on the income variables because there are fewer logical constraints to maintain. This software has been used for imputing income variables for the National Health Interview Survey (Schenker et al., 2007). However, the authors note that there were some inconsistencies in the final data.

This software could work for imputing the drug use variables, if multiple programs were developed to allow the complex logical constraints to be implemented. However, this approach defeats the software's primary advantage and would not necessarily simplify imputation procedures for the NSDUH. The new version of IVEware (Version 0.2) may resolve some of the issues that are mentioned above, but the new version of the software was not tested for this evaluation. Chapter 5 discusses an alternative option that is better suited for the NSDUH where the sequential regression procedure is used in combination with a hot-deck procedure.

This page intentionally left blank

5. Modified Predictive Mean Neighborhood Multiple Imputation

The modified predictive mean neighborhood multiple imputation (modPMN-MI) method is so named because it includes both simplifications and modifications to the predictive mean neighborhood (PMN) method to convert it to a multiple imputation (MI) method. One key feature of this method relates to "doubly protected" MI. Unlike most applications of MI, the survey design and the estimated probabilities of response were accounted for in the modeling process. As a result, even when the MI imputation model used to predict the missing values is incorrect, the resulting estimators are still nearly unbiased so long as the estimated probabilities of response are nearly unbiased. Conversely, when the estimated probabilities of response are biased but the MI imputation model holds, the resulting estimators are also nearly unbiased. Estimators with this property are said to be "doubly protected from nonresponse bias" (Kott & Folsom, 2010).

Using SUDAAN® procedures (RTI International, 2013), univariate regression models were fit for each variable selected for imputation. Upon completion of the imputation of one variable, both the observed and previously imputed data were used as predictor variables for the next regression model. Borrowing a key feature from IVEware, modPMN-MI allows the set of sequential univariate imputations to be repeated for a second cycle (known as "cycling") to ensure that all items in the set are able to serve as model predictors. This modified version of PMN simplifies the procedures for imputing categorical variables and eliminates the need for having a single donor for a set of variables.

5.1 Overview of modPMN-MI

Categorical variables imputed univariately,[24] such as all of the demographic variables involved in this evaluation, were handled using a simple model-based stochastic imputation (Procedure 1). Continuous variables imputed univariately, which did not occur in this evaluation but would occur if this method was applied to all National Survey on Drug Use and Health (NSDUH) variables that currently undergo imputation, would be handled using a simple model-based hot-deck imputation similar to PMN (Procedure 2).[25] Variables imputed multivariately, including all the drug measures, were handled using a combination of Procedures 1 and 2, depending on whether each variable in the set was categorical or continuous (Procedure 3). Procedure 3 also included a second cycle of imputation to ensure that each variable in the set had at least one chance to be a predictor in the imputation model for every other variable in the set. Variables imputed univariately in PMN were imputed univariately in modPMN-MI, and

[24] The 2011 imputation report of the NSDUH methodological resource book (Frechtel et al., 2013) draws a distinction between univariate matching and univariate assignment for PMN. "Univariate matching" means that only one predicted mean is used to measure the distance from the donors to the recipient. "Univariate assignment" means that the selected donor supplied values to the recipient for only one variable. For this methods evaluation, for PMN, "univariate" means "univariate assignment."

[25] A full description of Procedure 2 is included in this report because the steps in this procedure are followed when continuous variables are part of a variable set (Procedure 3). This occurred in this evaluation for drug frequency and age-at-first-use variables.

variables imputed multivariately in PMN were imputed multivariately in modPMN-MI. Table 5.1 presents the procedure that was used for each of the demographic and drug variables involved in this evaluation. It presents the variable and variable sets in the order in which they were imputed. Note that subsequent variables (as presented in the table) were used as predictors for prior variables.

Table 5.1 Order of Imputation by Variable and Procedure for modPMN-MI

Order	Group	Variable	Variable Type	Procedure
1	Demographics	Marital Status	Categorical	1
2	Demographics	Race	Categorical	1
3	Demographics	Hispanic/Latino Origin	Categorical	1
4	Demographics	Education Level	Categorical	1
5	Drugs	Lifetime Use (7 variables)	Categorical	3
6	Drugs	Recency (7 variables)	Categorical	3
7	Drugs	12-Month Frequency (6 variables)	Continuous	3
8	Drugs	30-Day Frequency (6 variables)	Continuous	3
9	Drugs	Age at First Use (7 variables)	Continuous	3

modPMN-MI = modified predictive mean neighborhood multiple imputation.

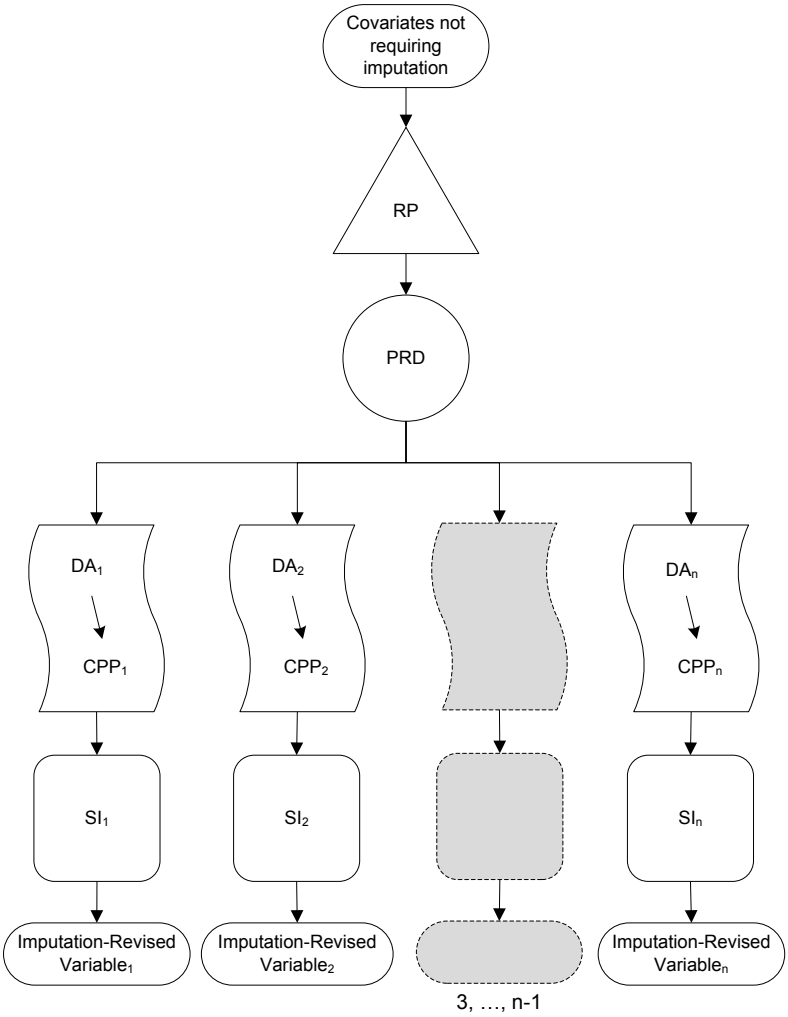
Note: The variable names are presented in Chapter 1 in Tables 1.1 and 1.2.

5.1.1 Categorical Variables, Univariate Framework: Procedures 0 and 1

Marital status was the first variable that was imputed in this evaluation. Only covariates that required no imputation were used in the models. Thus, although five imputations were completed, only one set of models needed to be fit. The steps for marital status are described below and graphically illustrated in Figure 5.1.

- **Step 0.1:** Use the generalized exponential model (GEM) (Frechtel et al., 2013, Section 2.3) to adjust the sampling weights for item nonresponse. (This step is RP in Figure 5.1, for "response propensity.")
- **Step 0.2:** Fit a polytomous logistic regression model. Save the parameter estimates and the variance-covariance matrix associated with the parameter estimates. (This step is PRD in Figure 5.1, for "prediction model.")
- **Step 0.3:** Draw a random vector of parameter estimates. Assume this random vector has a multivariate normal (MVN) distribution with the mean vector and variance-covariance matrix from Step 0.2. (This step is DA in Figure 5.1, for "data augmentation.")
- **Step 0.4:** For each item nonrespondent, use the vector drawn in Step 0.3 and the vector of predictors used in Step 0.2 to estimate the probability associated with each level of the outcome variable. (This step is CPP in Figure 5.1, for "calculate predicted probabilities.")
- **Step 0.5:** Randomly impute a value for each item nonrespondent using the probabilities from Step 0.4. (This step is SI in Figure 5.1, for "stochastic imputation.")
- **Step 0.6:** Repeat Steps 0.3 through 0.5 four times to produce a total of five imputed values for each item nonrespondent. The result is five proper imputations (Rubin, 1987) for marital status.

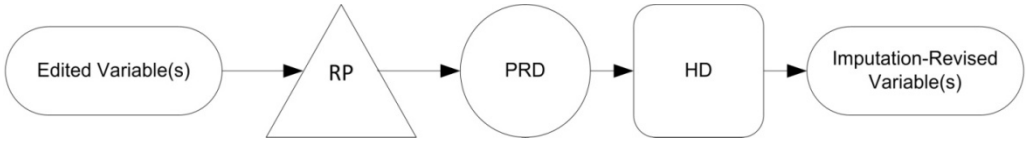
Figure 5.1 Procedure 0 of modPMN-MI



CPP = calculate predicted probabilities; DA = data augmentation; modPMN-MI = modified predictive mean neighborhood multiple imputation; PRD = prediction model; RP = response propensity; SI = stochastic imputation.

The differences between Procedure 0 and PMN are evident when compared with Figure 5.2, which shows the single response propensity/single prediction type of PMN. Note the insertion of the data augmentation step before the predicted means are calculated, the replacement of the hot-deck step with stochastic imputation, and the generation of five imputations instead of one.

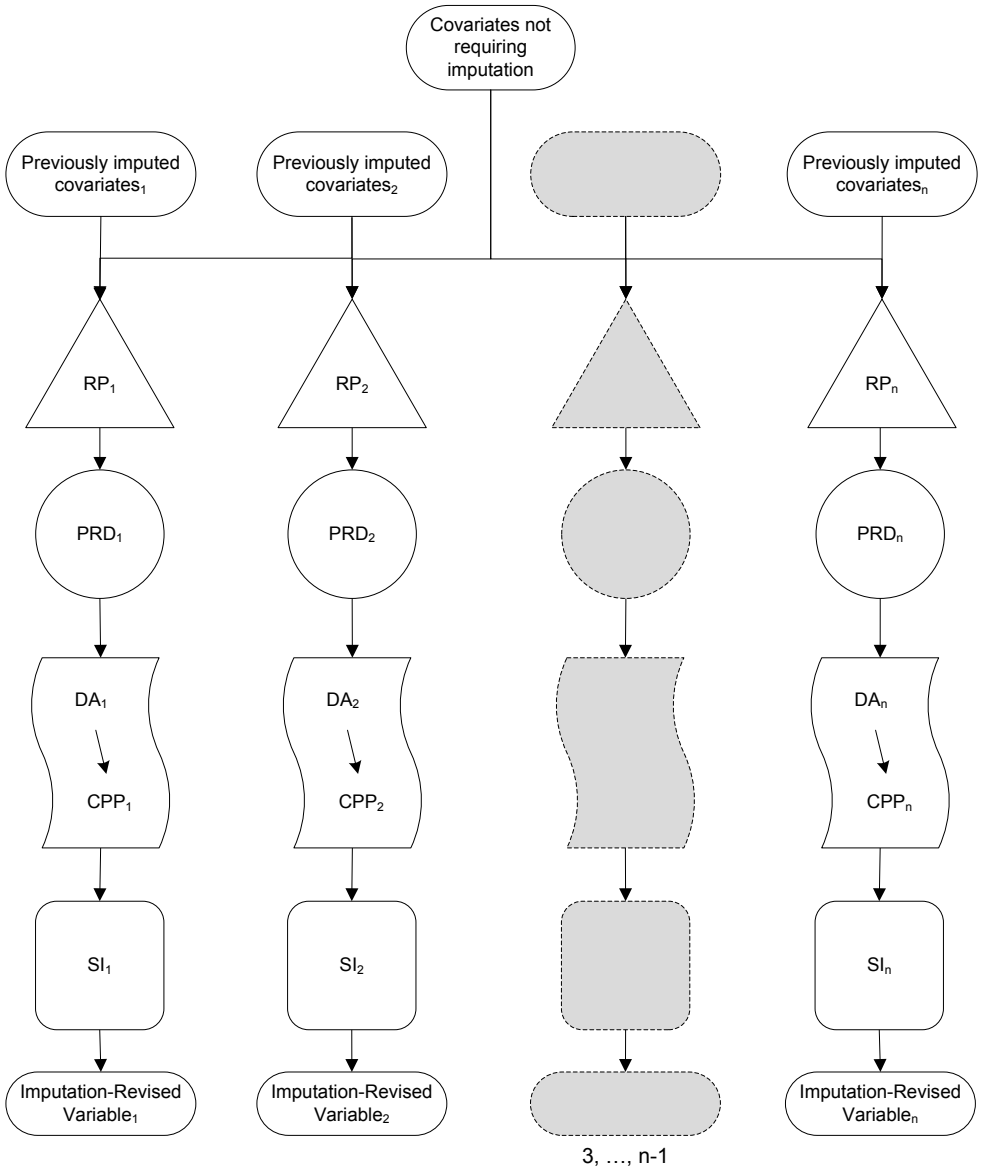
Figure 5.2 PMN Type 1: Single Response Propensity/Single Prediction



HD = hot deck; modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; PRD = prediction model; RP = response propensity.

For other categorical variables, a slightly different algorithm was used because variables that had already undergone imputation were used as predictor variables. For example, when the race variable was imputed, marital status was used as a predictor. For race and all subsequently imputed variables that used marital status as a predictor, the values already imputed for marital status were treated as known. These steps are graphically illustrated in Figure 5.3 and described below.

Figure 5.3 Procedure 1 of modPMN-MI



CPP = calculate predicted probabilities; DA = data augmentation; modPMN-MI = modified predictive mean neighborhood multiple imputation; PRD = prediction model; RP = response propensity; SI = stochastic imputation.

- **Step 1.1:** For Imputation 1 (i.e., first round of imputation out of five imputations[26]), use GEM to adjust weights for item nonresponse. For any predictor variables that have already undergone multiple imputations, use the imputed value from Imputation 1.
- **Step 1.2:** Fit a logistic regression model. For any predictor variables that have already undergone multiple imputations, use the imputed value from Imputation 1. Save the parameter estimates and the variance-covariance matrix associated with them.
- **Step 1.3:** Draw a random vector of parameter estimates. Assume this random vector has an MVN distribution with the mean vector and variance-covariance matrix from Step 1.2.
- **Step 1.4:** For each item nonrespondent, use the vector drawn in Step 1.3 and the vector of predictor variables used in Step 1.2 to estimate the probability associated with each level of the outcome variable.
- **Step 1.5:** Randomly impute a value for each item nonrespondent using the probabilities from Step 1.4.[27]
- **Step 1.6:** Repeat Steps 1.1 through 1.5 for Imputations 2 through 5 to produce a total of five imputed values for each item nonrespondent.

5.1.2 Continuous Variables, Univariate Framework: Procedure 2

Although this evaluation did not involve any continuous variables imputed univariately, it is useful to describe the steps that would be taken in that situation because the same steps were taken when continuous variables were imputed multivariately as in Procedure 3. The steps involved in such a situation are described below and graphically illustrated in Figure 5.4.

- **Step 2.1:** For Imputation 1, use GEM to adjust weights for item nonresponse. For any predictors that have already undergone multiple imputations, use the imputed value from Imputation 1.
- **Step 2.2:** Fit a linear regression model. For any predictors that have already undergone multiple imputations, use the imputed value from Imputation 1. Save the parameter estimates and the variance-covariance matrix associated with the parameter estimates.
- **Step 2.3:** Draw a random vector of parameter estimates. Assume this random vector has an MVN distribution with the mean vector and variance-covariance matrix from Step 2.2.
- **Step 2.4:** For item respondents and nonrespondents, use the vector drawn in Step 2.3 and the vector of predictors used in Step 2.2 to estimate the predicted mean for the outcome variable.
- **Step 2.5:** Implement the same hot-deck step as in PMN to randomly impute a value for each item nonrespondent, with the following enhancement: select the donor from the neighborhood with probability proportional to the adjusted weight instead of

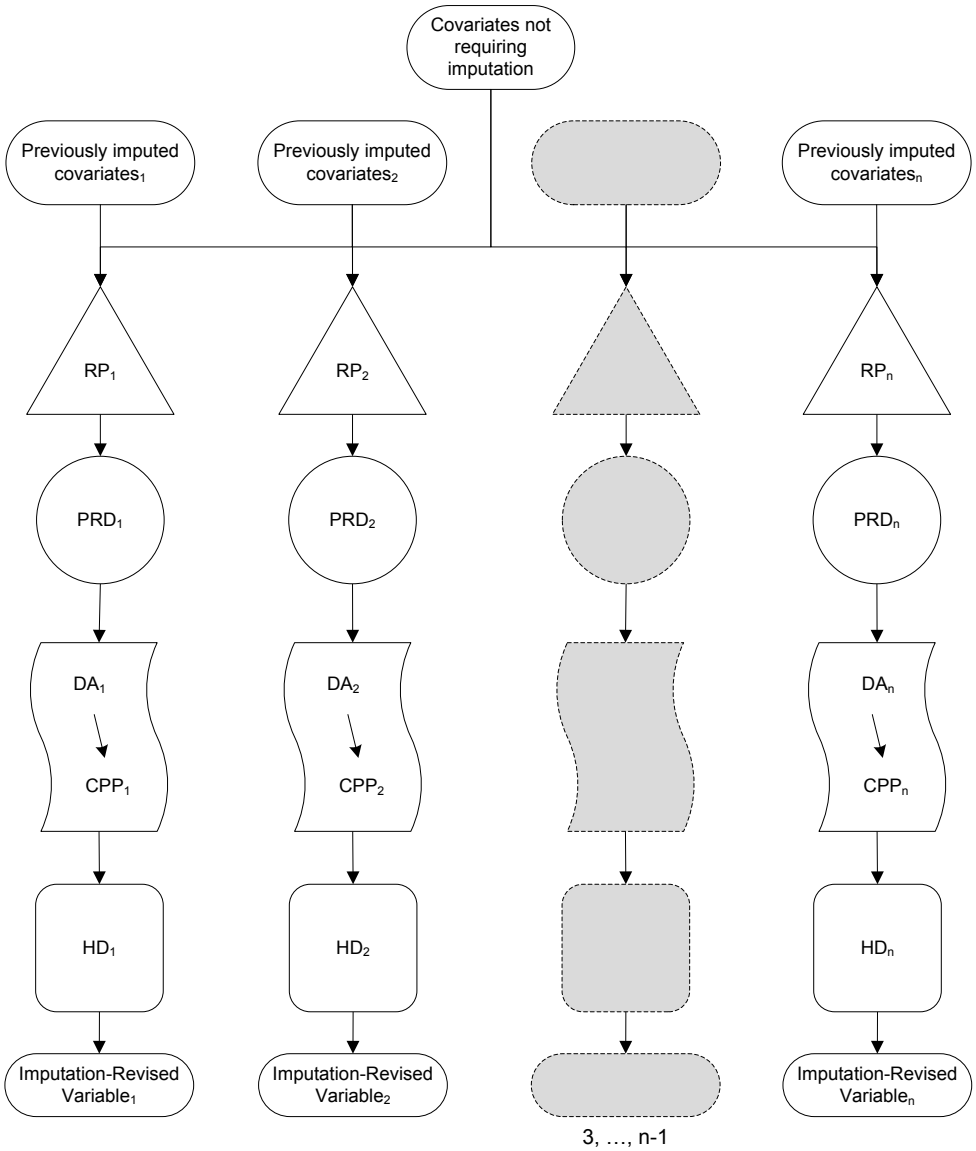
[26] Similar to IVEware, there were five imputations performed.

[27] In cases where logical constraints restrict the levels that are possible for a given item respondent, conditional probabilities were used. For example, some item nonrespondents for recency of drug use are known to be past year users, but it cannot be determined from their responses whether they are past month users. The probabilities used in Step 1.5 would then be conditional on item nonrespondents being past year users.

using the same selection probability for each donor. (This step is HD in Figure 5.4, for "hot deck.")

- **Step 2.6:** Repeat Steps 2.1 through 2.5 four times to produce a total of five imputed values for each item nonrespondent. For any predictors that have already undergone multiple imputations, continue to use the imputed value from Imputations 2 through 5.

Figure 5.4 Procedure 2 of modPMN-MI

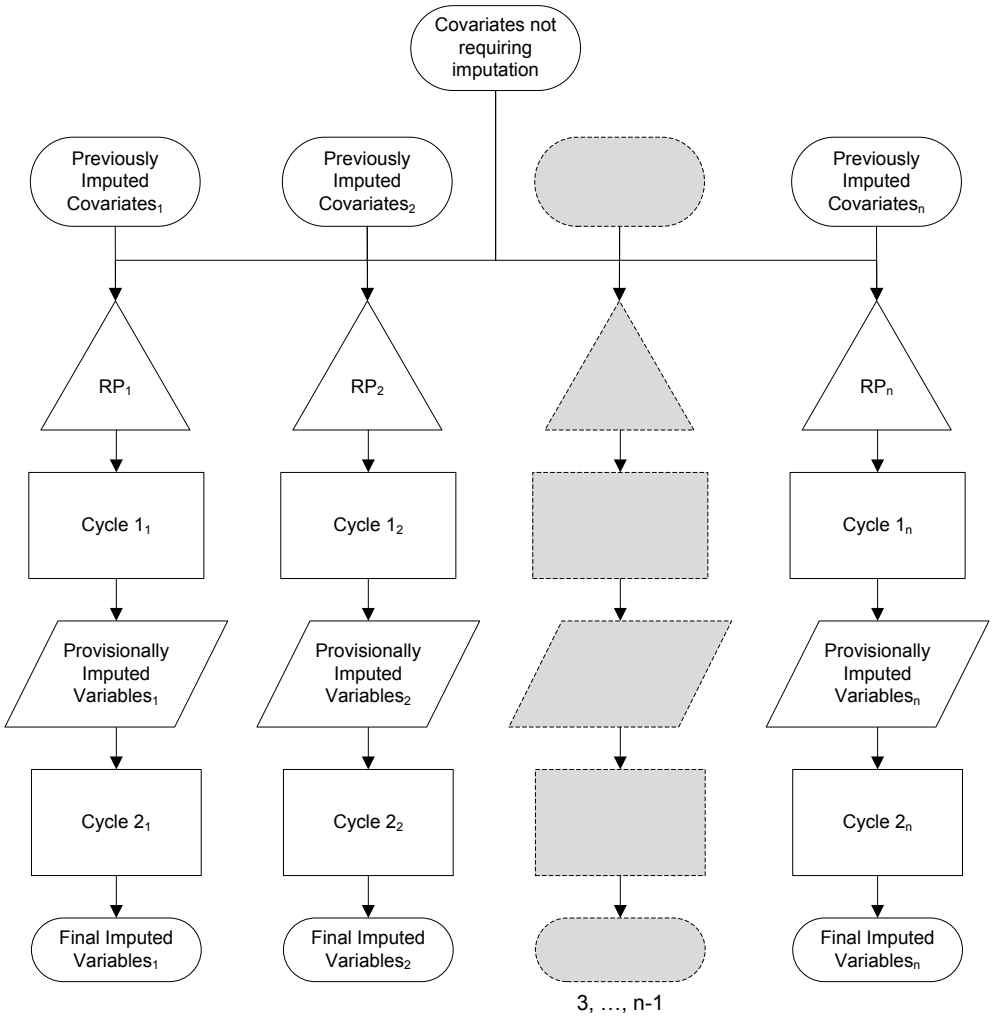


CPP = calculate predicted probabilities; DA = data augmentation; HD = hot deck; modPMN-MI = modified predictive mean neighborhood multiple imputation; PRD = prediction model; RP = response propensity.

5.1.3 Multivariate Framework: Procedure 3

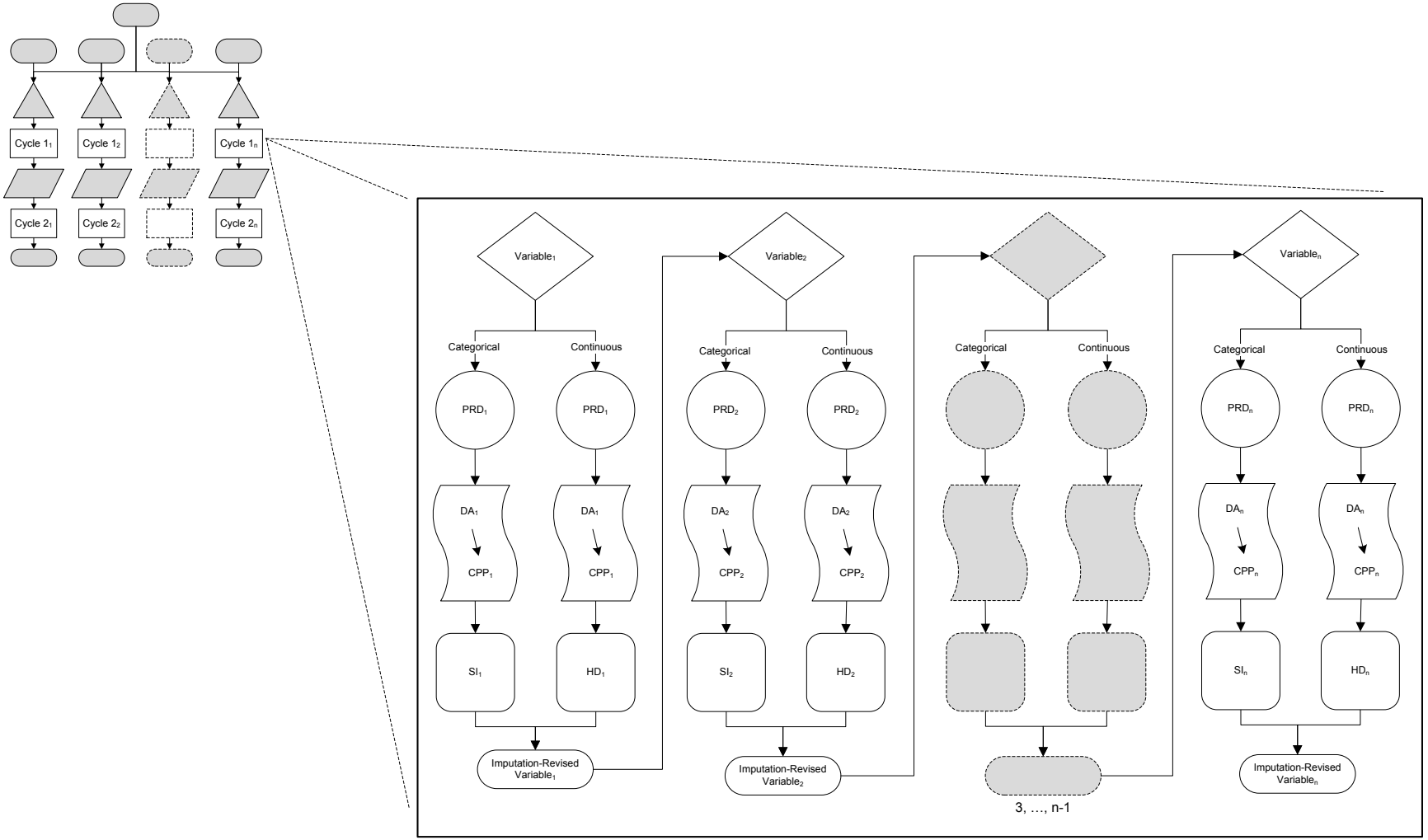
The steps taken for variables that were imputed multivariately (or part of a set), which included all drug variables involved in this evaluation, are described below. The general process is illustrated in Figure 5.5. A detailed illustration of a single cycle within a single imputation is shown in Figure 5.6.

Figure 5.5 Procedure 3 of modPMN-MI: Overview



modPMN-MI = modified predictive mean neighborhood multiple imputation; RP = response propensity.

Figure 5.6 Procedure 3 of modPMN-MI: Detailed Illustration of a Cycle



CPP = calculate predicted probabilities; DA = data augmentation; HD = hot deck; modPMN-MI = modified predictive mean neighborhood multiple imputation; PRD = prediction model; SI = stochastic imputation.

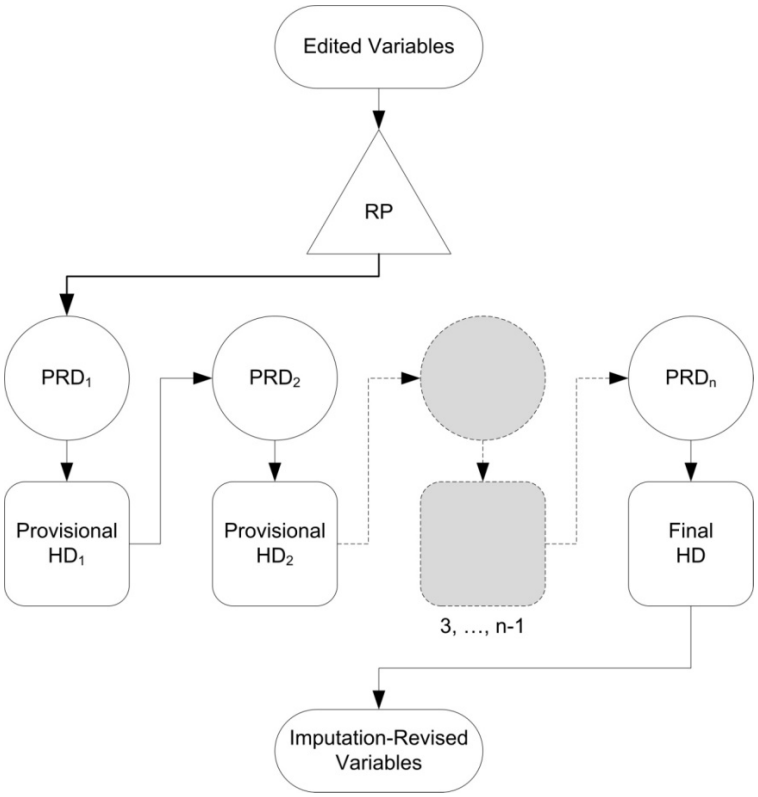
- **Step 3.1:** For Imputation 1, use GEM to adjust weights for item nonresponse. A unit respondent is in the domain of GEM if he or she is in the domain for any of the variables in the set. A unit respondent is considered an item respondent by GEM if he or she is an item respondent for all the variables in the set. For any predictors that have already undergone multiple imputations, use the imputed value from Imputation 1.[28]
- **Step 3.2:** Start with the first variable in the set. If it is categorical, follow Steps 1.2 through 1.5. If it is continuous, follow Steps 2.2 through 2.5. For the regression model, use a set of predictor variables similar to those used in PMN.
- **Step 3.3:** Repeat Step 3.2 for each variable in the variable set, following the sequence shown in Table 5.1. Again, use a set of predictor variables similar to those used in PMN, which are the same as the predictors used for the first variable in the sequence, plus the imputed versions of all variables in the set, which are imputed earlier in the sequence.
- **Step 3.4:** Repeat Steps 3.2 and 3.3 once (Cycle 2[29]), except enlarge the set of predictor variables to include imputed versions of all variables in the set other than the one being modeled. This follows the procedure used in IVEware (Raghunathan et al., 2001, p. 87). At the end of this step, the first imputation is complete.
- **Step 3.5:** Repeat Steps 3.1 through 3.4 four times to produce a total of five imputed values for each variable in the variable set for each item nonrespondent.

The differences between Procedure 3 and PMN are evident when compared with Figure 5.7, which shows the single response propensity/multiple prediction type of PMN. Note the inclusion of an additional cycle and the replacement of a final hot-deck step involving multivariate assignment with final imputations involving univariate assignment.

[28] This shortcut approach was not used for all drug sets in modPMN-MI. See Section 4.4.4.

[29] "Cycling" is the process by which variables later in the hierarchy aid in the imputation of variables earlier in the hierarchy. IVEware and modPMN-MI both use cycling.

Figure 5.7 PMN Type 3: Single Response Propensity/Multiple Prediction



HD = hot deck; PMN = predictive mean neighborhood; PRD = prediction model; RP = response propensity.

5.1.4 Detailed Descriptions of Procedures for modPMN-MI

5.1.4.1 Response Propensity Modeling

The first step in modPMN-MI, regardless of the type of procedure that was followed, was to perform a response propensity modeling step similar to the steps followed in PMN. These steps are noted as 0.1, 1.1, and 2.1 in Sections 5.1.1 and 5.1.2. However, there were a few modifications that were implemented for modPMN-MI. The first modification relates to how the adjustment factor calculated in GEM is applied to the design weight to form the response propensity-adjusted weight. In both PMN and modPMN-MI, the preliminary analytic weights are adjusted for item nonresponse before predictive modeling using GEM. However, the adjustment factor differs. In PMN, the adjustment factor is $1/p_k$, where p_k is the estimated probability of item response from GEM for the k^{th} item respondent. In modPMN-MI, the factor is $(1 - p_k)/p_k$. Appendix C describes the justification for this modification.

There were two additional modifications made to the response propensity modeling step for modPMN-MI as it compares with PMN for the drug variables: the inclusion of additional interaction terms in regression models and the aggregation across age groups for variables with small domains. The former is an improvement on PMN because it can better handle the case where the actual relationship between the propensity to respond and the predictors varies from variable to variable within the set of outcome variables. The latter is an improvement on PMN

because the response propensity model coefficients can be estimated with reduced variance when the number of observations used to fit the model is not so small.

As in PMN, in the multivariate case, a shortcut approach was often used where a single item response propensity model is applied to the whole set of variables of interest. The domain of the model included all unit respondents who were in the domain for any of the variables of interest. Unit respondents in the domain were considered item respondents if they had complete data across all variables of interest. Section 5.3.1 provides additional details on how the response propensity models were applied to groups of variables for certain drug measures.

5.1.4.2 Predictive Mean Modeling

In modPMN-MI, categorical outcome variables were modeled using logistic regression, and continuous outcome variables were modeled using linear regression. The sample design was taken into account through the use of SUDAAN, and the weights were adjusted for the response propensity. As stated in the previous section, the adjustment factor was different in modPMN-MI than in PMN.

The set of predictors used in modPMN-MI was similar to the set of predictors used in PMN. The strategy for selecting variables for an imputation model is somewhat different from the traditional strategy for selecting variables for analytical models. In the traditional approach, a variable is not included in an analytical model unless it is deemed significant, which places a high priority on model parsimony and model interpretability. For imputation models, by contrast, the primary goal is prediction, where imputation models with a large number of variables are often preferred over models with a smaller number. Some general guidelines for building an imputation model are discussed by Rubin (1996) that recommend the following categories of variables should be included: (1) variables that are considered important reporting variables (e.g., subdomains such as age and race), (2) variables that convey essential information about the complex sample design, and (3) variables that have large coefficients and large standard deviations. Including the last category may be important for creating multiple imputations with enough variability to reflect the actual uncertainty about the missing values.

Because modPMN-MI has the additional goal of estimating variance due to imputation, unlike PMN, the inclusion of the last category of variables is especially important for modPMN-MI. As a result, the models tended to be larger under modPMN-MI than they were under PMN. Certain warning messages are generated by SUDAAN under PMN, and most of these messages were ignored in modPMN-MI in an effort to include that last category of variables in the models.

Obtaining accurate standard errors of model coefficients was more important in modPMN-MI than it was in PMN because the standard errors were used in the next step (the data augmentation step). In PMN, only the predicted means are used. Because of this, an effort was made to reduce multicollinearity that occurred when powers of the age variable (i.e., age-squared or age-cubed term in the regression model) were used. Instead of powers of the age variable, age categories were formed from the individual age variable to help reduce multicollinearity.

5.1.4.3 Data Augmentation

Steps 0.3, 1.3, and 2.3 (Sections 5.1.1 and 5.1.2), which require a random beta coefficient vector to be drawn from an MVN distribution, are called "data augmentation steps" in MI literature. The data augmentation step makes modPMN-MI "proper" (Rubin, 1987) by accounting for the variance from the model-fitting step. For reasons described above, parsimony is not an important objective in the building of imputation models. The result is that over-fitted models are often developed. Occasionally, if the regression model was over-fitted, the SAS® software was not able to draw a random vector from an MVN distribution. Instead, it would report that the variance-covariance matrix is not positive definite. In these cases, the eigenvalues of the variance-covariance matrix were "corrected" to allow the SAS software to operate correctly. Appendix C presents details on how this correction was implemented.

The data augmentation step can be omitted if multiple imputations are not performed. However, the amount of variance inflation due to imputation would not be captured.

5.1.4.4 Calculation of Predicted Means

Predicted means for all unit respondents in the domain were calculated by multiplying the predictor matrix from the regression model by the "augmented" beta coefficient vector from the step described in section 5.1.4.3. Note that predicted means were calculated for both item respondents and item nonrespondents. If the outcome variable is categorical, the predicted means for item nonrespondents are used in a stochastic imputation step described in the next section (and the predicted means of item respondents are unused). If the outcome variable is continuous, the predicted means are used to map item respondents to each item nonrespondent based on the closeness of the predicted means, as described in Section 5.1.4.6. This is unlike PMN, where predicted means come directly from the model and the predictor matrix is simply multiplied by the regression coefficients as estimated by the model.

5.1.4.5 Imputation of Categorical Variables

In PMN, a hot-deck step is used to determine the final imputed values for categorical variables. In contrast, categorical variables are imputed stochastically (i.e., using the predicted values from regression models) in modPMN-MI. This guarantees that, over repeated imputations, the categorical probability distribution of the imputed values matches the model predictions. Adding the likeness constraints and making equal probability donor selections from a restricted neighborhood changes the distribution of possible imputes. However, it may not make estimates less biased, especially in cases where the likeness variables can be included in the predictive mean model. Additionally, imputing categorical variables stochastically can be considered a cost-saving measure, primarily because the stochastic imputation approach is much simpler and takes much less time than a hot-deck step, which requires decisions on likeness constraints and the order in which they need to be loosened.

One drawback of the stochastic imputation approach for categorical variables is that likeness constraints that are not covered by the model are not used in modPMN-MI, which may result in somewhat less accurate imputation results. Likeness constraints allow more flexibility than a model and can be loosened in any order, making it likely that recipients are matched to the

most similar donors available. Table 5.2 shows the likeness constraints used in PMN that are not covered by modPMN-MI for all categorical variables involved in this evaluation. One likeness constraint included in every hot-deck step is the "delta constraint," which ensures that the donor's predicted means are each within 5 percent of the recipient's predicted means. This constraint is purely model based, and therefore it is considered to be covered by stochastic imputation.

Logical constraints for categorical variables are covered in modPMN-MI using conditional probabilities. For example, if the outcome variable has four levels, but it is known that the item nonrespondent must have a final imputed value of 3 or 4, then the item respondent receives a final imputed value of 3 with probability $\frac{\hat{p}_3}{\hat{p}_3 + \hat{p}_4}$ and a final imputed value of 4 with probability $\frac{\hat{p}_4}{\hat{p}_3 + \hat{p}_4}$.

Table 5.2 PMN Likeness Constraints for Categorical Variables Not Covered in modPMN-MI

Variable	Likeness Constraints Not Covered by Regression Model	Comments
Marital Status	The donor's age must be within 3 years of the recipient's age.	The variables AGE, AGE^2, and AGE^3 are predictors in the model.
Race	If the recipient was Hispanic/Latino nonspecific, then the donor must be of Hispanic/Latino origin.	The segment-level variable for Hispanic/Latino origin concentration is a predictor in the model.
	If the recipient selected one or more Hispanic/Latino categories, including Mexican, Puerto Rican, Central or South American, Cuban, Dominican, and Spaniard, then the donor's Hispanic/Latino group value must be equal to one of the Hispanic/Latino groups mentioned by the recipient.	
	The donor must be Mexican (Hispanic/Latino or non-Hispanic/Latino).	
	The donor must be Cuban (Hispanic/Latino or non-Hispanic/Latino).	
	The donor must be Central or South American (Hispanic/Latino or non-Hispanic/Latino).	
	The donor must be Dominican (Hispanic/Latino or non-Hispanic/Latino).	
	The donor must be Spanish (Hispanic/Latino or non-Hispanic/Latino).	
Hispanic/Latino Origin	The segment of the donor must equal the segment of recipient.	The segment-level variables for Asian/Other Pacific Islander, Black/African American, American Indian/Alaska Native, Hispanic/Latino, and Owner Occupied concentration are predictors in the model.

Table 5.2 PMN Likeness Constraints for Categorical Variables Not Covered in modPMN-MI (continued)

Variable	Likeness Constraints Not Covered by Regression Model	Comments
Education Level	The segment of the donor must equal the segment of the recipient.	The segment-level variables for Asian/Other Pacific Islander, Black/African American, American Indian/Alaska Native, Hispanic/Latino, and Owner Occupied concentration are predictors in the model.
	The age of the donor must equal the age of the recipient.	The variables AGE, AGE^2, and AGE^3 are predictors in the model.
Lifetime Use	The state rank of the donor must equal the state rank of the recipient.	The state rank variable is a predictor in the model.
	The lifetime use of the donor must equal the lifetime use of the recipient for each nonmissing lifetime indicator.	The provisionally imputed lifetime drug use indicators are predictors in the model.
	If the recipient was missing the lifetime indicator(s) for any member of a family of drugs, then the donor's lifetime indicator(s) must agree with the recipient's nonmissing lifetime indicator(s) within that family.	
Recency	The state rank of the donor must equal the state rank of the recipient.	The state rank variable is a predictor in the model.

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood.

5.1.4.6 Imputation of Continuous Variables

For continuous variables, a hot-deck step is used for both PMN and modPMN-MI and is preferred to the stochastic imputation step because most continuous variables that are imputed in the NSDUH cannot reasonably be approximated by a normal distribution. The hot-deck donor selection of near neighbors, where "nearness" is defined by the distance from the donor's predicted mean to the recipient's predicted mean, has the potential to characterize the conditional distributions much better than a normal residual. For categorical variables, it is reasonable to assume that the missing items have a multinomial distribution based on the predicted probabilities associated with each outcome level.

However, one simple modification was introduced in modPMN-MI. In PMN, each donor in the neighborhood has an equal probability of being selected as the final donor. In modPMN-MI, the donor was selected from the neighborhood with probability proportional to the GEM-adjusted weight. See Appendix C for additional information on this approach.

Note that the problem with likeness constraints that occurs for categorical variables, described in Section 5.1.4.5, does not occur for continuous variables. This is because continuous variables are imputed in a hot-deck step that easily incorporates both likeness and logical constraints.

5.1.4.7 Cycling

In PMN, only one cycle is completed for each variable in a set. This makes the order in which the imputations are completed very important; that is, variables later in the sequence do not have a chance to be predictors for variables earlier in the sequence. In modPMN-MI, a second cycle is completed for all drug variables involved in this evaluation (Step 3.4 in Section 5.1.3), which uses imputations from the first cycle as predictors in the model. This process fine-tunes the predicted values for the drugs earlier in the sequence. Cycling was performed for drug usage variables where the cycling was done across drugs but within each drug measure (e.g., lifetime drug use, recency of drug use, frequency of drug use, and age at first use). In other words, all of the lifetime usage variables were imputed first, recency was imputed second, frequency was imputed third, and age at first use was imputed last.

Another difference between PMN and modPMN-MI is the procession from "intermediate" imputations to "final" imputations. In PMN, once models have been fit for all variables in the set and intermediate imputations are complete for each of them, the predicted values are used in a final hot-deck step in which the predicted values for all variables in the set are used to determine the neighborhood. The selected donor supplies values to the recipient for all missing variables in the set. By contrast, in modPMN-MI, there is no final hot-deck step where a single donor is used for all missing variables. Once the second cycle is complete, all imputations are complete. For example, if there are 10 continuous variables in the set and a recipient is missing all 10 of them, he or she may have as many as 10 different donors.

5.2 Imputation of Demographic Variables

All of the demographic variables involved in this evaluation were categorical variables that were imputed univariately using a simple model-based stochastic imputation. Similar to PMN, the four demographic variables were imputed in the following order: marital status, race, Hispanic/Latino origin, and education level. Steps 0.1 through 0.6 were followed for marital status and Steps 1.1 through 1.6 were followed for the other three variables. The predictors used in each model are summarized in Tables A.22 through A.25 in Appendix A.

5.3 Imputation of Drug Variables

The implementation of modPMN-MI for the drug variables differed from the implementation for the demographic variables for two reasons. First, three of the five drug usage measures involved in this evaluation were count or continuous variables, not categorical variables: 12-month frequency, 30-day frequency, and age at first use. Second, the drug variables were imputed as part of a set, instead of univariately, where the variable set was defined as all drugs for the given usage measure (e.g., recency, frequency, and age at first use). This was done to best exploit the correlations between the measures across drugs.

In modPMN-MI, the grouping was done across drugs and within measure for all measures, not just lifetime use (Table 5.3). This was done for the following reasons:

- When cycling through the variables in the set, all variables have a chance to be predictors for all other variables in the second cycle. This would be awkward if the

grouping was the same as for PMN because the recency and frequency variables are related in a hierarchical manner. For example, consider a respondent who is known to be a lifetime user of alcohol, but recency, 12-month frequency, and 30-day frequency are all missing. If the recency is imputed to past month use in the first cycle, then logically the 12-month frequency must be on the interval from 1 to 365 days and the 30-day frequency must be on the interval from 1 to 30 days. In the second cycle, if the two frequencies are used as predictors, then the respondent must logically remain a past month user. Grouping the variables within measure but across drugs forms sets of variables that are correlated but are not hierarchically related.

- For the continuous variables (12-month frequency, 30-day frequency, and age at first use), the logical constraints are simpler using this approach. The logical constraints in PMN involve complex interactions when more than one measure is missing.

For each set, Steps 3.1 through 3.5 were followed.

Table 5.3 Grouping of Drug Variables into Imputation Sets in modPMN-MI

Drug Variable	Lifetime Use	Recency	12-Month Frequency	30-Day Frequency	Age at First Use
Cigarettes Alcohol Inhalants Marijuana Pain Relievers Cocaine Heroin	Set 1	Set 2	Set 3 (cigarettes N/A)	Set 4 (pain relievers N/A)	Set 5

modPMN-MI = modified predictive mean neighborhood multiple imputation; N/A = not applicable.

In general, the sequence in which the drug variables are imputed within each variable set goes from the more common drugs (e.g., cigarettes) to the more rare drugs (e.g., heroin). Table 5.4 shows the sequence of imputation of the drugs and drug usage measures.

Table 5.4 Sequence of Imputation of Drug Variables, Within Each Variable Set for modPMN-MI

Drug Variable (Set)	Drugs (in order of imputation, within each set)						
	Cigarettes	Alcohol	Inhalants	Marijuana	Pain Relievers	Cocaine	Heroin
Lifetime Use	1	2	3	4	5	6	7
Recency	1	2	3	4	5	6	7
12-Month Frequency	N/A	1	2	3	4	5	6
30-Day Frequency	1	2	3	4	N/A	5	6
Age at First Use	1	2	3	4	5	6	7

modPMN-MI = modified predictive mean neighborhood multiple imputation; N/A = not applicable.

There is some concern about whether the 12-month frequency, 30-day frequency, and age-at-first-use variables are correlated strongly enough across drugs to justify grouping them into imputation sets. The correlation coefficients for every pair of drugs in each imputation set within the age group of 18 to 25 are shown in Tables 5.5, 5.6, and 5.7. (The tables for the other two age groups, 12 to 17 and 26 or older, look similar.) Notice that the tables for 12-month and 30-day frequency are not symmetric. This is because nonusers were handled differently for the rows and columns in these tables. In Table 5.5, respondents who had not used the drug in the last 12 months were not included in the row domain. However, they were included in the column domain but given a frequency of 0. Thus, the value of 0.07 in the second column of the first row can be interpreted as the correlation between the 12-month frequencies of alcohol and inhalants, given that the person used alcohol in the past 12 months. In contrast, the value of 0.18 in the first column of the second row would be interpreted as the correlation between the 12-month frequencies of alcohol and inhalants, given that the person used inhalants in the past 12 months. The last column shows the number of observations in the domain for each row.

Table 5.5 Correlation Coefficients for 12-Month Frequency Variables, for Respondents Aged 18 to 25

	Alcohol	Inhalants	Marijuana	Pain Relievers	Cocaine	Heroin	Total[1]
Alcohol	1.00	0.07	0.24	0.10	0.10	0.03	18,300
Inhalants	0.18	1.00	0.01	0.33	0.20	0.29	400
Marijuana	0.16	0.04	1.00	0.16	0.11	0.04	7,000
Pain Relievers	0.03	0.14	0.15	1.00	0.19	0.18	2,500
Cocaine	0.00	0.11	-0.02	0.15	1.00	0.37	1,000
Heroin	0.03	0.34	-0.07	0.00	0.52	1.00	100

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table 5.6 Correlation Coefficients for 30-Day Frequency Variables, for Respondents Aged 18 to 25

	Cigarettes	Alcohol	Inhalants	Marijuana	Cocaine	Heroin	Total[1]
Cigarettes	1.00	-0.02	0.00	0.11	0.04	0.04	8,200
Alcohol	0.17	1.00	0.06	0.23	0.11	0.02	14,300
Inhalants	0.02	0.35	1.00	0.18	0.01	0.03	100
Marijuana	0.22	0.12	0.03	1.00	0.05	0.06	4,300
Cocaine	0.26	0.10	0.08	0.01	1.00	0.43	300
Heroin	0.35	0.00	0.36	0.29	0.51	1.00	< 50

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

In contrast to Tables 5.5 and 5.6, Table 5.7 is symmetric. To be in the domain for a particular cell, a respondent had to be a lifetime user of both the drug corresponding to the row and the drug corresponding to the column. The number of respondents in the domain for each cell is listed in parentheses under the correlation coefficient, and this has been rounded to the nearest hundred to ensure respondent confidentiality.

Table 5.7 Correlation Coefficients for Age-at-First-Use Variables, for Respondents Aged 18 to 25

	Cigarettes	Alcohol	Inhalants	Marijuana	Pain Relievers	Cocaine	Heroin
Cigarettes	1.00 (14,600)	0.46 (14,000)	0.35 (2,000)	0.55 (10,400)	0.32 (4,800)	0.35 (3,000)	0.17 (400)
Alcohol	0.46 (14,000)	1.00 (19,900)	0.24 (2,200)	0.47 (11,800)	0.29 (5,300)	0.29 (3,000)	0.21 (400)
Inhalants	0.35 (2,000)	0.24 (2,200)	1.00 (2,200)	0.37 (2,000)	0.43 (1,500)	0.46 (1,100)	0.48 (200)
Marijuana	0.55 (10,400)	0.47 (11,800)	0.37 (2,000)	1.00 (12,000)	0.40 (4,600)	0.47 (3,000)	0.42 (400)
Pain Relievers	0.32 (4,800)	0.29 (5,300)	0.43 (1,500)	0.40 (4,600)	1.00 (5,400)	0.49 (2,200)	0.54 (400)
Cocaine	0.35 (3,000)	0.29 (3,000)	0.46 (1,100)	0.47 (3,000)	0.49 (2,200)	1.00 (3,100)	0.55 (400)
Heroin	0.17 (400)	0.21 (400)	0.48 (200)	0.42 (400)	0.54 (400)	0.55 (400)	1.00 (400)

Note: Counts have been rounded to the nearest hundred to ensure respondent confidentiality.
Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

In the tables above, there is no clear relationship between the 12-month and 30-day frequencies across drugs. However, there appear to be strong positive correlations between the age-at-first-use variables across all drugs. Because grouping the drugs into larger imputation sets is more labor intensive than imputing them univariately, and the frequency variables do not appear to be as strongly correlated across drugs, there does not appear to be much benefit to grouping the 12-month and 30-day frequency variables into imputation sets. There does, however, appear to be a benefit to imputing the age-at-first-use variables together.

For variables that require multivariate imputation in PMN, a shortcut approach for fitting the response propensity model is often used where a single model is applied to the whole variable set. This shortcut approach saves time by reducing the number of models that need to be fit, but it has some drawbacks. First, the actual relationship between the propensity to respond and the predictors may vary from variable to variable within the set, but the shortcut approach is too basic to account for that. Second, fewer observations are used to fit some of the predictive mean models because unit respondents who are item respondents for some (but not all) of the variables in the set are not used to fit any of the models. The use of additional observations would presumably result in a better, more robust model.

The first drawback can be circumvented by including more variable interactions in the response propensity model, and this was the second modification to the response propensity modeling step that was tested in modPMN-MI. Specifically, the extra interactions should be from each domain indicator variable crossed with each predictor variable.[30] This modification

[30] For example, for 12-month frequencies other than heroin, a single response propensity model was fit for the age group of 12 to 17 (Models 2 through 4 in Table 5.8). The predictors were the 5 domain indicators (past year usage of each of the five drugs), 9 demographic variables, and the 45 two-way interactions of the domain indicators and the demographic variables (see Table A.26).

was implemented for 12-month frequency, 30-day frequency, and age at first use but not for lifetime or recency. In effect, this approach results in a separate model within each domain. However, some of these extra variable interactions (and sometimes other predictor variables) were dropped from the model because of convergence problems. There were no modifications made in modPMN-MI to address the second drawback.

Table 5.8 describes the shortcut response propensity models for 12-month frequency and 30-day frequency of drug usage measures used for modPMN-MI. Different models were used based on the drug and age group. For lifetime, recency, and age at first use, shortcut models were used for all drugs within each age group, as is done frequently in PMN. Section 2.5.3 discusses the shortcut approach used in PMN.

To help explain the idea of a shortcut response propensity model, some brief examples are presented here. For 12-month frequency, the domain for Model 1 included all unit respondents who were past year users of heroin, regardless of age group. Members of the domain were considered item respondents if they reported a valid 12-month frequency for heroin. The domain for Model 4 included all unit respondents who were 26 or older and were past year users for any subset of alcohol, inhalants, marijuana, pain relievers, and cocaine. Members of the domain were considered item respondents if they reported valid 12-month frequencies for all of the five drugs for which they were past year users. Tables A.26 through A.28 present the response propensity model summaries for the marijuana drug usage variables.

Table 5.8 Response Propensity Models for 12-Month Frequency and 30-Day Frequency for modPMN-MI

Drug Measure	Drug Variable	Age Group		
		12-17	18-25	26+
12-Month Frequency	Heroin	Model 1		
	Alcohol	Model 2	Model 3	Model 4
	Inhalants			
	Marijuana			
	Pain Relievers			
	Cocaine			
30-Day Frequency	Inhalants	Model 1		
	Cocaine	Model 2		
	Heroin	Model 3		
	Cigarettes	Model 4	Model 5	Model 6
	Alcohol			
	Marijuana			

modPMN-MI = modified predictive mean neighborhood multiple imputation.

The second modification for the response propensity modeling step for modPMN-MI involved fitting the models, regardless of age group, when the number of observations used to fit the model was small (e.g., fewer than 100). In other words, a single response propensity model was fit for the entire sample (those aged 12 or older). This approach was taken for a few drug variables with small domains for 30-day frequency and 12-month frequency. These measures are especially subject to small domain size because of the scarcity of recent users for certain combinations of drugs and age groups. Table 5.9 shows the variable domain size (i.e., the

number of past year and past month users) for each drug for 12-month frequency and 30-day frequency. Because of small domain sizes, age groups were aggregated for heroin for 12-month frequency and for inhalants, cocaine, and heroin for 30-day frequency.

Table 5.9 Variable Domain Sizes, by Drug and Age Group for 12-Month Frequency and 30-Day Frequency for modPMN-MI

Drug Measure	Drug Variable	Age Group		
		12-17	18-25	26+
12-Month Frequency	Alcohol	7,300	17,400	16,800
	Inhalants	900	400	< 100*
	Marijuana	3,000	6,100	2,000
	Pain Relievers	1,500	2,600	1,000
	Cocaine	300	1,400	500
	Heroin	< 50*	< 100*	< 50*
30-Day Frequency	Cigarettes	2,300	8,300	6,400
	Alcohol	3,700	13,600	13,300
	Inhalants	300	< 100*	< 50*
	Marijuana	1,600	3,600	1,200
	Cocaine	< 100*	400	200
	Heroin	< 50*	< 50*	< 50*

modPMN-MI = modified predictive mean neighborhood multiple imputation.

Note: Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

*These domains had fewer than 100 members, so a single response propensity model was fit across the three age groups.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

In summary, the response propensity step for modPMN-MI differs from PMN in the use of predictor variables, which includes interactions with domain indicator variables such as alcohol recency by age. Additionally, modPMN-MI aggregates across age groups for small domains, whereas PMN maintains separate age groups. The response propensity models for modPMN-MI are very similar to their PMN counterparts. In fact, the PMN models were the starting point for the modPMN-MI models. Tables A.26 through A.28 present both the response propensity and predictive mean model summaries for the marijuana usage variables.

5.4 Comparison of modPMN-MI with PMN

5.4.1 Summary of Statistical Tests Comparing Estimates Based on PMN with Estimates Based on modPMN-MI

Similar to the analysis presented in Chapters 3 and 4, the estimates based on PMN were compared with estimates based on modPMN-MI. Tables 5.10 and 5.11 present the results of these comparisons along with the weighted percentage of imputed data. Although PMN and modPMN-MI used different imputation algorithms, statistically there were not many significant differences between the estimates. Race was the only significant variable (Table 5.10) among the demographic variables. For the drug variables, alcohol 12-month frequency and 30-day frequency (Table 5.11) were the only significant variables. The percentages imputed for these alcohol frequency variables were higher (2.12 percent for 12-month frequency and 1.06 percent for 30-day frequency) than for any other drug frequency variables, proving a potential

explanation for why significant differences were found for these variables and not the rest. As noted in previous chapters, the magnitude of the differences is relatively small. As shown in Table F.9 in Appendix F, the PMN estimate for alcohol 12-month frequency is 86.9 days as compared with the modPMN-MI estimate of 86.7 days, and the difference in the means is only 0.2 days. Section 6.2 discusses the meaningfulness of these significant differences by examining bias ratios and confidence intervals for coverage probabilities. Tables F.1 through F.8 present the comparisons for each level of the demographic variables and note the significant differences. Tables F.9 through F.16 present the comparisons for the drug variables and note the significant differences.

Similar to IVEware, these differences may be contributed to the different set of predictor variables used and the cyclical nature of modPMN-MI. It should be noted that there were no significant differences in the recency variables. The lack of significant differences supports using the predicted values from the regression models as the final imputed values, as compared with PMN where donor values are used for the final imputed values.

Table 5.10 Comparisons of PMN and modPMN-MI Imputed Estimates for Demographic and Drug Recency Variables

Variable	Number of Categories	Weighted Percentage Imputed	P-Value for Chi-square Test of Interaction with Method
			PMN vs. modPMN-MI
Demographics			
Marital Status	4	0.03	0.8669
Hispanic/Latino Origin	2	0.05	0.3194
Race	4	2.52	< 0.0001
Education Level	4	0.05	0.4402
Recency			
Cigarettes	5	0.27	0.7922
Alcohol	4	0.90	0.2200
Inhalants	4	0.30	0.7185
Marijuana	4	0.39	0.5671
Pain Relievers	4	0.63	0.3387
Cocaine	4	0.33	0.3087
Heroin	4	0.04	0.5984

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood.

Note: The weighted percentage imputed is based on the 2007 imputation indicators for cases noted as logically assigned or statistically imputed.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table 5.11 Comparisons of PMN and modPMN-MI Imputed Estimates for Frequency and Age-at-First-Use Variables

Variable	Weighted Percentage Imputed	PMN Estimate[1]	Difference of Means: PMN vs. modPMN-MI[1]
30-Day Frequency for Past Month Users			
Cigarettes	0.20	22.6	0.0
Alcohol	1.06	8.4	0.0[a]
Inhalants	0.04	4.0	-0.1
Marijuana	0.16	12.9	0.0
Cocaine	0.12	6.0	0.1
Heroin	0.00	15.5	1.6
12-Month Frequency for Past Year Users			
Alcohol	2.12	86.9	0.2[a]
Inhalants	0.16	28.6	-0.1
Marijuana	0.71	101.9	-0.3
Pain Relievers	0.51	46.2	2.0
Cocaine	0.36	43.3	0.3
Heroin	0.03	92.8	-0.1
Age at First Use for Lifetime Users			
Cigarettes	0.66	15.7	0.0
Alcohol	1.21	17.0	0.0
Inhalants	0.51	17.3	0.0
Marijuana	0.28	18.0	0.0
Pain Relievers	0.99	22.1	0.0
Cocaine	0.33	21.9	0.0
Heroin	0.02	22.9	0.0

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood.

Note: The weighted percentage imputed is based on the 2007 imputation indicators for cases noted as logically assigned or statistically imputed.

[1] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

[a] Difference is statistically significant at the 0.05 level.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

5.4.2 Differences and Similarities between PMN and modPMN-MI

This section attempts to quantify the similarities and differences between PMN and modPMN-MI. Using the criteria outlined in Chapter 2, features of PMN are compared and contrasted with modPMN-MI.

5.4.2.1 Methodological Steps for modPMN-MI

- The modPMN-MI method contains the same key processes as PMN (response propensity adjustment, predictive mean modeling, and hot-deck imputation). However, there are additional steps performed in modPMN-MI that are not performed in PMN including the data augmentation and cycling steps. Moreover, the response propensity modeling step used in modPMN-MI is modified as discussed in

Section 5.1.4.1. For categorical variables, the hot-deck step is replaced by stochastic imputation, and cycling only occurs in the multivariate case.

- PMN and modPMN-MI use SUDAAN models to account for the sample design. This means that the standard errors of parameter estimates are accurate, which leads to a more informed model-fitting experience when predictors need to be dropped. Although modPMN-MI is patterned after IVEware, it uses item nonresponse-adjusted survey weights to fit the predictive mean models at each step in the cycle, whereas IVEware uses unweighted regressions.

5.4.2.2 Complexity of Data Consistency and Order of Imputation for modPMN-MI

- Similar to PMN, modPMN-MI ensures the same level of consistency in post-imputation data distributions because all logical and likeness constraints for continuous variables are maintained in the hot-deck step, and for categorical variables, conditional probabilities are used if the imputed value must be restricted (see Section 5.1.4.5, last paragraph). Because continuous variables are imputed in sets with drug usage measures (e.g., all 12-month frequency variables made up a set), the number of logical constraints are reduced because there were not as many complex interactions that needed to be maintained when more than one drug measure was missing. The consistency for categorical variables is also maintained because conditional probabilities are used to ensure that the final imputed value is restricted to levels consistent with other variables.
- In modPMN-MI, it would be fairly easy to allow for a reordering of the variables to be imputed because the variables are imputed in sets (see Table 5.3, reproduced below for convenience) that relate to the drug usage measures as compared with PMN, which was imputed in sets based across drugs as shown in Table 5.12. Similar to PMN, the logical and likeness constraints used in modPMN-MI would need to be developed based on any new reordering.

(Reproduced) Table 5.3 Grouping of Drug Variables into Imputation Sets in modPMN-MI

Drug Variable	Lifetime Use	Recency	12-Month Frequency	30-Day Frequency	Age at First Use
Cigarettes Alcohol Inhalants Marijuana Pain Relievers Cocaine Heroin	Set 1	Set 2	Set 3 (cigarettes N/A)	Set 4 (pain relievers N/A)	Set 5

modPMN-MI = modified predictive mean neighborhood multiple imputation; N/A = not applicable.

Table 5.12 Grouping of Drug Variables into Imputation Sets in PMN

Drug Variable	Lifetime Use	Recency	12-Month Frequency	30-Day Frequency	Age at First Use
Cigarettes	Set 1	Set 2 (12-Month Frequency N/A)			Set 3
Alcohol		Set 4			Set 5
Inhalants		Set 6			Set 7
Marijuana		Set 8			Set 9
Pain Relievers		Set 10 (30-Day Frequency N/A)			Set 11
Cocaine		Set 12			Set 13
Heroin		Set 14			Set 15

N/A = not applicable; PMN = predictive mean neighborhood.

Note: This table is a simplified version of Table 5.1 in Frechtel et al. (2013). It represents what PMN might look like if PMN was restricted to the drug variables discussed in this methods study.

5.4.2.3 Issues for Implementation of modPMN-MI

- Similar to PMN, modPMN-MI requires two sets of model-fitting exercises for each variable that requires imputation: one for the response propensity model and one for the predictive mean model. However, modPMN-MI actually requires additional time to implement because there is a second cycle of predictive mean models involved for variables imputed multivariately. If MI is considered, then the number of models involved is further multiplied by the number of imputations. The response propensity models also require additional time because of the larger set of predictors used in these models.
- After the decision on how to group the variables requiring imputation into sets and how to order the variables within sets is completed, the program development for modPMN-MI can be based on PMN programs. However, the time required for developing programs for the data augmentation and cycling steps is substantial. Additionally, substantial time is required to develop programs to perform the MI aspect of this method.
- When compared with PMN, modPMN-MI requires additional programs to be developed and implemented, additional model-fitting adjustments, and additional quality control checks. However, some of the time spent on model fitting would be saved because the hot-deck step is not used for categorical variables.

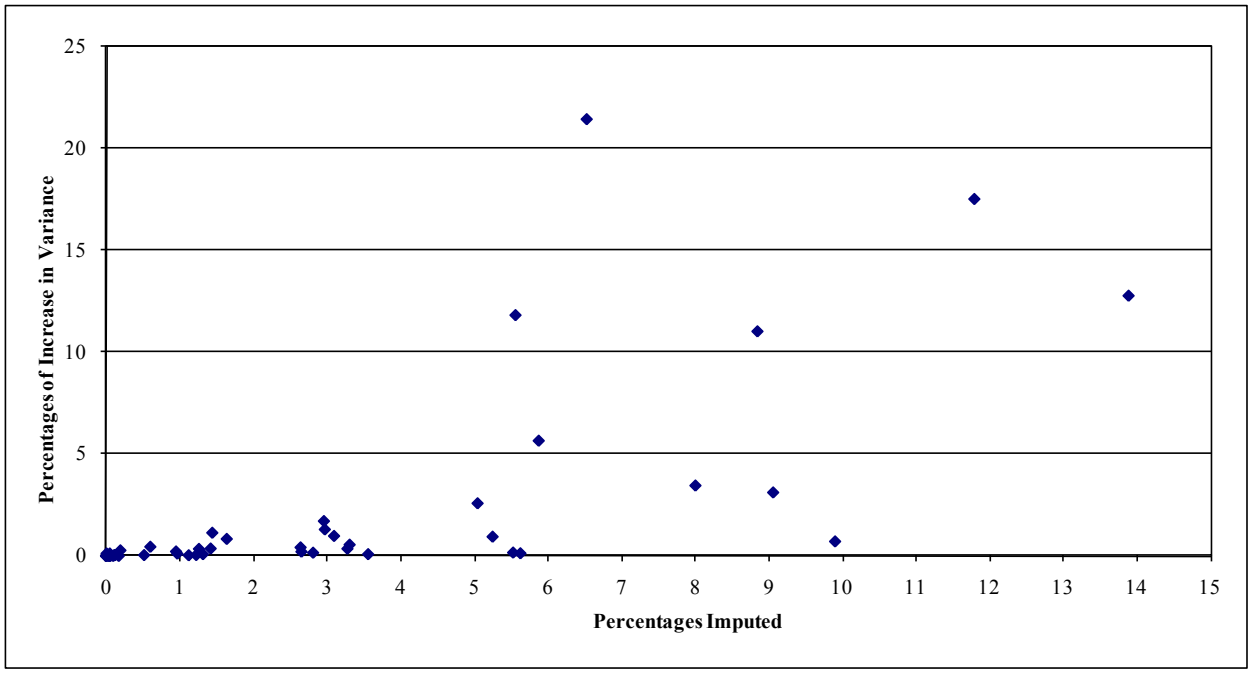
5.4.3 Variance Inflation Due to Imputation for modPMN-MI

Similar to IVEware, the amount of variance inflation due to imputation was calculated for modPMN-MI and is presented in Table B.2 in Appendix B. Figure 5.8 shows that the relative increase in variance is generally low (less than 5 percent) for modPMN-MI. The relative increase in variance due to imputation ranges from very low percentages (less than 1 percent) to relatively high percentages (more than 10 percent):

- Inhalants past month recency (37.19 percent)
- American Indian/Alaska Native (31.90 percent)
- Heroin 30-day frequency (21.86 percent)
- Heroin recency for more than 30 days ago but less than 12 months (21.41 percent)

- Cocaine 30-day frequency (17.5 percent)
- Heroin 12-month frequency (16.85 percent)
- Pain relievers 12-month frequency (12.76 percent)
- Heroin past month recency (11.81 percent)
- Inhalants recency for more than 30 days ago but less than 12 months (11.01 percent)

Figure 5.8 Relative Percentages of Increase in Variance as a Function of the Percentages of Imputed Data for modPMN-MI



modPMN-MI = modified predictive mean neighborhood multiple imputation.

Note: This figure excludes two extreme relative percentage increases for American Indian/Alaska Native (31.90 percent) and past month inhalants recency (37.19 percent) shown in Table B.2 in Appendix B.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Similar to IVEware, using this measure of relative increase in variance due to imputation, there appears to be some linear relation, which indicates that as the percentage imputed increases, the variance increases for some but not all of the estimates. By examining the amount of variance inflation using the data from modPMN-MI, it can be concluded that the relative increase in variance due to imputation is typically small (with a majority of the relative increases being lower than 5 percent), thus supporting the assumption that this variance is ignorable.

5.5 Summary and Options

During the implementation of modPMN-MI, a list of possible additional modifications for modPMN-MI was developed. Two highly recommended modifications are described below.

5.5.1 Cycling through Hispanic/Latino Origin and Race Variables

The Hispanic/Latino origin and race variables are known to be highly correlated, and despite this strong correlation, these two variables are not imputed together in either PMN or modPMN-MI. In both methods, race is imputed first and used as a predictor for Hispanic/Latino origin. PMN includes some likeness constraints for race that exploit the correlation. Hispanic/Latino recipients must have Hispanic/Latino donors, and recipients whose Hispanic/Latino group (e.g., Mexican) is known must have donors from the same Hispanic/Latino group. Table 5.2 provides a complete list of these constraints. These constraints are important because most of the item nonrespondents for race are of Hispanic/Latino origin. For example, in the 2007 NSDUH, about 98 percent of respondents who underwent imputation for race were of Hispanic/Latino origin. As implemented in this evaluation, modPMN-MI has no convenient way to use Hispanicity to assist with race imputation because (1) it was imputed after race and the missing values prevented it from being used as a predictor in the model, and (2) no constraints are used for categorical variables like race. For modPMN-MI, it is recommended to cycle through the Hispanic/Latino origin and race variables using the steps outlined in Section 5.1.

5.5.2 Bounding the Weights

The new method of adjusting weights for item nonresponse (Section 5.1.4.1) can result in extreme variability among the adjusted weights. Some observations can have an enormous influence on the response propensity model, whereas others have practically no influence on it. This is especially true when the item response rate is high, as it often is in the NSDUH. This results in predicted probabilities of item response very close to one for the majority of unit respondents. As p_k approaches one, $(1 - p_k)/p_k$ approaches zero.

An example of this occurred for the education level variable. In the 2007 NSDUH, there were only three item nonrespondents for education level in the age group of 18 to 25. In one of the five imputations for education level, the 75th percentile of the adjusted weight was 0.000008, meaning that 75 percent of the observations used in the response propensity model essentially had no impact. The largest adjusted weight was 57.26. A mitigating factor is that when there are so few item nonrespondents, the imputation procedure has very little impact on any analyses involving the outcome variable. For the next implementation of modPMN-MI, ways to impose bounds on these weights should be evaluated.

6. Statistical Evaluation Results

This chapter presents the statistical evaluation results of comparing the estimates from the five different imputation methods listed below.

- predictive mean neighborhood (PMN) (Chapter 2)
- simple weighted sequential hot deck (simple WSHD) (Chapter 3)
- complex weighted sequential hot deck (complex WSHD) (Chapter 3)
- IVEware (Chapter 4)
- modified predictive mean neighborhood multiple imputation (modPMN-MI) (Chapter 5)

Included is an examination of the differences between weighted estimates before and after imputation for each method and an evaluation of the significant differences among the weighted estimates from each method.

6.1 Before and After Imputation Distributions for Demographic Variables

The first comparisons examined were changes in the demographic variable distributions before and after imputation among the imputation methods. The before and after imputation unweighted frequency counts and weighted percentages for each method are displayed in Tables E.1 through E.4 in Appendix E. The first columns in Tables E.1 through E.4 display the number of item nonrespondents. For the race variable (Table E.1), there were 1,860 nonrespondents, which allowed for some variability across the imputation methods. One noteworthy difference was in the number of American Indians/Alaska Natives assigned using simple WSHD (84) as compared with the frequencies assigned using PMN (398), modPMN-MI (330), and IVEware (900).[31] For marital status (Table E.3) and education level (Table E.4), the number of item nonrespondents (18 and 10, respectively) was quite small. The number of item nonrespondents for Hispanic/Latino origin (Table E.2) was also relatively small (109). Given these relatively small numbers of item nonrespondents, the before and after weighted percentages for these three demographic variables were generally consistent across the different imputation methods.

The after imputation estimates for each imputation method (presented in the last columns of Tables E.1 through E.4) appeared similar for Hispanic/Latino origin, marital status, and education level, a result not surprising given the small number of imputed cases. However, this was not the case for the race variable. In particular, the weighted percentages for American Indian/Alaska Native ranged from 1.2 percent for PMN to 2.3 percent for IVEware. Similar variation existed for white and black/African American. Estimated percentages for white ranged from 80.5 percent for IVEware to 81.6 percent for PMN, and estimated percentages for black/African American ranged from 12.3 percent for IVEware to 12.5 percent for simple WSHD.

[31] This number is based on the first set of imputations. Each round of imputations for the IVEware programs produced similar estimates.

6.2 Before and After Imputation Distributions for Drug Variables

Similar to the demographic variables, the before and after imputation unweighted frequency counts and weighted percentages for the drug variables were calculated. For the drug frequency and age-at-first-use variables, the before and after imputation unweighted sample sizes and before and after imputation weighted mean estimates were calculated. In calculating the weighted percentages, only the subset of respondents for whom the variable is relevant or applicable contributed to the sum of the weights included in the corresponding denominators. Similarly, the unweighted frequencies represent the number of imputed or logically assigned cases for only those respondents for whom the variable is relevant. Note that, for IVEware, imputed data were available for only cigarettes, alcohol, and marijuana. The drug variable distributions are displayed in Tables E.5 through E.18.

The number of imputed values for the drug recency and frequency questions varied considerably across drug variables, ranging from less than 5 imputed cases for heroin 30-day frequency (Table E.18) to more than 2,000 imputed cases for alcohol 12-month frequency (Table E.13). Because each imputation method could have imputed different values for lifetime drug use and thus changed subsequent imputations of recency of use and frequency of use, the number of respondents flagged as imputed varied across the different imputation methods. Similarly, different imputation values for drug recency affected the number of imputations for the drug frequency variables.

Although some variation occurred in the distribution of imputed values across the levels of the recency variables, the numbers were generally too small to make a large difference in after imputation weighted percentages. Likewise, for the drug frequency variables, the before and after imputation weighted mean estimates were not strikingly different, with the exception of inhalants, cocaine, and heroin 12-month frequency. Compared with the before imputation estimate, the imputed weighted mean estimate of days using inhalants (Table E.14) and cocaine (Table E.17) increased by 5 or more days. The imputed weighted mean estimate of days using heroin in the past year (Table E.18) decreased by approximately 5 or more days. Even with these differences, the after imputation frequencies across imputation methods were rather similar.

6.3 Significant Differences among Imputation Methods for Demographic and Drug Variables

Using the repeated measures model described in Appendix D, differences among the imputation methods for the demographic and drug variables were tested. The analyses were performed for those aged 12 or older (i.e., the entire National Survey on Drug Use and Health [NSDUH] sample) and were split by age group (12 to 17, 18 to 25, and 26 or older) because PMN is typically performed within each of these age groups. First, a global test was conducted to determine whether significant differences existed between imputation methods. If the global test was significant, then all possible pairwise comparisons were tested to gain a better understanding of which methods differed significantly from each other.

The results of the global significance testing and pairwise testing for the demographic and drug variables can be found in Tables F.1 through F.16 in Appendix F. The estimates for each variable, across all ages and by age group, are listed first and then followed by the p-values

for the global and pairwise tests. The demographic variable results are shown in Tables F.1 through F.8, and the drug variable results are shown in Tables F.9 through F.16. For categorical variables such as demographic and drug recency variables, the pairwise comparisons were conducted within each level of the variable. For example, for white, differences in the imputation results were tested for PMN, simple and complex WSHD, modPMN-MI, and IVEware, but tests for differences between white and Asian/Other Pacific Islander were not tested.

For the demographic variables, race of the respondent was the only variable statistically significant at the global level ($\alpha = 0.05$). This result held for all three age groups as well as across all ages combined. Among the drug variables, many were found to have significant global differences ($\alpha = 0.05$), as detailed below.

- For those aged 12 or older, the following drug variables showed significance: cigarettes recency, cigarettes 30-day frequency, alcohol recency, alcohol 12-month frequency, alcohol 30-day frequency, marijuana recency, and marijuana 12-month frequency.
- For the age group of 12 to 17, the following drug variables showed significance: cigarettes recency, cigarettes 30-day frequency, cigarettes age at first use, alcohol recency, alcohol 12-month frequency, inhalants recency, marijuana recency, and pain relievers age at first use.
- For the age group of 18 to 25, the following drug variables showed significance: cigarettes recency, alcohol 12-month frequency, alcohol 30-day frequency, and marijuana recency.
- For the age group of 26 or older, the following drug variables showed significance: alcohol recency, alcohol 30-day frequency, alcohol 12-month frequency, and marijuana recency.

Not many significant differences were found among imputation methods for inhalants or cocaine. The comparison among methods for the heroin recency variable could not be performed due to the lack of differences between the estimates as a result of the small number of item nonrespondents that needed to be imputed.

6.3.1 Pairwise Differences among Imputation Methods for Demographic Variables

The demographic variables were imputed using only one of the WSHD approaches (simple). Hence, for the demographic models, four imputation methods were compared. Many statistically significant differences ($\alpha = 0.05$) existed across the imputation methods for each level of race, for all ages, and for each age group. These pairwise comparisons are presented in Tables F.2, F.4, F.6, and F.8. For the comparisons of those aged 12 or older where each level of race was statistically significant at the global level, the results of the pairwise comparisons (Table F.2) are summarized in Figure 6.1. In this figure and in subsequent figures found in Appendix G, the estimates for each imputation method are ordered from smallest to largest where estimates that are not statistically significant are linked together by an underline.

As shown in Figure 6.1, the differences were not consistent across the levels of race. For white, PMN imputation (81.6 percent) was significantly higher than the other three methods. WSHD (81.4 percent) and modPMN-MI (81.3 percent) were not found to be significantly

different from each other, but both imputation methods were significantly higher than IVEware (80.5 percent). IVEware (12.3 percent) and PMN (12.3 percent) for black/African American were not found to be significantly different from each other, though both methods were significantly lower than modPMN-MI (12.4 percent), and all three methods were lower than WSHD (12.5 percent). For Asian/Other Pacific Islander, the only differences found among the imputation methods were between PMN (4.9 percent) and WSHD (4.9 percent) and between PMN (4.9 percent) and IVEware (4.9 percent). For American Indian/Alaska Native, significant differences were not found between PMN (1.2 percent) and WSHD (1.2 percent), but both were significantly lower than modPMN-MI (1.5 percent) and IVEware (2.3 percent), and IVEware was found to be significantly higher than the other three methods. When split by age group (Tables F.4, F.6, and F.8), the results were fairly similar, with the exception of the age group of 26 or older where no significant differences were found for Asian/Other Pacific Islander. The complementary summary figure for the age groups can be found in Figure G.1 in Appendix G.

Figure 6.1 Pairwise Comparisons of Imputation Methods for Race: 12 Years or Older, Percentages

Results for White:

IVEware$_{80.5}$ < modPMN-MI$_{81.3}$ WSHD$_{81.4}$ < PMN$_{81.6}$

Results for Black/African American:

IVEware$_{12.3}$ PMN$_{12.3}$ < modPMN-MI$_{12.4}$ < WSHD$_{12.5}$

Results for Asian/Other Pacific Islander:

PMN$_{4.9}$ modPMN-MI$_{4.9}$ WSHD$_{4.9}$ IVEware$_{4.9}$

Results for American Indian/Alaska Native:

PMN$_{1.2}$ WSHD$_{1.2}$ < modPMN-MI$_{1.5}$ < IVEware$_{2.3}$

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Note: An underline indicates imputation methods that were not found to be statistically different from each other.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

6.3.2 Pairwise Differences among Imputation Methods for Drug Variables

Because of processing difficulties, IVEware was used to impute only cigarettes, alcohol, and marijuana (see Chapter 4 for details on the IVEware imputation process). Thus, the models for these drugs involved all five imputation methods under consideration, whereas the models for cocaine, inhalants, and heroin involved only the remaining four methods (i.e., all but IVEware).

Figure 6.2 presents a summary of the significance results of the pairwise comparisons (shown in Table F.10) by imputation method for cigarettes recency for all age groups. For cigarettes recency use within the past 30 days, PMN (24.2 percent), complex WSHD (24.2 percent), modPMN-MI (24.2 percent), and simple WSHD (24.3 percent) were not significantly different from each other. IVEware (24.3 percent) was different from all of the other four methods and produced a slightly higher 30-day frequency estimate. Similarly, recency for more than 12 months ago but within the past 3 years was not significantly different among complex WSHD (3.9 percent), simple WSHD (3.9 percent), PMN (4.0 percent), and modPMN-MI (4.0 percent), but significant differences were found between these methods and IVEware (3.9 percent). Similar patterns were found for marijuana recency as shown in Figure G.4. However, the patterns for alcohol recency (Figure G.3) were slightly different with significant differences for past month recency between PMN (51.1 percent) and complex WSHD (51.3 percent), and past month recency for all methods except PMN were significantly different from IVEware (51.1 percent). The summary figures for pairwise comparison by age group are presented in Appendix G: cigarettes recency in Figure G.2, alcohol recency in Figure G.3, marijuana recency in Figure G.4, and inhalants recency in Figure G.5. For the recencies for the other three drugs (pain relievers, cocaine, and heroin), the global tests did not find any differences among the four imputation methods (PMN, simple WSHD, complex WSHD, and modPMN-MI). This is likely due both to the dominance of the "never used" category for these rarer drugs and to the absence of results for IVEware, which tended to be the method that was different from the others most frequently.

Figure 6.2 Pairwise Comparisons of Imputation Methods for Cigarettes Recency: 12 Years or Older, Percentages

Results for Within Past 30 Days:

PMN$_{24.2}$ Complex WSHD$_{24.2}$ modPMN-MI$_{24.2}$ Simple WSHD$_{24.3}$ < IVEware$_{24.3}$

Results for More than 12 Months Ago but within Past 3 Years:

IVEware$_{3.9}$ < Complex WSHD$_{3.9}$ Simple WSHD$_{3.9}$ PMN$_{4.0}$ modPMN-MI$_{4.0}$

Results for More than 3 Years Ago:

IVEware$_{32.8}$ Simple WSHD$_{32.8}$ Complex WSHD$_{32.9}$ PMN$_{32.9}$ modPMN-MI$_{32.9}$

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Note: An underline indicates imputation methods that were not found to be statistically different from each other.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Figures G.6 through G.8 summarize the pairwise comparison results for 12-month frequency of use, 30-day frequency of use, and age at first use. As shown in Figure G.6, only alcohol and marijuana showed significant differences for the global test for 12-month frequency of use across all ages. For alcohol 12-month frequency, simple WSHD (86.3 days) was significantly lower than complex WSHD (86.7 days), modPMN-MI (86.7 days), IVEware (86.8 days), and PMN (86.9 days). Additionally, PMN and modPMN-MI were significantly different from each other. Similar patterns were seen among the three different age groups for alcohol 12-month frequency where simple WSHD (86.3 days) produced significantly lower estimates than all other methods. Note that, for the age group of 12 to 17, simple WSHD (36.0 days) and complex WSHD (36.6 days) were not significantly different from each other.

Similar to 12-month frequency of use, only cigarettes and alcohol showed significant differences for 30-day frequency of use in the global comparison test. As seen in Figure G.7, cigarettes use for those aged 12 or older for IVEware (22.6 days) was significantly lower than for all other methods. For the age group of 12 to 17, the same pattern was seen; that is, the results generated by IVEware were significantly lower than all the other methods, and none of the other methods were found to be significantly different from each other. By contrast, different results were seen for alcohol 30-day frequency for those aged 12 or older where PMN (8.4 days) versus IVEware (8.4 days) and PMN (8.4 days) versus simple WSHD (8.4 days) were the only pairs statistically different from each other. For the age group of 18 to 25, simple WSHD (7.3 days) was significantly lower than the other four methods.

Significant differences for age at first use were not very common. Among all the statistical tests, age-at-first-use variables were only statistically significant for cigarettes and pain relievers in the age group of 12 to 17 (Figure G.8). For pain relievers, PMN (13.3 years), complex WSHD (13.4 years), and modPMN-MI (13.4 years) were not found to be statistically different. However, all three methods were significantly lower than simple WSHD (13.4 years). For cigarettes, PMN (12.6 years) was found to be significantly lower than complex WSHD (12.6 years), IVEware (12.6 years), and simple WSHD (12.7 years), but differences were not found with modPMN-MI (12.6 years).

It is not surprising to see few significant differences for frequency and age at first use because of the small domains. For small domains, there is limited power to detect differences between the methods because of the small sample sizes. For 12-month frequency, only past year users are in the domain; for 30-day frequency, only past month users are in the domain; and for age at first use, only lifetime users are in the domain. The repeated measures analyses only include observations that are in the domain. For the rare drugs, the domains are smaller than for the more common drugs, so it is also unsurprising that most of the few differences seen were for common drugs like cigarettes and alcohol.

6.4 Bias Ratios and Confidence Intervals for Coverage Probabilities for the PMN Imputation Method versus Other Imputation Methods

Because of the large sample size and high correlation within the repeated measures, there were many significant differences among the imputation methods. The number of comparisons between estimates by pairs of imputation methods ranged from 140 to more than 230 comparisons (as discussed later in Table 6.3). Of these comparisons, the number of significant

differences among estimates ranged from 6 to 36, depending on which two imputation methods were compared. To help identify the more meaningful differences, bias ratios and actual coverage probabilities for the biased 95 percent confidence intervals were calculated, assuming PMN represented the least biased or approximately unbiased method.[32] The bias ratio is defined as the ratio of the estimated bias to the standard error of the estimate. Bias ratios give a measure of the magnitude of the bias and how confidence interval coverage might be affected by the bias. For example, a bias ratio of 0.10 or less would result in the probability of an error (of more than 1.96 standard deviations from the mean) of 0.0511 as compared with the usual 0.05 significance level. However, as the bias ratio increases, the effect becomes more serious. When the bias ratio is 1, the total probability of error increases to 0.17, which is more than 3 times the usual 0.05. Based on Cochran's "working rule," a bias ratio of less than 0.2 can be interpreted as being small enough to have only a modest effect on the accuracy of the estimate (Cochran, 1977, pp. 12-15). The methodology for computing the bias ratios and coverage probabilities is described in Appendix D.

The bias ratios and 95 percent confidence interval coverage probabilities for the demographic variables with statistically significant PMN versus other imputation method comparisons can be found in Table 6.1. For demographic variables, race was the only variable with significant differences. For American Indian/Alaska Native, the first comparison (PMN versus simple WSHD) was not significant. However, the second comparison (PMN versus IVEware) was significant, and the bias ratio for IVEware compared with PMN was rather large (10.03). Because of this large ratio, the coverage probability was 0.00 percent, meaning that the interval was almost guaranteed not to capture the population mean (assuming it was represented by PMN). The third comparison (PMN versus modPMN-MI) was also significant and the bias ratio was large (3.63), resulting in a confidence interval coverage of only 0.73 percent. However, for Asian/Other Pacific Islander, the bias ratios for PMN versus simple WSHD and PMN versus IVEware were 0.25 and 0.27, respectively, with the coverage probability equal to approximately 94 percent for both comparisons, a minimal loss of coverage. Among the remaining significant comparisons, the bias ratios for white were largest between PMN and IVEware (-2.71), PMN and modPMN-MI (-0.87), and PMN and simple WSHD (-0.67). For black/African American, the bias ratio for PMN versus simple WSHD was the largest (0.53), followed by a small bias ratio for PMN versus modPMN-MI (0.10).

[32] As discussed in Appendix D, this assumption is probably not true, but it is made because PMN is the current NSDUH imputation method.

Table 6.1 Race Variable Bias Ratios and 95 Percent Confidence Intervals for Coverage Probabilities for PMN versus Other Methods: 12 Years or Older

Variable	PMN vs. Simple WSHD		PMN vs. IVEware		PMN vs. modPMN-MI	
	Bias Ratio	95% CI Coverage Probability	Bias Ratio	95% CI Coverage Probability	Bias Ratio	95% CI Coverage Probability
Race						
American Indian/Alaska Native	N/A	N/A	10.03	0.00	3.63	0.73
Asian/Other Pacific Islander	0.25	94.27	0.27	94.18	N/A	N/A
Black/African American	0.53	91.70	N/A	N/A	0.10	94.89
White	-0.67	89.65	-2.71	22.75	-0.87	86.08

CI = confidence interval; modPMN-MI = modified predictive mean neighborhood multiple imputation; N/A = not applicable, which indicates that the comparison was not significant; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

For the drug variables with significant differences when comparing PMN with the other methods, the bias ratios and 95 percent confidence interval coverage probabilities are shown in Table 6.2. Only the levels with statistically significant results within the drug recency variables are displayed. For example, only three levels of the five-level cigarettes recency variable are displayed. None of the bias ratios for the drug variables were as extreme as the American Indian/Alaska Native ratios noted above. The bias ratios and loss of coverage were the highest for PMN versus IVEware for marijuana recency of more than 30 days ago but within the past 12 months (-0.67 bias ratio; 89.73 percent confidence interval) and for PMN versus simple WSHD for alcohol 12-month frequency (-0.61 bias ratio; 90.65 percent confidence interval). Other notable coverage probability reductions when PMN was compared with IVEware were the following:

- marijuana 12-month frequency (92.51 percent)
- cigarettes recency more than 30 days ago but within past 12 months (93.11 percent)
- alcohol recency more than 12 months ago (93.55 percent)
- marijuana recency within past 30 days (93.52 percent)
- marijuana recency more than 12 months ago (93.53 percent)
- cigarettes recency within past 30 days (94.28 percent)
- alcohol 30-day frequency (94.44 percent)

Alcohol recency within the past 30 days for PMN versus complex WSHD and alcohol 12-month frequency for PMN versus modPMN-MI were the remaining two comparisons with notable coverage probability reductions (94.16 percent and 94.46 percent, respectively). The rest of the statistically significant comparisons had bias ratios too small to be of much concern.

Table 6.2 Drug Variable Bias Ratios and 95 Percent Confidence Intervals for Coverage Probabilities for PMN versus Other Methods: 12 Years or Older

Variable	PMN vs. Simple WSHD		PMN vs. Complex WSHD		PMN vs. IVEware		PMN vs. modPMN-MI	
	Bias Ratio	95% CI Coverage Probability	Bias Ratio	95% CI Coverage Probability	Bias Ratio	95% CI Coverage Probability	Bias Ratio	95% CI Coverage Probability
Cigarettes Recency								
Within Past 30 Days	*	*	*	*	0.25	94.28	*	*
More than 12 Months Ago but within Past 3 Years	*	*	*	*	-0.40	93.11	*	*
More than 3 Years Ago	*	*	*	*	-0.09	94.91	*	*
Cigarettes	*	*	*	*			*	*
30-Day Frequency	*	*	*	*	-0.19	94.58	*	*
Alcohol Recency	*						*	*
Within Past 30 Days	*	*	0.27	94.16	*	*	*	*
More than 12 Months Ago	*	*	-0.12	94.83	0.35	93.55	*	*
Alcohol								
12-Month Frequency	-0.61	90.60	*	*	*	*	-0.22	94.46
30-Day Frequency	*	*	*	*	-0.22	94.44	-0.16	94.69
Marijuana Recency	*	*	*	*				
Within Past 30 Days	*	*	*	*	-0.36	93.52	*	*
More than 30 Days Ago but within Past 12 Months	*	*	*	*	-0.67	89.73	*	*
More than 12 Months Ago	*	*	*	*	0.36	93.53	*	*
Marijuana	*	*	*	*			*	*
12-Month Frequency	*	*	*	*	0.46	92.51	*	*

CI = confidence interval; modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

Note: An asterisk indicates that the comparison was not significant.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

6.5 Two-Way Drug Comparisons

To assess the question of whether there were differences in estimates between two different drug variables, the following two-way comparisons were examined:

- percentage of lifetime drug users of a specified drug given no other lifetime drug use
- percentage of lifetime cocaine or heroin use given cigarettes, alcohol, or marijuana lifetime use

These two examples were used to evaluate whether the imputation was able to maintain the correlations that exist for polydrug use. Appendix H presents the estimates (Tables H.1 and H.2) based on the imputed data from each of the different imputation methods.

Similar to the analyses described in Section 6.3, global tests of significance were performed before conducting pairwise tests. For both two-way comparisons, none of the global tests were statistically significant, and thus no pairwise tests were conducted. Also, as before, the statistical tests for heroin recency (used to create heroin lifetime use) could not be conducted due to the similarity of the estimates stemming from the extremely small number of item nonrespondents, a problem also seen for lifetime cocaine users given no other lifetime drug use in the age group of 18 to 25 (Table H.1).

The estimates for the first comparison (Table H.1) were extremely similar across imputation methods. Hence, the p-values from global testing for significant differences among imputation methods were insignificant. Similar results were found for the second comparison (Table H.2).

6.6 Summary

The different imputation methods did produce some significant differences, but for the majority of the estimates, statistically significant differences were not found and many of the differences were small and potentially not substantively meaningful. Furthermore, when only considering PMN compared with the other four imputation methods,[33] the percentage of significant differences (out of all the statistical tests conducted) was low. The numbers and percentages of significant differences by comparison are shown in Table 6.3.

[33] In the previous sections of this chapter, the discussion is based on all possible comparisons among all imputation methods. In this summary section, the discussion is based only on the possible comparisons between PMN and the other imputation methods.

Table 6.3 Numbers and Percentages of Significant Differences for PMN versus Other Methods

Age Group by Variable	Total Differences among all Methods	PMN vs. Simple WSHD	PMN vs. Complex WSHD	PMN vs. IVEware	PMN vs. modPMN-MI
12+					
Demographic	9	3	N/A	3	3
Drug	15	1	2	10	2
12-17					
Demographic	7	2	N/A	3	2
Drug	13	3	3	6	1
18-25					
Demographic	9	3	N/A	3	3
Drug	6	2	0	4	0
26+					
Demographic	7	2	N/A	3	2
Drug	6	1	1	4	0
Total Differences by Method	72	17	6	36	13
Total Number of Comparisons	N/A	232	184	140	232
Percentage of Total Differences by Method	N/A	7.3	3.3	25.7	5.6

modPMN-MI = modified predictive mean neighborhood multiple imputation; N/A = not applicable; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

Note: The percentage of total differences by comparison was calculated by dividing the number of significant test results by the total number of possible comparisons where the total includes tests for each level of categorical variables. Appendix F denotes these significant differences.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

IVEware comparisons against PMN had the most significant differences. In total, there were 140 PMN versus IVEware comparisons, which included tests conducted

- across all ages,
- within each of the three age groups,
- for each level of the demographic and recency variables, and
- for the drug frequency and age-at-first-use variables.

Of these 140 comparisons, 36 (25.7 percent) were statistically significant. Similarly, the percentages of significant differences were calculated for the comparisons between PMN and the other methods. The percentages were much lower for the other comparisons: 3.3 percent of the statistical tests between PMN and complex WSHD were significant; 5.6 percent were significant between PMN and modPMN-MI; and 7.3 percent were significant between PMN and simple WSHD.

For the demographic variables, across all comparisons and regardless of age, race was the only variable with significant differences. Across all ages and when comparing PMN with the other imputation methods, there were a total of nine significant differences (three for PMN

versus simple WSHD, three for PMN versus IVEware, and three for PMN versus modPMN-MI) within the different levels of race as shown in Table 6.3. The results were very similar when split by age group: a total of seven significant differences for 12 to 17, nine total significant differences for 18 to 25, and seven total significant differences for 26 or older.

There were more significant differences for the drug variables, though there were also more comparisons. For those aged 12 or older, comparing only PMN with the other imputation methods, there were 15 significant differences spread across the different levels of the drug recency variables, the age-at-first-use variables, and the frequency variables. These differences were found for alcohol, cigarettes, and marijuana recency; alcohol and marijuana 12-month frequency; and alcohol and cigarettes 30-day frequency. Very few differences were found for the rarer drugs (inhalants, pain relievers, cocaine, and heroin). This was due in part to the absence of results for IVEware, which tended to be the method that was different from the others most frequently, and in part to the small domains and imbalanced distributions that tend to be associated with rarer drugs.

The number of significant differences within each age group dropped in comparison with those aged 12 or older. Again, only including the comparisons with PMN, there were 13 significant differences for the age group of 12 to 17 and 6 significant differences for the age groups of 18 to 25 and 26 or older. For 12 to 17, alcohol, cigarettes, and inhalants recency had significant differences within at least one of the drug recency levels. Alcohol 12-month frequency and cigarettes 30-day frequency as well as cigarettes and pain relievers age at first use also had significant differences for 12 to 17. Cigarettes and marijuana recency and alcohol 12-month frequency and alcohol 30-day frequency were the only variables with significant differences for 18 to 25. For 26 or older, alcohol and marijuana recency and alcohol 12-month frequency were the only variables with significant differences.

Although these differences were statistically significant, substantively one might question if they were meaningful. Among the demographic variables, the largest difference between the estimates was found between PMN and IVEware within the white and American Indian/Alaska Native levels of the race variable. These larger differences were found for the age groups of 12 or older, 12 to 17, and 18 to 25. The most notable percentage point difference for these comparisons was for 18 to 25: 1.9 for white (PMN estimate 78.6 minus IVEware estimate 76.7) and 1.7 for American Indian/Alaska Native (PMN estimate 1.4 minus IVEware estimate 3.1) (Table F.5). Within the drug variables, none of the differences were as large as the noted race differences. The largest difference was found for those aged 12 or older for marijuana 12-month frequency (Table F.9): IVEware's 12-month frequency (102.9 days) was 1.0 days less than PMN's 12-month frequency (101.9 days), a difference that is considerably small relative to the actual estimates. Given the percentage imputed for race (2.5 percent) as compared with the percentage imputed for the other variables (see Tables 1.1 and 1.2), it is not surprising that the race variable is the only variable with notable differences across the imputation methods.

Some of the differences between IVEware and PMN and modPMN-MI relate to the model-fitting procedures. One of the strengths of IVEware and modPMN-MI as compared with PMN relates to the sequence in which the regression models are developed. Both IVEware and modPMN-MI develop models by cycling through predictor variables after provisionary imputations have been completed, thus allowing for a more robust model-fitting process and

producing similar estimates across the multiple imputations. In PMN, the predetermined hierarchy for imputing the drug variables does not allow for additional predictors to be used.

The sampling weights of unit respondents for the NSDUH vary substantially. PMN and modPMN-MI utilize SUDAAN® models (RTI International, 2013), which account for the sample design. This means that the standard errors of parameter estimates are accurate, which leads to a more informed model-fitting experience when predictors need to be dropped. IVEware does not use the sampling weights to determine the parameter estimates for the regression models. As a result, IVEware may yield parameter estimates that vary substantially from those that are based upon the weighted regression models used in PMN because of the differential weighting problem. Although modPMN-MI is patterned after IVEware, it uses item nonresponse-adjusted survey weights to fit the predicted mean models at each step in the cycle, whereas IVEware uses unweighted regressions. This variation could also explain some of the differences in the estimates between the post-imputation distributions of these two methods.

This page intentionally left blank

7. Imputation of Race, Hispanicity, and Age Variables

7.1 Introduction

This chapter and the next explore the use of additional data in imputations beyond the responses to the National Survey on Drug Use and Health (NSDUH) interview questions. Chapter 7 focuses on the age, race, and Hispanicity data available from the screener questionnaire, which is a short interview conducted with an adult in the dwelling unit (DU) that asks a few basic questions about each member of the DU. Chapter 8 evaluates whether interview data collected from the other member of the DU (when two members of a dwelling unit are selected to be interviewed and have completed interviews) can be used to assist with imputation. If there is a strong correlation overall or across relevant subgroups between the interview and screener data and/or interview and pair member data, there is a possibility that the noninterview data sources can be used in a deterministic imputation method that would be an improvement on the predictive mean neighborhood (PMN) method as well as a cost-saving simplification.

Most of Chapter 7 focuses on race and ethnicity for two reasons: (1) the screener asks each member of the DU a simplified version of the race question and the Hispanic/Latino origin question, so screener data are usually available for these variables; and (2) in Section 6.1, differences were noted between the imputation methods for race, suggesting that the imputation method can be improved and/or warrants further investigation. Current PMN procedures use the race of the householder (i.e., the screener respondent) as a covariate in the race models instead of in a deterministic manner. Current PMN procedures also use the screener data in the editing of the age variable. The consistency between screener age and interview age is also examined in this chapter.

Section 7.2 describes the NSDUH screener and interview data sources in more detail, such as what data are collected, how the screener compares with the interview, and how frequently an interview is completed by the screener respondent (called a "dual respondent" in the chapter). Section 7.3 reports the findings of a short literature review and "meta-survey," in which staff from other major national surveys were asked about their imputation methods for race and ethnicity, whether a screener was done, and whether the screener data were used in editing and/or imputation. Section 7.4 describes the feasibility of implementing a deterministic imputation method for race and Hispanicity. Section 7.5 is a detailed investigation of alternative race/ethnicity imputation methods, many of which involve the screener data. Section 7.6 is a brief assessment of the utility of the screener data in the editing of the age variable. Section 7.7 summarizes the findings of the analyses reported in the chapter.

7.2 NSDUH Screener and Interview Data

This section briefly describes the screener and interview process and the variables that are collected from the screening and interview questionnaires.

7.2.1 Screener Variables

The DUs, such as a house, an apartment, assisted living quarters, or student housing, are randomly selected from a sampled area segment of a census tract. Professionally trained interviewers confirm the DU address and conduct a screener interview with an available resident of the DU who is aged 18 or older. Interviewers collect information from that person about all people aged 12 years or older who reside in the DU most of the time for the current year's quarter (January-March, April-June, July-September, or October-December). Information is collected starting with the head of house (also known as the householder) and then continuing with the oldest resident to the youngest resident (12 years or older). Information collected for each eligible resident includes age, relation to the householder, gender, Hispanicity, race, and active military duty status. After information is collected for each resident, interviewers ask the screener respondent to confirm the information given. The screener data related to age, gender, and race/ethnicity are described below.

- **Screener Age:** Age for each eligible DU resident is collected by stating "Please tell me the age of this person on his or her last birthday." Respondents do not have the option of refusing to provide a response. If the respondent does not know the resident's age, then the respondent is asked to put the resident in one of the following age categories: 12 to 17, 18 to 25, 26 to 34, 35 to 49, and 50 or older.
- **Screener Gender:** Gender is collected by asking "Is this person male or female?" The screener respondent has the option of selecting "Male," "Female," or "Refuse." If the screener respondent chooses "Refuse," then gender is assigned a missing value.
- **Screener Hispanicity:** During the screener process, Hispanic/Latino origin is collected by asking the screener respondent the following question: "Is he/she of Hispanic, Latino, or Spanish origin? (That is, do any of these groups describe his/her national origin or ancestry – Puerto Rican, Cuban, Cuban-American, Mexican, Mexican-American, Chicano, Central or South American, or origin in some other Spanish-Speaking country?" The screener respondent may choose "Yes," "No," "Unknown," or "Refuse." If the screener respondent chooses "Unknown" or "Refuse," then the Hispanic/Latino origin is assigned a missing value.
- **Screener Race:** Race is collected by asking "Is he/she White, Black or African American, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, or Asian?" The respondent may choose "Unknown," "Refused," or one or more of the following options: white, black/African American, American Indian/Alaska Native, Native Hawaiian/Other Pacific Islander, Asian, or Other. If the screener respondent chooses "Unknown" or "Refuse," then a missing value is assigned to the screener race variable.

7.2.2 Interview Variables

Once the screening process is complete, a predetermined algorithm based on age and household age composition is used for each sample dwelling unit (SDU) to select zero, one, or two people to participate in the interview process. A variety of information is collected during the interview through audio computer-assisted self-interviewing (ACASI) including demographics, drug use, household income, and insurance. The collection of the age, gender, race, and Hispanicity data during the interview is described below.

- **Interview Age:** After a respondent enters his or her birth date in the first part of the interview, he or she has multiple opportunities to change his or her age in response to consistency checks throughout the interview. Therefore, it is possible for the age reported by the respondent at the beginning of the interview (CALCAGE) to be different from the age reported at the end of the interview (NEWAGE). The final age variable, AGE, is determined using these two variables and three other sources: the age calculated from the final edited interview date (INTDATE) and the raw birth date (AGE1), the age corresponding to the "self" in the interview household roster (if it existed), and the pre-interview screener age. There were no missing values for interview age in the 2009 NSDUH. Of the 68,700 interview respondents, 68,694 (99.99 percent) had identical and nonmissing CALCAGE and NEWAGE values.
- **Interview Gender:** As with surveys since 2002, it was mandatory in the 2009 survey for an interviewer to enter the respondent's gender in question QD01. As a result, it was not possible to have missing values for this question. To maintain continuity with the 1999-2001 surveys, the variable name IRSEX was used to describe gender in the 2009 survey. However, it was not necessary to create an imputation indicator, because IRSEX and QD01 were equivalent.
- **Interview Hispanicity:** In the 2009 survey, two core questions (QD03 and QD04) focused on the respondent's ethnicity. Question QD03 asks about Hispanic/Latino origin and question QD04 asks about the Hispanic/Latino group. A respondent is administered QD04 only if he or she answered "Yes" to QD03. The QD03 question asks "Are you of Hispanic, Latino, or Spanish origin or descent?" The respondent may choose "Yes," "No," "Refuse," or "Unknown." The QD04 question is asked only if the respondent indicates that he or she is of Hispanic/Latino origin (QD03 = 1) and is phrased as "Which of these Hispanic, Latino, or Spanish groups best describes you?" The respondent may choose "Unknown," "Refuse," or one or more of the following options: (1) Mexican/Mexican American/Mexicano/Chicano, (2) Puerto Rican, (3) Central/South American, (4) Cuban/Cuban American, (5) Dominican (from the Dominican Republic), (6) Spanish (from Spain), or (7) Other (Specify).
- **Interview Race:** In the 2009 survey, two core questions focused on the respondent's race (QD05 and QD05ASIA). The QD05 question asks "Which of these groups describes you?" The respondent may choose "Unknown," "Refuse," or one or more of the following options: white, black/African American, American Indian/Alaska Native (American Indian includes North American, Central American, and South American Indians), Native Hawaiian/Other Pacific Islander, Asian (e.g., Asian Indian, Chinese, Filipino, Japanese, Korean, and Vietnamese), or Other (Specify). If the respondent chooses "Asian," then the QD05ASIA follow-up question is asked and is phrased as "Which of these groups describes you?" The respondent may choose "Unknown," "Refuse," or one or more of the following options: Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, or Other (Specify).

7.2.3 Screener and Interview Respondents and Data

As discussed before, the screener respondent can be any adult household resident aged 18 or older, whereas the interview respondent has to be a randomly sampled household member aged 12 or older. The sampling selection probability varies across the different age groups, with

12 to 17 and 18 to 25 being the age group categories with the highest sampling rates.[34] Because the screener questions are answered by an adult aged 18 or older from the SDU and a maximum of two members are selected, it is likely that the screener respondent and interview respondent are two different people. In the 2009 NSDUH, 31.8 percent of the interview respondents were also the screener respondents (Table 7.1). To find out whether screener race data could be used to edit the missing interview race/ethnicity, it is of interest to see how well the screener data agree with the interview data based on the condition that the interview respondent was also the screener respondent. For discussion purposes, a respondent who answered both screener and interview questions is called a dual respondent. Logic suggests that screener data would agree more with interview data when the screener respondent and the interview respondent are the same person. One aspect of this evaluation was to examine whether screener respondents who were also the interview respondents could serve as deterministic imputation donors when interview respondents' race and Hispanicity were missing, as well as whether screener respondents who were not the interview respondents could potentially serve as proxy representatives.

Table 7.1 Screener Respondents among Interview Respondents, 2009 NSDUH

Dual Respondents (same person answering both screener and interview questions)	Frequency[1]	Unweighted Percentage[2]
Yes	21,800	31.8
No or Unknown	46,900	68.2
Total	68,700	100.0

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Percentages have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

7.3 Literature Review of Imputation Methods for Other National Surveys

Federal agencies that sponsor various surveys were contacted in order to gain insight on imputation methods for race and Hispanicity used by other national surveys similar to the NSDUH. Four surveys were identified for investigation: National Health Interview Survey (NHIS), National Health and Nutrition Examination Survey (NHANES), the American Community Survey (ACS), and the National Crime Victimization Survey (NCVS). Each survey representative (see notes for Tables 7.5 and 7.6) shared information about methodologies used for imputing race and Hispanicity and helped assemble population estimates based on the respective survey. A brief overview of each of the surveys is presented below.

7.3.1 Summary of the Selected Surveys

The **National Survey on Drug Use and Health (NSDUH)** is an annual nationally representative household survey of the civilian, noninstitutionalized population of the United States involving face-to-face interviews of roughly 70,000 randomly selected individuals aged 12 or older.[35] Data from the NSDUH provide national and state-level estimates and track trends on

[34] Refer to the 2009 sample design report of the NSDUH methodological resource book (MRB) for details on the sample allocations by different age groups (Morton, Martin, Chromy, Foster, & Hirsch, 2010).

[35] NSDUH website: https://www.samhsa.gov/data/

the use of tobacco products, alcohol, illicit drugs, and mental health in the United States. The NSDUH is sponsored by the Substance Abuse and Mental Health Services Administration (SAMHSA).

The **National Health Interview Survey (NHIS)** is an annual face-to-face study of the noninstitutionalized population of the United States.[36] It is sponsored by the National Center for Health Statistics, whereas the U.S. Census Bureau collects the data. The main objective of the NHIS is to monitor the health of the United States population through the collection and analysis of data on a broad range of health topics. In 2010, 90,000 people were interviewed from 35,000 households. The study produces trend data based on cross-sectional data collection.

The **National Health and Nutrition Examination Survey (NHANES)** combines a face-to-face interview and physical examinations to assess the health and nutritional status of adults and children in the United States.[37] This survey has been sponsored by the National Center for Health Statistics (NCHS) since the early 1960s. The NHANES III, conducted between 1988 and 1994, collected health information from about 40,000 participants. Since 1999, NHANES has been an annual survey selecting a nationally representative sample of about 5,000 people where the survey health-related topics change from year to year.

The **American Community Survey (ACS)** is a nationwide continuous survey where a monthly sample of housing units is chosen to produce estimates of population and housing characteristics. The ACS can be used to provide small-area estimates (e.g., census tracts and block groups) for 1-, 3-, and 5-year periods.[38] This survey is sponsored by the U.S. Census Bureau. Each year since 2006, the ACS has collected data from nearly 2 million housing units and 145,000 group quarters.

7.3.2 Information Requested from Representatives for Selected Surveys

The representatives contacted for each selected survey were asked to report the weighted proportions for race broken down in the following three ways: race (white, black/African American, other) for all individuals, race including Hispanicity (non-Hispanic/Latino white, non-Hispanic/Latino black/African American, non-Hispanic/Latino other, and Hispanic/Latino[39]), and race (white, black/African American, other) only for individuals who identified themselves as Hispanic/Latino. Individuals were also asked to provide the weighted rate of imputed race, imputed Hispanicity, and imputed race only for individuals who identified themselves as Hispanic/Latino. The staff working on these four surveys were also asked to provide information about their race and Hispanicity imputation methodology and their confidence in their imputation process as well as how they report race and Hispanicity. In summary, responses to the following questions were requested:

- Do you impute race and Hispanicity?
- What are your national weighted estimates for race categories and Hispanicity?

---

[36] NHIS website: https://www.cdc.gov/nchs/nhis/about_nhis.htm

[37] NHANES website: https://www.cdc.gov/nchs/nhanes/index.htm

[38] ACS website: https://www.census.gov/programs-surveys/acs/

[39] The Office of Management and Budget defines Hispanic/Latino as "a person of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin regardless of race."

- If you do impute race and Hispanicity, do you impute them separately (globally) or within the Hispanic/Latino category?
- If you do impute race and Hispanicity, what information do you use to impute each one?
- Do you have a screening process? If so, do you use the screener variable to logically edit, impute, or include in a prediction model?
- How confident are you of your imputed race and Hispanicity data?
- In your regular tables, is Hispanic/Latino presented as a separate race/ethnicity category? If not, then are you ever asked for tables with Hispanic/Latino distributed among race categories?
- Do you/can you break down the missing Hispanicity by race?
- Do you/can you break down the missing race by Hispanicity?

Finally, the survey representatives were asked if any literature is available that explains in more detail the imputation methodology and processing. The following sections discuss the responses to these questions as well as the race and Hispanicity imputation rates.

7.3.3 Imputation Methodologies for Race and Hispanicity Used in National Surveys

All four surveys impute race and Hispanicity separately, using variations of hot-deck procedures. The methods of imputation, however, are different. The NSDUH uses demographic variables and census block estimates, such as percentage of Hispanics in segment, to impute race and Hispanicity. The NSDUH has screener data, but screener data are not used to impute race or Hispanicity. The NHANES III is the only other survey that has screener data and uses the screener race and Hispanicity values to logically assign values to missing race or Hispanicity interview data. If screener and race are missing, the NHANES III uses census block estimates to impute race and Hispanicity.

Both NHIS and the ACS do not have a screening process. If there are household members with nonmissing Hispanic/Latino and race values, then the ACS uses the relationship of a household member to impute the missing race or Hispanicity. If the individual cannot be matched by relationship to another household member or if all members of the household have missing race and/or Hispanic/Latino values, the ACS matches the participant's last name to a Hispanic/Latino surname file to impute Hispanicity. Race is then imputed via a hot-deck procedure using age, gender, and Hispanicity as likeness constrains. Similar to the ACS, NHIS first tries to match individuals who are missing race and Hispanicity to a relative in the household who has nonmissing values for race and Hispanicity. If this is not possible, NHIS uses census block estimates to conduct a hot-deck procedure.

7.3.4 Race and Hispanicity Distribution among Surveys

The estimates for race and Hispanicity presented in Tables 7.2 through 7.6 were obtained from the survey representatives and the respective study's public use file (PUF) data. The tables present the weighted race and Hispanicity distribution. Table 7.2 shows the weighted percentages for race and Hispanicity for each survey. All four surveys report race and Hispanicity, but race regardless of ethnicity is most frequently reported in these surveys. The weighted estimates are very similar across all four surveys.

Table 7.2 Weighted Distribution of Race/Ethnicity

Race/Ethnicity	NSDUH[1]	NHIS[2]	NHANES[3]	ACS[4]
Non-Hispanic/Latino White	67.4	65.39	64.62	65.35
Non-Hispanic/Latino Black/African American	11.9	12.80	12.06	12.10
Non-Hispanic/Latino Other or Two or More Races	6.4	5.62	7.39	7.12
Hispanic/Latino	14.3	16.20	15.94	15.43

[1] Data are from the 2009 NSDUH Restricted-Use Analytic File. Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

[2] Data are from the 2010 NHIS Public Use Microdata file from DataFerrett: https://dataferrett.census.gov/LaunchDFA.html.

[3] Data are from the NHANES 2009-2010 Public Use Data File: https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2009.

[4] Data are from the 2007-2009 ACS 3-year estimates Public Use Microdata Sample from DataFerrett.

The race distribution shown in Table 7.3 is similar in three of the four surveys, but the ACS has a larger percentage of two or more races (13.03 percent versus about 6 percent) and a smaller white percentage (74.59 percent versus about 80 percent) than the other surveys. The 1998 NHANES III survey was used because it provides the separate race and Hispanicity distributions that are needed to compare with the other studies. The most recent continuous NHANES survey (2009-2010) only provides a combined race/Hispanicity variable.

Table 7.3 Weighted Distribution of Race

Race	NSDUH[1]	NHIS[2]	NHANES[3]	ACS[4]
White	80.6	80.30	82.66	74.59
Black/African American	12.2	13.40	12.83	12.38
Other or Two or More Races	7.2	6.30	4.51	13.03

[1] Data are from the 2009 NSDUH Restricted-Use Analytic File. Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

[2] Data are from the 2010 NHIS Public Use Microdata file from DataFerrett: https://dataferrett.census.gov/LaunchDFA.html.

[3] Data are from the NHANES III updated Household data files (1998) from DataFerrett.

[4] Data are from the 2007-2009 ACS 3-year estimates Public Use Microdata Sample from DataFerrett.

The race distributions for only Hispanic/Latino participants are shown in Table 7.4. The weighted white, black/African American, and other race distributions among Hispanics/Latinos, are similar between the NSDUH and NHIS. The NHANES III has a slightly smaller white proportion (88.52 percent versus 92.5 percent), whereas the ACS race distribution among Hispanics/Latinos is drastically different from the other three surveys. The ACS has a much smaller proportion of white Hispanic/Latino (59.80 percent versus about 92 percent) and a larger proportion of other or two or more races (38.33 percent versus about 5 percent). The differences in these proportions for two or more races may be related to how the ACS codes multiple races by allowing respondents to provide two, three, or more races. If respondents are included in the multiple race categories, then the proportions for white and black/African American are smaller for the ACS as seen in Tables 7.3 and 7.4.

Table 7.4 Weighted Distribution for Race among Hispanic/Latino

Race	NSDUH[1]	NHIS[2]	NHANES[3]	ACS[4]
White	92.5	92.05	88.52	59.80
Black/African American	2.0	3.72	4.53	1.79
Other or Two or More Races	5.5	4.23	6.95	38.33

[1] Data are from the 2009 NSDUH Restricted-Use Analytic File. Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

[2] Data are from the 2010 NHIS Public Use Microdata file from DataFerrett: https://dataferrett.census.gov/LaunchDFA.html.

[3] Data are from the NHANES III updated Household data files (1998) from DataFerrett.

[4] Data are from the 2007-2009 ACS 3-year estimates Public Use Microdata Sample from DataFerrett.

7.3.5 Item Nonresponse Rates

Table 7.5 shows the unweighted item nonresponse rates for Hispanicity and race separately and for those participants missing race who identify themselves as Hispanic/Latino. Although the item nonresponse rates are different for all four studies, the patterns are the same. For all studies, the percentage missing Hispanicity is lower than the percentage missing race. In addition, the percentage missing race within Hispanic/Latino is much higher than the overall percentage missing race.

Table 7.5 Unweighted Item Nonresponse Rates for Race and Hispanicity

Race/Hispanicity	NSDUH[1]	NHIS[2]	NHANES[3]	ACS[4]
Missing Hispanicity	0.14	0.07	Less than 10%	0.60
Missing Race	4.18	11.50	About 10%	5.50
Missing Race among Hispanic/Latino	21.51	43.70	About 25%	Although exact numbers are not available, ACS staff indicated that the proportion of respondents who are missing race but identified as Hispanic/Latino is much higher than among those respondents who identify themselves as non-Hispanic/Latino.

[1] Information is from the 2011 imputation report of the NSDUH methodological resource book (Frechtel et al., 2013).

[2] Information is from the 2010 NHIS Public Use Microdata file and e-mails and phone conversations with the Office of Surveillance, Epidemiology and Laboratory Services, National Center for Health Statistics.

[3] Information is from e-mails and phone conversations with the National Center for Health Statistics.

[4] Information is from e-mails and phone conversations with the U.S. Census Bureau.

Table 7.6 provides the weighted item nonresponse rates for Hispanicity and race separately and for those participants missing race who identify themselves as Hispanic/Latino. The percentages missing Hispanicity are very similar and are lower than 1.00 percent. The overall percentages missing race are similar as well, between 4.20 percent and about 10 percent. However, the percentage of Hispanic/Latino missing race is much higher in the NHIS (42.32 percent) than the other studies (about 25 percent).

Table 7.6 Weighted Item Nonresponse Rates for Race and Hispanicity

Race/Hispanicity	NSDUH[1]	NHIS[2]	NHANES[3]	ACS[4]
Missing Hispanicity	0.06	0.06	Less than 10%	0.60
Missing Race	4.20	7.22	About 10%	5.50
Missing Race among Hispanic/Latino	22.37	42.32	About 25%	Although exact numbers are not available, ACS staff indicated that the proportion of respondents who are missing race but identified as Hispanic/Latino is much higher than among those respondents who identify themselves as non-Hispanic/Latino.

[1] Information is from the 2011 imputation report of the NSDUH methodological resource book (Frechtel et al., 2013).

[2] Information is from the 2010 NHIS Public Use Microdata file and e-mails and phone conversations with the Office of Surveillance, Epidemiology and Laboratory Services, National Center for Health Statistics.

[3] Information is from e-mails and phone conversations with the National Center for Health Statistics.

[4] Information is from e-mails and phone conversations with the U.S. Census Bureau.

7.3.6 Summary

Although all four surveys impute race and Hispanicity separately with a hot-deck procedure, each study uses different information to impute both variables. With the exception of the NHANES III, the weighted distributions of race/ethnicity are relatively similar. In addition, with the exception of the ACS, the weighted distribution of Hispanicity and race within Hispanic/Latino are similar. All four studies have a very small proportion missing Hispanicity (less than 1 percent), a moderate proportion missing race (4 percent to 11 percent), and a large proportion of Hispanic/Latino respondents missing race (about 23 percent to 42 percent). Although none of the survey representatives outright said they were not confident in their imputation methods, there was indication that confidence in imputing Hispanicity and race within non-Hispanic/Latino was higher than imputing race for individuals who identify themselves as Hispanic/Latino.

7.4 Assessing the Feasibility of Using NSDUH Screener Data for Imputation

Since the 1999 NSDUH, PMN has been used to impute the interview race/ethnicity variables, and this method does not use the NSDUH screener data to help with editing of the missing race/ethnicity data. Chapter 3 of the 2011 imputation report of the NSDUH methodological resource book (MRB; Frechtel et al., 2013) describes the details of the editing and imputation methods for race/ethnicity. In PMN, the race and Hispanic/Latino origin variables are imputed using a logistic regression model where the screener race and Hispanicity information is used as a predictor variable in the form of a three-level household variable (Hispanic/Latino, non-Hispanic/Latino black/African American, or non-Hispanic/Latino non-black/African American). Besides this utilization, interview race/ethnicity variables are edited and imputed independently of the screener race/ethnicity data.

This section describes an investigation of the use of screener and interview data in the editing and imputation procedures for race and Hispanic/Latino origin. It discusses the feasibility of implementing a deterministic imputation procedure for race and Hispanic/Latino origin and presents possible scenarios for evaluation. The following questions describe the main focus of the evaluation.

- How often does age differ between the screener and interview data?
- Can screener data be used to edit and impute missing values for race and Hispanic/Latino origin?
- Would a household member's proxy report of interview respondent race or Hispanic/Latino origin be a better impute than a random PMN donor selection?

7.4.1 Screener and Interview Race and Hispanic/Latino Origin

Table 7.7 shows the item response count and unweighted distribution based on the 2009 interview and screener data for race and Hispanic/Latino origin variables. Screener Hispanic/Latino origin was missing only 33 cases (0.05 percent), but interview Hispanic/Latino origin was missing 116 cases (0.17 percent). Similar to the Hispanic/Latino origin variable, interview race had a higher rate of missing cases in the interview data than in the screener data. The screener data had only 30 cases missing race (0.04 percent), but the interview data had nearly 100 times the number of cases missing race (2,987 cases, or 4.35 percent). Also note that race tends to have a much higher missing rate among the Hispanic/Latino respondents than the non-Hispanic/Latino respondents. This is especially true among the interview Hispanic/Latino respondents, where 23.12 percent of the Hispanics/Latinos had missing race compared with only 0.85 percent of the non-Hispanics/Latinos with missing race information.

Nonmissing Hispanic/Latino origin and race between the two data sources were compared to find out whether screener and interview data have similar distributions. Table 7.8 shows that both data sources had a very similar distribution for reported Hispanic/Latino origin: approximately 10,500 cases (15.3 percent) for Hispanic/Latino in the screener data and about 10,800 cases (15.7 percent) for Hispanic/Latino in the interview data. The distribution for the reported race is similar for most of the race categories. However, the distribution of white and American Indian/Alaska Native varied by 2 to 3 percentage points. The reported race for white decreased from about 54,100 cases (78.8 percent) in the screener data to around 49,600 (75.5 percent) in the interview data. In contrast, the reported race for American Indian/Alaska Native increased from approximately 900 cases (1.4 percent) in the screener data to nearly 2,400 cases (3.6 percent) in the interview data. Table 7.8 also shows the percentage difference based on the interview data. The percentage difference for two groups, American Indian/Alaska Native and Native Hawaiian/Other Pacific Islander, was considerably higher than the other race categories. If the screener data were to be used as a proxy for interview race, the interview race categories American Indian/Alaska Native and Native Hawaiian/Other Pacific Islander would be underestimated.

Table 7.7 Hispanic/Latino Origin and Race Item Response Summary, by Screener Data and Interview Data, 2009 NSDUH

	Total Completed Interviews		Reported Values		Logically Assigned or Imputed Values	
	Frequency	Unweighted Percentage	Frequency	Unweighted Percentage	Frequency	Unweighted Percentage
Hispanic/Latino Origin						
Screener Data[1]	68,700	100.00	68,667	99.95	33	0.05
Interview Data[2]	68,700	100.00	68,584	99.83	116	0.17
Race						
Screener Data[1]	68,700	100.00	68,670	99.96	30	0.04
Interview Data[2]	68,700	100.00	65,713	95.65	2,987	4.35
Race among Nonmissing Hispanic/Latino						
Screener Data[1]	10,498	100.00	10,482	99.85	16	0.15
Interview Data[2]	10,759	100.00	8,272	76.88	2,487	23.12
Race among Nonmissing Non-Hispanic/Latino						
Screener Data[1]	58,169	100.00	58,164	99.99	5	0.01
Interview Data[2]	57,825	100.00	57,331	99.15	494	0.85

[1] The methodology used for imputing the screener Hispanic/Latino origin and race is documented in the 2009 person-level sampling weight calibration report of the NSDUH methodological resource book (MRB; Chen et al., 2011).

[2] The methodology used for imputing the interview Hispanic/Latino origin and race is documented in the 2009 MRB imputation report (Ault et al., 2011).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

Table 7.8 Comparison of Screener and Interview Race and Hispanic/Latino Origin Distributions among Respondents with No Missing Values in Screener and Interview Data, 2009 NSDUH

	Screener Data		Interview Data		Percent Difference Based on Interview Data[2]
	Frequency[1]	Unweighted Percentage[2]	Frequency[1]	Unweighted Percentage[2]	
Hispanic/Latino Origin					
Hispanic/Latino	10,500	15.3	10,800	15.7	-2.5
Non-Hispanic/Latino	58,200	84.7	57,800	84.3	0.5
Total	68,700	100.0	68,600	100.0	--
Race					
White	54,100	78.8	49,600	75.5	4.4
Black/African American	8,500	12.4	8,700	13.2	-5.9
American Indian/Alaska Native	900	1.4	2,400	3.6	-62.4
Native Hawaiian/Other Pacific Islander	300	0.4	400	0.7	-34.5
Asian	2,400	3.4	2,200	3.4	0.2
Two or More Races	2,500	3.6	2,400	3.6	-1.7
Total	68,700	100.0	65,700	100.0	--

-- Not available.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Percentages have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

7.4.2 Correlation between Reported Screener and Interview Hispanic/Latino Origin

Responses from the screening and interview were matched and compared to determine how well the two data sources agreed. Of the 67,000 interview cases, around 100 were missing Hispanic/Latino origin from at least one source. Table 7.9 shows the correlation between Hispanic/Latino origin screener and interview data for cases with both screener and interview data. When respondents do not have missing values in both screener and interview data, the agreement is quite strong for Hispanic/Latino (95.0 percent) and non-Hispanic/Latino (99.5 percent), though it is lower for the Hispanic/Latino subgroup. The table also shows that in screener interviews people are more likely to be classified as non-Hispanic/Latino, though they identify themselves as Hispanic/Latino in the interview than vice versa (5.1 percent versus 0.5 percent).

Table 7.9 Correlation between Screener and Interview Hispanic/Latino Origin Using Nonmissing Screener and Interview Data, 2009 NSDUH

Screener Hispanic/Latino Origin	Total Completed Interviews[1]	Interview Hispanic/Latino Origin			
		Hispanic/Latino		Non-Hispanic/Latino	
		Frequency[1]	Unweighted Percentage[2]	Frequency[1]	Unweighted Percentage[2]
Nonmissing Screener and Interview Data					
Hispanic/Latino	10,500	10,200	95.0	300	0.5
Non-Hispanic/Latino	58,100	500	5.1	57,500	99.5
Total	68,600	10,800	100.0	57,800	100.0

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Percentages have been rounded to the nearest tenth to ensure respondent confidentiality. Unweighted percentages are conditioned on dual respondent status.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

To investigate whether screener Hispanic/Latino origin can be used as the interview Hispanic/Latino origin when the interview data are missing, the level of agreement between the two sources was evaluated while considering whether the interview respondent was also the screener respondent (or a dual respondent). As presented in Table 7.1, the majority of interview respondents (68.22 percent) were not positively identified as the screener respondents, and it is expected that Hispanic/Latino origin agreement would be lower for non-dual respondents than for dual respondents. Table 7.10 shows the correlation between nonmissing screener and interview Hispanic/Latino origin by dual and non-dual respondents. Similar to data presented in Table 7.9, dual respondents had higher Hispanic/Latino origin agreement rates, with 97.6 percent for Hispanic/Latino and 99.8 percent for non-Hispanic/Latino. The Hispanic/Latino origin agreement rates are slightly lower for non-dual respondents: 94.0 percent for Hispanic/Latino and 99.4 percent for non-Hispanic/Latino.

Table 7.10 Correlation between Nonmissing Screener and Interview Hispanic/Latino Origin among Dual and Non-Dual Respondents, 2009 NSDUH

Screener Hispanic/Latino Origin	Total Completed Interviews[1]	Interview Hispanic/Latino Origin			
		Hispanic/Latino		Non-Hispanic/Latino	
		Frequency[1]	Unweighted Percentage[2]	Frequency[1]	Unweighted Percentage[2]
Dual Respondents					
Hispanic/Latino	2,900	2,800	97.6	< 50	0.3
Non-Hispanic/Latino	19,000	100	2.5	18,900	99.8
Total	21,800	2,900	100.0	18,900	100.0
Non-Dual Respondents					
Hispanic/Latino	7,600	7,400	94.0	200	0.6
Non-Hispanic/Latino	39,100	500	6.0	38,600	99.4
Total	46,700	7,900	100.0	38,900	100.0

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Percentages have been rounded to the nearest tenth to ensure respondent confidentiality. Unweighted percentages are conditioned on dual respondent status.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

Based on the two above tables, it appears that the rate of misclassification of people identified as Hispanic/Latino during the interview is lower when the same person responds to both the screening and the interview (2.5 percent versus 6.0 percent). However, overall the information presented in Tables 7.9 and 7.10 demonstrates that the interview-reported Hispanic/Latino origin strongly agrees with the screener-reported Hispanic/Latino origin. In addition, proxy reports for Hispanic/Latino origin are highly reliable. This association suggests that it may be feasible to deterministically impute missing interview Hispanic/Latino origin with screener data because there is lower missingness in screener data than in interview data (so a large portion of respondents could inherit the Hispanic/Latino origin from the screener data). However, this would only be true if the deterministic method performs better than the current PMN method or any other method. The screener data are not collected as rigorously as the interview data and thus may not be as reliable as desired for performing deterministic imputation. It is not certain that the deterministic imputation using screener data is any better than an imputation from the PMN method. The results of a deterministic imputation compared with the results of a PMN imputation may produce slightly different results. Section 7.5 evaluates the deterministic method and five other methods against the current method.

7.4.3 Correlation between Reported Screener and Interview Race

Table 7.11 shows the correlation between screener and interview race when neither screener race nor interview race was missing. The percentages presented are conditional on the screener race category; that is, the interview race percentages sum to 100 within each screener race category to describe the extent to which the interview percentage agrees with the associated screener race category.

Table 7.12 shows the distribution of the racial groups separately within Hispanic/Latino and non-Hispanic/Latino cases for the nonmissing data. The agreement rates for non-Hispanic/Latino respondents are relatively high for the race categories of white (99.2 percent), black/African American (96.9 percent), and Asian (92.2 percent), but they are more moderate for American Indian/Alaska Native (76.5 percent), Native Hawaiian/Other Pacific Islander (61.2 percent), and two or more races (66.9 percent). Therefore, if the moderate agreement among American Indian/Alaska Native, Native Hawaiian/Other Pacific Islander, and two or more races are acceptable, it is reasonable to assume that a deterministic imputation for interview race based on available screener race would be better suited for non-Hispanic/Latino cases because the correlation between the racial categories is strong.

In contrast, the correlation for Hispanic/Latino cases is not as strong, especially for the different nonwhite subgroups. Within Hispanic/Latino, black/African American cases agree 57.5 percent of the time, Asian cases agree 50.5 percent of the time, and two or more races agree 56.4 percent of the time. There is only 7.4 percent agreement for American Indian/Alaska Native and 14.9 percent agreement for Native Hawaiian/Other Pacific Islander. Although white cases agree 96.5 percent of the time, the screener race is concentrated in one interview race category, namely white. For instance, 88.8 percent of cases where the respondent reported being American Indian/Alaska Native in the interview were reported as white during the screening, and 76.6 percent of the cases where the respondent reported being Native Hawaiian/Other Pacific Islander in the interview were reported as white during the screening. This concentration of cases where the respondent reported white during the screening process would bias the race category toward

white if deterministic imputation (i.e., using the screener race as the interview race when the interview race is missing) were used.

7.4.4 Summary

Similar to Table 7.10, Table 7.13 shows the correlation between the screener and interview data for dual and non-dual respondents. This comparison describes how well the screener respondent agreed with the interview respondent for nonmissing interview race responses. The agreement rates for white, black/African American, and Asian are very high (99.4 percent, 97.0 percent, 93.7 percent, respectively) for dual respondents (i.e., respondents who answered both screener and interview questions). However, other racial groups had much lower agreement rates, particularly American Indian/Alaska Native (42.0 percent) and Native Hawaiian/Other Pacific Islander (50.0 percent). It should be noted that the questions on the screener and interview are very similar, but the screener race questions do not include the detailed Asian race questions (see sections 7.2.1.1 and 7.2.1.2). It is not clear why the same respondents classify themselves differently. The agreement rates follow a similar pattern for non-dual respondents, with high agreement rates for white, black/African American, and Asian (98.6 percent, 94.5 percent, and 88.9 percent, respectively) and even lower agreement rates for American Indian/Alaska Native (29.6 percent) and Native Hawaiian/Other Pacific Islander (39.2 percent). Therefore, using a deterministic imputation based on a proxy report for race may be better for white, black/African American, and Asian, compared with other race categories.

In summary, the interview respondents tend to be classified by the screeners as white, but classify themselves as one of the non-white race categories. This racial classification difference is especially pronounced among the Hispanic/Latino subgroup. However, screener race might still be a useful predictor of interview race instead of or in addition to variables currently used in PMN, including the three-level household variable and the segment-level racial population-related variables. Section 7.4.3 results also demonstrate that implementing a deterministic imputation would most likely not improve the imputation results of interview race and Hispanic/Latino origin. For interview race, using the screener race for deterministic imputation could lead to higher percentages of white and lower percentages of other race categories imputed. Because the missing rate is low for Hispanic/Latino origin, the impact of deterministic imputation is minimal.

Table 7.11 Correlation of Nonmissing Screener and Interview Race, 2009 NSDUH

Screener Race	Total[1]	Interview Race: White Frequency[1]	White Unweighted Percentage[2]	Black/African American Frequency[1]	Black/African American Unweighted Percentage[2]	American Indian/Alaska Native Frequency[1]	American Indian/Alaska Native Unweighted Percentage[2]	Native Hawaiian/Other Pacific Islander Frequency[1]	Native Hawaiian/Other Pacific Islander Unweighted Percentage[2]	Asian Frequency[1]	Asian Unweighted Percentage[2]	Two or More Races Frequency[1]	Two or More Races Unweighted Percentage[2]
Nonmissing Screener and Interview Data													
White	51,500	49,000	98.9	200	2.6	1,500	61.1	200	38.8	100	3.2	500	21.4
Black/African American	8,500	< 50	0.1	8,300	95.3	< 50	1.2	< 50	2.6	< 50	0.1	200	6.4
American Indian/Alaska Native	900	100	0.1	< 50	0.1	800	32.4	< 50	1.2	< 50	0.1	100	3.1
Native Hawaiian/Other Pacific Islander	300	< 50	0.0	< 50	0.0	< 50	0.1	200	42.1	< 50	2.1	100	2.1
Asian	2,100	< 50	0.0	< 50	0.0	< 50	0.4	< 50	4.0	2,000	90.4	< 50	1.8
Two or More Races	2,400	400	0.9	200	2.0	100	4.7	< 50	11.3	100	4.1	1,600	65.3
Total	65,700	49,600	100.0	8,662	100.0	2,400	100.0	400	100.0	2,200	100.0	2,400	100.0

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Percentages have been rounded to the nearest tenth to ensure respondent confidentiality. Unweighted percentages are conditioned on whether the interview and screener race were missing or not.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

Table 7.12 Correlation of Nonmissing Screener and Interview Race, by Hispanic/Latino Origin (as determined during the interview), 2009 NSDUH

Screener Race	Total[1]	Interview Race: White Frequency[1]	White Unweighted Percentage[2]	Black/African American Frequency[1]	Black/African American Unweighted Percentage[2]	American Indian/Alaska Native Frequency[1]	American Indian/Alaska Native Unweighted Percentage[2]	Native Hawaiian/Other Pacific Islander Frequency[1]	Native Hawaiian/Other Pacific Islander Unweighted Percentage[2]	Asian Frequency[1]	Asian Unweighted Percentage[2]	Two or More Races Frequency[1]	Two or More Races Unweighted Percentage[2]
Non-Hispanic/Latino from Interview Data													
White	44,200	43,500	99.2	100	1.5	100	12.5	< 50	12.4	100	2.7	400	19.1
Black/African American	8,300	< 50	0.1	8,100	96.9	< 50	2.3	< 50	3.6	< 50	0.1	100	7.1
American Indian/Alaska Native	800	< 50	0.1	< 50	0.1	700	76.5	< 50	0.4	< 50	0.1	100	3.2
Native Hawaiian/Other Pacific Islander	200	< 50	0.0	< 50	0.0	< 50	0.2	200	61.2	< 50	2.0	< 50	2.1
Asian	2,100	< 50	0.0	< 50	0.0	< 50	0.8	< 50	6.4	2,000	92.2	< 50	1.7
Two or More Races	1,900	300	0.6	100	1.6	100	7.6	< 50	16.0	100	2.9	1,400	66.9
Total	57,400	43,800	100.0	8,300	100.0	900	100.0	300	100.0	2,100	100.0	2,000	100.0
Hispanic/Latino from Interview Data													
White	7,300	5,600	96.5	100	29.8	1,300	88.8	100	76.6	< 50	14.4	100	34.5
Black/African American	200	< 50	0.2	200	57.5	< 50	0.5	< 50	1.1	< 50	0.0	< 50	2.5
American Indian/Alaska Native	200	< 50	0.5	< 50	0.3	100	7.4	< 50	2.3	< 50	0.0	< 50	2.7
Native Hawaiian/Other Pacific Islander	< 50	< 50	0.0	< 50	0.3	< 50	0.1	< 50	14.9	< 50	6.2	< 50	1.9
Asian	100	< 50	0.0	< 50	0.0	< 50	0.1	< 50	0.6	< 50	50.5	< 50	1.9
Two or More Races	500	200	2.8	< 50	12.1	< 50	3.1	< 50	4.6	< 50	28.9	200	56.4
Total	8,300	5,800	100.0	300	100.0	1,500	100.0	200	100.0	100	100.0	400	100.0

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Percentages have been rounded to the nearest tenth to ensure respondent confidentiality. Unweighted percentages are conditioned on whether the interview and screener race were missing or not.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

Table 7.13 Correlation of Nonmissing Screener and Interview Race among Dual and Non-Dual Respondents, 2009 NSDUH

Screener Race	Total[1]	Interview Race: White Frequency[1]	White Unweighted Percentage[2]	Black/African American Frequency[1]	Black/African American Unweighted Percentage[2]	American Indian/Alaska Native Frequency[1]	American Indian/Alaska Native Unweighted Percentage[2]	Native Hawaiian/Other Pacific Islander Frequency[1]	Native Hawaiian/Other Pacific Islander Unweighted Percentage[2]	Asian Frequency[1]	Asian Unweighted Percentage[2]	Two or More Races Frequency[1]	Two or More Races Unweighted Percentage[2]
Dual Respondents													
White	16,600	16,100	99.4	100	1.8	300	52.9	100	46.6	< 50	2.5	100	15.2
Black/African American	2,700	< 50	0.0	2,700	97.0	< 50	1.3	< 50	1.7	< 50	0.0	< 50	3.6
American Indian/Alaska Native	300	< 50	0.1	< 50	0.1	200	42.0	< 50	0.0	< 50	0.1	< 50	3.3
Native Hawaiian/Other Pacific Islander	100	< 50	0.0	< 50	0.0	< 50	< 50	100	50.0	< 50	1.8	< 50	1.2
Asian	700	< 50	0.0	< 50	0.0	< 50	0.4	< 50	0.9	700	93.7	< 50	0.7
Two or More Races	600	100	0.5	< 50	1.0	< 50	3.5	< 50	0.9	< 50	1.9	400	76.0
Total	20,900	16,200	100.0	2,800	100.0	500	100.0	100	100.0	700	100.0	600	100.0
Non-Dual Respondents													
White	34,900	32,900	98.6	200	2.9	1,200	63.6	100	35.9	100	3.5	400	23.4
Black/African American	5,800	< 50	0.1	5,600	94.5	< 50	1.1	< 50	2.9	< 50	0.1	100	7.3
American Indian/Alaska Native	700	100	0.2	< 50	0.0	500	29.6	< 50	1.6	< 50	0.1	100	3.1
Native Hawaiian/Other Pacific Islander	200	< 50	0.0	< 50	0.1	< 50	0.2	100	39.2	< 50	2.3	< 50	2.4
Asian	1,400	< 50	0.1	< 50	0.0	< 50	0.4	< 50	5.2	1,300	88.9	< 50	2.1
Two or More Races	1,800	300	1.0	100	2.4	100	5.1	< 50	15.2	100	5.1	1,100	61.8
Total	44,800	33,400	100.0	5,900	100.0	1,800	100.0	300	100.0	1,500	100.0	1,800	100.0

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Percentages have been rounded to the nearest tenth to ensure respondent confidentiality. Unweighted percentages are conditioned on whether the interview and screener race were missing or not.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

7.5 Imputation Evaluation of Race and Hispanicity Using Alternative Imputation Methods

7.5.1 Overview

Based on the findings presented in Section 7.4, five alternative imputation methods were chosen to investigate (1) the impact of using the screener race and Hispanic/Latino origin variables as covariates in imputation models to impute the interview race and Hispanic/Latino origin variables; (2) the outcome of using stochastic model-based imputation (i.e., the method described in Section 5.1.1, but without the data augmentation step); and (3) the impact of imputing the interview Hispanic/Latino origin variable before the interview race variable, which reversed the order of imputation used in the current PMN method. Three variables (IRHOIND, IRNWRACE, and IRHOGRP4) directly comparable with the PMN imputed variables were imputed using the selected alternative imputation methods. In addition, a recoded race and Hispanic/Latino variable NEWRACE2 was created using IRHOIND and IRNWRACE imputed from the alternative imputation methods. Following are the descriptions and levels of these variables:

Imputation-Revised Hispanic/Latino Origin (IRHOIND)

1 Hispanic/Latino
2 Non-Hispanic/Latino

Imputation-Revised Race (IRNWRACE)

1 White
2 Black/African American
3 American Indian/Alaska Native
4 Native Hawaiian
5 Other Pacific Islander
6 Native Hawaiian/Other Pacific Islander
7 Chinese
8 Filipino
9 Japanese
10 Asian Indian
11 Korean
12 Vietnamese
13 Other Asian
14 Asian Multiple Categories
15 Two or More Races

Imputation-Revised Hispanic/Latino Origin Group (IRHOGRP4)

1 Puerto Rican
2 Mexican
3 Cuban
4 Other

5 Central/South American
6 Dominican
7 Spanish (from Spain)
99 LEGITIMATE SKIP, Respondent Is Not Hispanic/Latino

Race/Hispanicity Recode (NEWRACE2)

1 Non-Hispanic/Latino White
2 Non-Hispanic/Latino Black/African American
3 Non-Hispanic/Latino Native American/Alaska Native
4 Non-Hispanic/Latino Native Hawaii/Other Pacific Islander
5 Non-Hispanic/Latino Asian
6 Non-Hispanic/Latino Two or More Races
7 Hispanic/Latino

The five alternative imputation methods to obtain these variables were tested using the 2010 NSDUH data. Most of the investigation focuses on the correlation between the Hispanic/Latino origin variable (IRHOIND) and the race variable (IRNWRACE). The Hispanic/Latino group variable (IRHOGRP4) was only imputed in Method 5 in order to gain more insight on how Hispanic/Latino origin influences race imputation results. The layout of the staggered changes in each method allowed the performance of a controlled comparison with the current PMN method, and the comparisons are summarized in Table 7.14. The five methods are outlined below and described in detail in Section 7.5.2. The comparison of the test results with the current PMN imputation results are summarized in Section 7.5.3.

- **Method 1:** Imputing Hispanic/Latino origin before race using PMN.
- **Method 2:** Imputing Hispanic/Latino origin before race using PMN and adding screener Hispanic/Latino origin and race variables in the models.
- **Method 3:** Imputing Hispanic/Latino origin before race using model-based imputation and adding screener Hispanic/Latino origin and race variables in the models.
- **Method 4:** Imputing Hispanic/Latino origin before race using model-based imputation and adding screener Hispanic/Latino origin and race variables in the models, and then repeating the imputation process one more time to get the final imputed Hispanic/Latino origin and race.
- **Method 5:** Adding screener Hispanic/Latino origin and race variables in the models and repeating the imputation process for Hispanic/Latino origin, race, and Hispanic/Latino group, which are cycled through twice.

Table 7.14 Alternative Imputation Methods Comparison Summary

Method	Imputation Method	Cycling	Variables Imputed (in order)	Additional Covariates
1	PMN	No	IRHOIND IRNWRACE	None
2	PMN	No	IRHOIND IRNWRACE	Screener Hispanic/Latino Origin, Screener Race
3	Stochastic*	No	IRHOIND IRNWRACE	Screener Hispanic/Latino Origin, Screener Race
4	Stochastic	Yes	IRHOIND IRNWRACE	Screener Hispanic/Latino Origin, Screener Race
5	Stochastic	Yes	IRHOIND IRNWRACE IRHOGRP4	Screener Hispanic/Latino Origin, Screener Race

PMN = predictive mean neighborhood.

*Stochastic imputation used here follows four steps: (1) adjust sampling weights for item nonresponse; (2) fit a polytomous logistic regression model; (3) estimate the probability associated with each level of the outcome variable; (4) impute a value for each item nonrespondent using the probability from step 3.

7.5.2 Descriptions of Alternative Imputation Methods

7.5.2.1 Method 1

In Method 1, the current PMN imputation was used, with a change in the order of the variables that were imputed: Hispanic/Latino origin (IRHOIND) and race (IRNWRACE). Specifically, Hispanic/Latino origin was imputed before race, and then used as a covariate in the subsequent race imputation model in PMN. This method is intuitively rational because the missing rate for the Hispanic/Latino origin question (0.14 percent) is much lower than the missing rate for the race question (3.45 percent).[40] This implies that the imputation of the Hispanic/Latino origin variable is more straightforward. Also, in the interview questionnaire, the question of Hispanic/Latino origin comes before the question about race. Finally, it was thought that the Hispanic/Latino origin would be a good predictor for imputing race.

Method 1 was used to test whether switching the imputation order of Hispanic/Latino origin and race variables could improve race imputation models in PMN. Testing of Method 1 was conducted separately within three age groups: 12 to 17, 18 to 15, and 26 or older. The Wald statistic results of using the imputed interview Hispanic/Latino origin are summarized in Table 7.15. The results show that the imputed Hispanic/Latino origin carries fairly strong predictive power for race.

[40] Missing rates are based on the 2010 NSDUH analytic results.

Table 7.15 Method 1: Race Predictive Mean Model Wald Statistics Summary

Age Group	Imputed Hispanic/Latino Origin Level	Wald F	P-Value	Rank of Wald F (relative to other covariates)
12-17	Hispanic/Latino	115.94	0.00	2 (out of 16)
18-25	Hispanic/Latino	146.36	0.00	1 (out of 17)
26+	Hispanic/Latino	132.31	0.00	1 (out of 16)

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

7.5.2.2 Method 2

Method 2 is identical to Method 1 but includes the Hispanic/Latino origin and race variables from the screener data as covariates.[41] This approach takes advantage of the insight gained from the analysis provided in Section 7.4.3, which concluded that there is a high degree of correlation between the race and Hispanic/Latino origin variables from the screener data and the race and Hispanic/Latino variables from the interview data. Because these variables are correlated, it is reasonable to assume that the screener data variables are good explanatory variables for the corresponding variables in the interview data. The screener data variables thus would improve the predictive power of the predictive mean model.

In practice, most of the covariates from the screener had to be removed from the predictive mean model because of convergence problems. In all three age groups of race models, the screener race levels for white, black/African American, American Indian/Alaska Native, Other Pacific Islander, and Asian were included in the starting covariate list, and the screener race level for two or more races was used as a reference cell. However, the only screener level that was kept in the model was white because it is a strong predictor for interview race, as shown in Table 7.16. The imputed Hispanic/Latino origin had to be removed from the race models for all three age groups when screener race was included.

Table 7.16 Method 2: Race Predictive Mean Model Wald Statistics Summary

Age Group	Screener Race Level	Wald F	P-Value	Rank of Wald F
12-17	White	472.12	0.00	1 (out of 16)
18-25	White	469.15	0.00	1 (out of 17)
26+	White	443.17	0.00	1 (out of 16)

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

7.5.2.3 Method 3

Method 3 uses a model-based stochastic imputation method in lieu of the hot-deck imputation used in the current PMN method. In Method 3, the screener Hispanic/Latino origin variable and race variable are used as covariates in the models, similar to Method 2. Method 3 was also tested within the three age groups, as was done for Methods 1 and 2. Method 3 is similar to the univariate version of the modified predictive mean neighborhood multiple

[41] The screener race variable and screener Hispanic/Latino origin variable used in testing are imputed when necessary.

imputation (modPMN-MI) method, as described in Chapter 5, but without the data augmentation step.

For Hispanic/Latino origin (IRHOIND), a binary variable, the stochastic imputation was based on a logistic regression that simplifies the PMN algorithm by utilizing the predicted mean of a given observation as a direct indicator of the probability of that observation. For the respondent with a missing Hispanic/Latino origin value, a random value is selected from a uniform distribution representing the range of the probabilities. The value is compared with the predicted mean of that observation. If the random value is less than or equal to the predicted mean, then IRHOIND of this observation is assigned to Hispanic/Latino. Otherwise, it is assigned to non-Hispanic/Latino.

For the detailed race variable (IRNWRACE), a polytomous logistic regression model was used. The imputation process is similar to the one for IRHOIND but is more complex due to the large number of categories (15 levels) that the imputation model has to fit. It was not possible to obtain 15 levels of predicted means from the model, because certain levels of the race variable were quite sparse and, as a result, the model would not converge.[42]

A two-step model-based imputation approach was designed to resolve this issue. All Asian specific, Native Hawaiian, and Other Pacific Islander categories were combined into one general Asian/Native Hawaiian/Other Pacific Islander level. This created a five-level response variable that included (1) white only, (2) black/African American only, (3) American Indian/Alaska Native only, (4) Asian/Native Hawaiian/Other Pacific Islander, and (5) two or more races. As before, the predicted mean for each level was assumed to be a direct indicator of the probability of the observation. A random value was selected from a table distribution derived from the predicted means, which represented the range of probabilities for each level. A value of 1 to 5 was then selected based on the table distribution, and this number was given as the imputed race level for the observation. If the imputed race for the observation was not Asian/Native Hawaiian/Other Pacific Islander (level 4), the process was complete. For observations where the race was imputed as Asian, a second imputation step was done. The second step repeated the operation of the first step except that the finer Asian categories were imputed based on the weighted frequencies derived from nonmissing data instead of the predicted mean. The finer Asian categories include Native Hawaiian, Other Pacific Islander, Native Hawaiian/Other Pacific Islander, Chinese, Filipino, Japanese, Asian Indian, Korean, Vietnamese, Other Asian, and Asian multiple categories. The results of these two imputation steps were combined to create the final imputed variable (IRNWRACE).

Because all steps except the final imputation step in Method 3 (Table 7.17) are similar to those in Method 2, the model Wald statistics shared almost identical results to the Method 2 results. Also, the imputed Hispanic/Latino origin had to be removed from the race models for all three age groups when screener race was included.

[42] One way to deal with the sparse cells is to use several years of data to fit the model. One problem with this approach is that racial demographics in the United States have changed over time. See Section 3.4 of the 2008 MRB imputation report (Ault et al., 2010). To account for this, the models would have to include covariates for the year and any significant interactions of the year and the other covariates. The model selection process might become time-consuming.

Table 7.17 Method 3: Race Predictive Mean Model Wald Statistics Summary

Age Group	Screener Race Level	Wald F	P-Value	Rank of Wald F
12-17	White	472.12	0.00	1 (out of 16)
18-25	White	469.15	0.00	1 (out of 17)
26+	White	443.18	0.00	1 (out of 16)

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

7.5.2.4 Method 4

Method 4 is defined as a cyclic model-based stochastic imputation method. It employed a similar approach to Method 3 except that the steps of Hispanic/Latino origin imputation and race imputation were repeated after the first round of imputation. Method 4 is similar to the multivariate version of modPMN-MI, as described in Chapter 5, but, again, without the data augmentation step. In the second cycle, the imputed value of race as determined from the first cycle was used as a covariate in the final imputation models for Hispanic/Latino origin. The final imputed value of Hispanic/Latino was in turn used as a covariate for imputing the final race value. The final Asian/Native Hawaiian/Other Pacific Islander race categories were not imputed during the first cycle. If the final imputed race was Asian, then the finer Asian category imputation method described in Method 3 was performed. The interview Hispanic/Latino origin variable and interview race variable are good explanatory variables for each other, and including them as covariates in the cycling scheme could make better predicted models. Method 4 was tested separately within the three age groups.

The modeling steps from Cycle 1 were identical to Methods 2 and 3 and thus produced similar Wald statistics. The intermediate imputed Hispanic/Latino origin was included in the race model, but it had to be removed because of convergence problems in Cycle 1. However, the final imputed Hispanic/Latino origin was kept in the race models for all three age groups in Cycle 2. Table 7.18 demonstrates that both the screener race level for white and the imputed Hispanic/Latino origin are strong predictors for interview race.

Table 7.18 Method 4: Race Predictive Mean Model Wald Statistics Summary

Age Group	Cycle 1				Cycle 2							
	Screener Race Level	Wald F	P-Value	Rank of Wald F	Screener Race Level	Wald F	P-Value	Rank of Wald F	Imputed Hispanic/ Latino Origin Level	Wald F	P-Value	Rank of Wald F
12-17	White	472.01	0.00	1 (out of 16)	White	484.59	0.00	1 (out of 17)	Hispanic/ Latino	123.29	0.00	2 (out of 17)
18-25	White	469.15	0.00	1 (out of 17)	White	482.77	0.00	1 (out of 18)	Hispanic/ Latino	145.72	0.00	2 (out of 18)
26+	White	443.20	0.00	1 (out of 16)	White	461.28	0.00	1 (out of 17)	Hispanic/ Latino	135.68	0.00	2 (out of 17)

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

7.5.2.5 Method 5

Method 5 was motivated by the observation (Section 7.4.3) that the majority of the missing race respondents are of Hispanic/Latino origin. This implies that the race information obtained from the Hispanic/Latino group variable may be of value as a predictor for race. The method proceeds as in Method 4, until the step of the final Hispanic/Latino origin. Instead of imputing the final race variable, the Hispanic/Latino group variable was then imputed using a polytomous logistic regression model with the final imputed Hispanic/Latino origin and imputed race from the first step as covariates. The result of this imputation is labeled as intermediate imputed Hispanic/Latino group. The intermediate imputed Hispanic/Latino group variable was used as a covariate in the final imputation of race, followed by the final imputation of the Hispanic/Latino group with the final imputed race in the model.

Because of the small domain of the Hispanic/Latino group variable, age groups are combined in imputing IRHOGRP4. Also, because of the sparsity of some levels in the Hispanic/Latino group variable, levels were collapsed to form a four-level response variable for the imputation model: (1) Mexican, (2) Puerto Rican, (3) Other, and (4) Cuban. Adapted from the two-step model-based imputation technique for imputing finer Asian/Native Hawaiian/Other Pacific Islander categories in Methods 3 and 4, the Hispanic/Latino group variable was first imputed to one of the four levels. If the imputed value fell to category 3, then one of the four levels (Central/South American, Dominican, Spanish from Spain, and Other Hispanic/Latino) was assigned. The finer Hispanic/Latino group was imputed based on the weighted frequencies of the four levels derived from nonmissing data. Both the race finer categories and Hispanic/Latino group finer categories are imputed during the last cycle of imputation.

Again, the outcome of the Cycle 1 race modeling was identical to Methods 2 and 3 and to Method 4, Cycle 1. The intermediate imputed Hispanic/Latino origin had to be removed from the Cycle 1 race model because of convergence problems. In Cycle 2, the intermediate imputed Hispanic/Latino groups were included in the final imputation of race. The levels included were Mexican, Puerto Rican, and Central/South American, and the level for Other Hispanic/Latino was used as a reference cell in the model. The final imputed Hispanic/Latino origin could be kept in the Cycle 2 race models. The only screener level that was kept in the model for both cycles was white. All Hispanic/Latino group levels remained in the final cycle of the race model. The screener race level for white and the imputed Hispanic/Latino group are strong predictors for interview race, as shown in Table 7.19.

Table 7.19 Method 5: Race Predictive Mean Model Wald Statistics Summary

Cycle 1				
Age Group	**Screener Race Level**	**Wald F**	**P-Value**	**Rank of Wald F**
12-17	White	472.02	0.00	1 (out of 16)
18-25	White	469.15	0.00	1 (out of 17)
26+	White	443.19	0.00	1 (out of 16)

Cycle 2								
Age Group	**Screener Race Level**	**Wald F**	**P-Value**	**Rank of Wald F**	**Intermediate Hispanic/ Latino Group Levels**	**Wald F**	**P-Value**	**Rank of Wald F**
12-17	None*	--	--	--	Mexican	100.28	0.000	2 (out of 18)
					Puerto Rican	13.38	0.000	8 (out of 18)
					Central/South American	88.63	0.000	3 (out of 18)
18-25	White	501.50	0.00	1 (out of 20)	Mexican	113.81	0.000	2 (out of 20)
					Puerto Rican	14.55	0.000	8 (out of 20)
					Central/South American	132.70	0.000	3 (out of 20)
26+	White	448.99	0.00	1 (out of 18)	Mexican	31.20	0.000	3 (out of 20)
					Puerto Rican	3.76	0.005	16 (out of 20)

*Because of convergence problems, no levels of screener race were used as covariates in the 12-17 age group interview race predictive mean model.

-- Not available.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

7.5.3 Imputation Results Summary and Comparisons

Tables for each of the imputed variables are provided in the sections below. In each section, the frequency table of complete data is followed by the frequency table of imputed data for all five methods. This in turn is followed by the unweighted and weighted frequency tables for all cases. The tables for NEWRACE2 are included to clarify the effect of the Hispanic/Latino race on the groupings.

7.5.3.1 Imputation-Revised Hispanic/Latino Origin (IRHOIND)

The results for the Hispanic/Latino origin variable (IRHOIND) are tabulated below. Tables 7.20 and 7.21 show that, for imputed cases, the unweighted percentage obtained by the current PMN method (11.2 percent) most closely approximates the unweighted percentage obtained in completed cases (16.0 percent). The unweighted percentages decrease from Method 1 to Method 5, thus making the Hispanic/Latino origin distribution more different from that of respondents in the interview data. These results demonstrate the advantage of the current PMN imputation method for the IRHOIND variable. Tables 7.22 and 7.23 show the unweighted and weighted effects of the imputation methods on all cases. Because there were only a few imputed cases, the effects of the imputation methods were negligible.

Table 7.20 IRHOIND, Completed Cases Only (Unweighted)

Level	Frequency[1]	Unweighted Percentage
Hispanic/Latino	10,900	16.0
Non-Hispanic/Latino	57,500	84.0
Total	68,400	100.0

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

Table 7.21 IRHOIND, Imputed Cases Only (Unweighted)

Level	Current PMN		Method 1		Method 2		Method 3		Method 4		Method 5	
	Frequency	Unweighted Percentage	Frequency	Unweighted Percentage	Frequency	Unweighted Percentage	Frequency	Unweighted Percentage	Frequency	Unweighted Percentage	Frequency	Unweighted Percentage
Hispanic/ Latino	11	11.2	10	10.2	9	9.2	7	7.1	7	7.1	7	7.1
Non-Hispanic/ Latino	87	88.8	88	89.8	89	90.8	91	92.9	91	92.9	91	92.9
Total	98	100.0	98	100.0	98	100.0	98	100.0	98	100.0	98	100.0

PMN = predictive mean neighborhood.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

Table 7.22 IRHOIND, All Cases (Unweighted)

Level	Current PMN		Method 1		Method 2		Method 3		Method 4		Method 5	
	Frequency[1]	Unweighted Percentage	Frequency[1]	Unweighted Percentage	Frequency[1]	Unweighted Percentage	Frequency[1]	Unweighted Percentage	Frequency[1]	Unweighted Percentage	Frequency[1]	Unweighted Percentage
Hispanic/ Latino	10,900	16.0	10,900	16.0	10,900	16.0	10,900	16.0	10,900	16.0	10,900	16.0
Non-Hispanic/ Latino	57,600	84.0	57,600	84.0	57,600	84.0	57,600	84.0	57,600	84.0	57,600	84.0
Total	68,500	100.0	68,500	100.0	68,500	100.0	68,500	100.0	68,500	100.0	68,500	100.0

PMN = predictive mean neighborhood.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

Table 7.23 IRHOIND, All Cases (Weighted)

Level	Current PMN		Method 1		Method 2		Method 3		Method 4		Method 5	
	Estimate (Thousands)[1]	Percentage	Estimate (Thousands)[1]	Percentage	Estimate (Thousands)[1]	Percentage	Estimate (Thousands)[1]	Percentage	Estimate (Thousands)[1]	Percentage	Estimate (Thousands)[1]	Percentage
Hispanic/ Latino	36,769	14.5	36,764	14.5	36,780	14.5	36,767	14.5	36,778	14.5	36,778	14.5
Non-Hispanic/ Latino	216,850	85.5	216,855	85.5	216,839	85.5	216,852	85.5	216,842	85.5	216,842	85.5
Total	253,619	100.0	253,619	100.0	253,619	100.0	253,619	100.0	253,619	100.0	253,619	100.0

PMN = predictive mean neighborhood.

[1] Estimates have been rounded to the nearest thousand to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

7.5.3.2 Imputation-Revised Race (IRNWRACE) and Hispanic/Latino/Race Recode (NEWRACE2)

This section summarizes the testing results of a 15-level detailed imputed race variable (IRNWRACE) and a Hispanic/Latino race recode (NEWRACE2). Tables 7.24 through 7.27 show the results of the various imputation methods on IRNWRACE, and Tables 7.28 through 7.30 show the results on NEWRACE2. Tables 7.26, 7.27, 7.29, and 7.30 show either weighted or unweighted results of these two variables on all respondents including both missing and nonmissing cases. Because of the relative small size of missing in race, when results are tabulated on all cases, either weighted or unweighted, the differences in results are barely detectable across different testing methods. The discussion below focuses on the imputed cases for Tables 7.24, 7.25, and 7.28.

Table 7.24 shows the unweighted distribution of the completed cases only by Hispanic/Latino origin for NEWRACE2. Black/African American consists of 13.4 percent of the total respondents, but among Hispanic/Latino, only 4.7 percent of respondents are black/African American. There are 3.6 percent of total respondents who are American Indian/Alaska Native, whereas the percentage increased to 16.8 percent among Hispanic/Latino. In Table 7.25, the most noticeable differences are the trend of increasing percentages of white for Methods 1 to 5. The current PMN method and Method 1 produced the lowest percentages of white, at about 69 percent, whereas the white percentages for Methods 3 to 5 are much higher. This trend is largely due to the use of the screener Hispanic/Latino origin and race variables in the imputation models. Results discussed in Section 7.4.3 showed that screener race has a higher percentage of whites (78.8 percent) than interview race (74.7 percent). Because of multicollinearity, when screener race is used in the model, the screener race indicator variable for the level "white" was usually kept in the model, whereas the rest of the levels had to be removed. Screener data have more influence on the imputed interview race in the model-based Methods 3 to 5, which have higher percentages of observations imputed as white than the PMN-based Method 2.

The percentage of observations imputed as white peaks for Method 3 (81.7 percent) and drops by 6 percentage points for Method 4 and by almost 9 percentage points for Method 5 (Table 7.25). One explanation for this change is that the inclusion of either the Hispanic/Latino origin variable or the Hispanic/Latino group variable improves the race model performance. Model summary results from Tables 7.18 and 7.19 in Section 7.5.2 indicate that Hispanicity has high predictive power for the missing interview race. Cycle 1 of Method 4 and Method 5 had similar results to Method 3, but during Cycle 2, interview Hispanicity information was used in the race model. This change is probably the cause for the difference in the outcomes of Methods 4 and 5 versus Method 3.

Also in Table 7.25, percentages of observations imputed as black/African American are relatively low, with 4.4 percent for Method 1 decreasing to 1.5 percent for Method 4 and increasing back up to 3.3 percent for Method 5, as compared with 13.4 percent of overall completed cases for black/African American shown in Table 7.24. This could be explained because most respondents with missing race were Hispanic/Latino.

As shown in Table 7.28, among approximately 2,400 missing race respondents, about 2,300 (98.9 percent) are Hispanic/Latino, but fewer than 50 cases (1.1 percent) are non-

Hispanic/Latino. Therefore, the distribution of imputed black/African American will align with the low black/African American distribution among nonmissing Hispanic/Latino respondents. Table 7.24 illustrates that for the respondents not missing race, the percentage of black/African American was 4.7 percent among Hispanics/Latinos, whereas it was 14.7 percent among non-Hispanics/Latinos. The decreasing trajectory of black/African American concentration from Method 1 to Method 5 follows the same reasoning behind the increasing concentration of whites from Method 1 to Method 5 discussed above. The variations of the American Indian/Alaska Native distribution can be explained using the same reasoning discussed above. Changes in distribution for the other racial categories are subtle across methods. When accounting for all cases, both the unweighted and weighted distribution changes are minimal across all methods. In summary, including screener race in the model could potentially impute higher percentages of white. The Hispanic/Latino group is a good predictor for imputing race, and including the Hispanic/Latino group variable in the race imputation process results in an imputed racial distribution that is mostly similar to the distribution of the interview respondents.

Table 7.24 IRNWRACE, Completed Cases Only (Unweighted)

	Hispanic/Latino		Non-Hispanic/Latino		Total	
Level	Frequency[1]	Unweighted Percentage	Frequency[1]	Unweighted Percentage	Frequency[1]	Unweighted Percentage
White	6,000	70.0	43,400	75.4	49,400	74.7
Black/African American	400	4.7	8,500	14.7	8,900	13.4
American Indian/Alaska Native	1,400	16.8	1,000	1.7	2,400	3.6
Native Hawaiian	< 50	0.4	100	0.2	200	0.2
Other Pacific Islander	100	1.4	200	0.3	300	0.5
Native Hawaiian/Other Pacific Islander	< 50	0.0	< 50	0.0	< 50	0.0
Chinese	< 50	0.2	500	0.8	500	0.7
Filipino	< 50	0.5	400	0.8	500	0.7
Japanese	< 50	0.2	100	0.2	100	0.2
Asian Indian	< 50	0.0	700	1.2	700	1.0
Korean	< 50	0.0	200	0.4	200	0.3
Vietnamese	< 50	0.1	200	0.3	200	0.3
Other Asian	< 50	0.1	200	0.3	200	0.3
Asian Multiple Categories	< 50	0.1	100	0.1	100	0.1
Two or More Races	500	5.6	2,000	3.6	2,500	3.8
Total	8,600	100.0	57,500	100.0	66,100	100.0

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

Table 7.25 IRNWRACE, Imputed Cases Only (Unweighted)

Level	Current PMN		Method 1		Method 2		Method 3		Method 4		Method 5	
	Frequency	Unweighted Percentage	Frequency	Unweighted Percentage	Frequency	Unweighted Percentage	Frequency	Unweighted Percentage	Frequency	Unweighted Percentage	Frequency	Unweighted Percentage
White	1,622	68.7	1,628	69.0	1,722	73.0	1,927	81.7	1,786	75.7	1,723	73.0
Black/African American	104	4.4	110	4.7	61	2.6	53	2.2	36	1.5	79	3.3
American Indian/Alaska Native	472	20.0	474	20.1	466	19.7	290	12.3	431	18.3	403	17.1
Native Hawaiian	8	0.3	5	0.2	3	0.1	2	0.1	1	0.0	4	0.2
Other Pacific Islander	33	1.4	27	1.1	41	1.7	4	0.2	4	0.2	3	0.1
Native Hawaiian/ Other Pacific Islander	0	0.0	1	0.0	0	0.0	1	0.0	0	0.0	0	0.0
Chinese	2	0.1	4	0.2	3	0.1	11	0.5	4	0.2	6	0.3
Filipino	6	0.3	7	0.3	3	0.1	5	0.2	2	0.1	8	0.3
Japanese	7	0.3	5	0.2	1	0.0	0	0.0	0	0.0	0	0.0
Asian Indian	5	0.2	5	0.2	2	0.1	10	0.4	12	0.5	12	0.5
Korean	1	0.0	0	0.0	0	0.0	1	0.0	2	0.1	3	0.1
Vietnamese	0	0.0	1	0.0	1	0.0	0	0.0	3	0.1	6	0.3
Other Asian	2	0.1	0	0.0	0	0.0	2	0.1	4	0.2	6	0.3
Asian Multiple Categories	1	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Two or More Races	97	4.1	93	3.9	57	2.4	54	2.3	75	3.2	107	4.5
Total	2,360	100.0	2,360	100.0	2,360	100.0	2,360	100.0	2,360	100.0	2,360	100.0

PMN = predictive mean neighborhood.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

Table 7.26 IRNWRACE, All Cases (Unweighted)

Level	Current PMN		Method 1		Method 2		Method 3		Method 4		Method 5	
	Frequency[1]	Unweighted Percentage	Frequency[1]	Unweighted Percentage	Frequency[1]	Unweighted Percentage	Frequency[1]	Unweighted Percentage	Frequency[1]	Unweighted Percentage	Frequency[1]	Unweighted Percentage
White	51,000	74.5	51,000	74.5	51,100	74.7	51,300	75.0	51,200	74.8	51,100	74.7
Black/African American	9,000	13.1	9,000	13.1	8,900	13.1	8,900	13.0	8,900	13.0	9,000	13.1
American Indian/Alaska Native	2,900	4.2	2,900	4.2	2,900	4.2	2,700	3.9	2,800	4.1	2,800	4.1
Native Hawaiian	200	0.2	200	0.2	200	0.2	200	0.2	200	0.2	200	0.2
Other Pacific Islander	300	0.5	300	0.5	300	0.5	300	0.4	300	0.4	300	0.4
Native Hawaiian/ Other Pacific Islander	< 50	0.0	< 50	0.0	< 50	0.0	< 50	0.0	< 50	0.0	< 50	0.0
Chinese	500	0.7	500	0.7	500	0.7	500	0.7	500	0.7	500	0.7
Filipino	500	0.7	500	0.7	500	0.7	500	0.7	500	0.7	500	0.7
Japanese	100	0.2	100	0.2	100	0.2	100	0.2	100	0.2	100	0.2
Asian Indian	700	1.0	700	1.0	700	1.0	700	1.0	700	1.0	700	1.0
Korean	200	0.3	200	0.3	200	0.3	200	0.3	200	0.3	200	0.3
Vietnamese	200	0.3	200	0.3	200	0.3	200	0.3	200	0.3	200	0.3
Other Asian	200	0.3	200	0.3	200	0.3	200	0.3	200	0.3	200	0.3
Asian Multiple Categories	100	0.1	100	0.1	100	0.1	100	0.1	100	0.1	100	0.1
Two or More Races	2,600	3.8	2,600	3.8	2,600	3.8	2,600	3.8	2,600	3.8	2,600	3.9
Total	68,500	100.0	68,500	100.0	68,500	100.0	68,500	100.0	68,500	100.0	68,500	100.0

PMN = predictive mean neighborhood.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

Table 7.27 IRNWRACE, All Cases (Weighted)

Level	Current PMN		Method 1		Method 2		Method 3		Method 4		Method 5	
	Estimate (Thousands)[1]	Percentage	Estimate (Thousands)[1]	Percentage	Estimate (Thousands)[1]	Percentage	Estimate (Thousands)[1]	Percentage	Estimate (Thousands)[1]	Percentage	Estimate (Thousands)[1]	Percentage
White	204,032	80.4	202,808	80.0	203,177	80.1	203,348	80.2	202,935	80.0	202,789	80.0
Black/ African American	31,168	12.3	31,193	12.3	31,163	12.3	31,131	12.3	31,080	12.3	31,173	12.3
American Indian/ Alaska Native	2,483	1.0	3,546	1.4	3,250	1.3	3,032	1.2	3,481	1.4	3,394	1.3
Native Hawaiian	183	0.1	223	0.1	180	0.1	179	0.1	185	0.1	212	0.1
Other Pacific Islander	671	0.3	707	0.3	744	0.3	650	0.3	670	0.3	651	0.3
Native Hawaiian/ Other Pacific Islander	3	0.0	4	0.0	3	0.0	5	0.0	3	0.0	3	0.0
Chinese	2,207	0.9	2,216	0.9	2,207	0.9	2,273	0.9	2,220	0.9	2,210	0.9
Filipino	2,205	0.9	2,217	0.9	2,219	0.9	2,204	0.9	2,214	0.9	2,221	0.9
Japanese	704	0.3	704	0.3	721	0.3	702	0.3	702	0.3	702	0.3
Asian Indian	3,674	1.4	3,692	1.5	3,666	1.4	3,686	1.5	3,721	1.5	3,679	1.5
Korean	915	0.4	907	0.4	907	0.4	942	0.4	912	0.4	938	0.4
Vietnamese	706	0.3	707	0.3	707	0.3	706	0.3	711	0.3	722	0.3
Other Asian	986	0.4	967	0.4	967	0.4	983	0.4	972	0.4	973	0.4
Asian Multiple Categorical	121	0.0	121	0.0	121	0.0	121	0.0	121	0.0	121	0.0
Two or More Races	3,559	1.4	3,606	1.4	3,586	1.4	3,657	1.4	3,692	1.5	3,829	1.5
Total	253,619	100.0	253,619	100.0	253,619	100.0	253,619	100.0	253,619	100.0	253,619	100.0

PMN = predictive mean neighborhood.

[1] Estimates have been rounded to the nearest thousand to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

Table 7.28 NEWRACE2, Imputed Non-Hispanic/Latino Cases Only (Unweighted)

Level	Current PMN		Method 1		Method 2		Method 3		Method 4		Method 5	
	Frequency	Unweighted Percentage	Frequency	Unweighted Percentage	Frequency	Unweighted Percentage	Frequency	Unweighted Percentage	Frequency	Unweighted Percentage	Frequency	Unweighted Percentage
White	14	0.6	12	0.5	16	0.7	17	0.7	16	0.7	16	0.7
Black/African American	4	0.2	6	0.3	2	0.1	2	0.1	2	0.1	2	0.1
American Indian/Alaska Native	0	0.0	0	0.0	2	0.1	1	0.0	0	0.0	0	0.0
Native Hawaiian/Other Pacific Islander	0	0.0	0	0.0	0	0.0	0	0.0	2	0.1	1	0.0
Asian	7	0.3	6	0.3	5	0.2	5	0.2	3	0.1	4	0.2
Two or More Races	0	0.0	1	0.0	0	0.0	0	0.0	2	0.1	2	0.1
Hispanic/Latino	2,335	98.9	2,335	98.9	2,335	98.9	2,335	98.9	2,335	98.9	2,335	98.9
Total	2,360	100.0	2,360	100.0	2,360	100.0	2,360	100.0	2,360	100.0	2,360	100.0

PMN = predictive mean neighborhood.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

Table 7.29 NEWRACE2, All Cases (Unweighted)

Level	Current PMN		Method 1		Method 2		Method 3		Method 4		Method 5	
	Frequency[1]	Unweighted Percentage	Frequency[1]	Unweighted Percentage	Frequency[1]	Unweighted Percentage	Frequency[1]	Unweighted Percentage	Frequency[1]	Unweighted Percentage	Frequency[1]	Unweighted Percentage
White	43,400	63.4	43,400	63.4	43,400	63.4	43,400	63.4	43,400	63.4	43,400	63.4
Black/African American	8,500	12.4	8,500	12.4	8,500	12.4	8,500	12.4	8,500	12.4	8,500	12.4
American Indian/Alaska Native	1,000	1.4	1,000	1.4	1,000	1.4	1,000	1.4	1,000	1.4	1,000	1.4
Native Hawaiian/Other Pacific Islander	300	0.5	300	0.5	300	0.5	300	0.5	300	0.5	300	0.5
Asian	2,300	3.4	2,300	3.4	2,300	3.4	2,300	3.4	2,300	3.4	2,300	3.4
Two or More Races	2,100	3.0	2,100	3.0	2,100	3.0	2,100	3.0	2,100	3.0	2,100	3.0
Hispanic/Latino	10,900	16.0	10,900	16.0	10,900	16.0	10,900	16.0	10,900	16.0	10,900	16.0
Total	68,500	100.0	68,500	100.0	68,500	100.0	68,500	100.0	68,500	100.0	68,500	100.0

PMN = predictive mean neighborhood.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

Table 7.30 NEWRACE2, All Cases (Weighted)

Level	Current PMN		Method 1		Method 2		Method 3		Method 4		Method 5	
	Estimate (Thousands)[1]	Weighted Percentage	Estimate (Thousands)[1]	Weighted Percentage	Estimate (Thousands)[1]	Weighted Percentage	Estimate (Thousands)[1]	Weighted Percentage	Estimate (Thousands)[1]	Weighted Percentage	Estimate (Thousands)[1]	Weighted Percentage
White	170,049	67.0	170,051	67.0	170,055	67.1	170,078	67.1	170,060	67.1	170,065	67.1
Black/ African American	30,233	11.9	30,235	11.9	30,209	11.9	30,210	11.9	30,210	11.9	30,210	11.9
American Indian/ Alaska Native	1,205	0.5	1,205	0.5	1,219	0.5	1,207	0.5	1,205	0.5	1,205	0.5
Native Hawaiian/ Other Pacific Islander	731	0.3	731	0.3	731	0.3	731	0.3	759	0.3	750	0.3
Asian	11,454	4.5	11,454	4.5	11,448	4.5	11,448	4.5	11,420	4.5	11,429	4.5
Two or More Races	3,178	1.3	3,179	1.3	3,178	1.3	3,178	1.3	3,188	1.3	3,182	1.3
Hispanic/ Latino	36,769	14.5	36,764	14.5	36,780	14.5	36,767	14.5	36,778	14.5	36,778	14.5
Total	253,619	100.0	253,619	100.0	253,619	100.0	253,619	100.0	253,619	100.0	253,619	100.0

PMN = predictive mean neighborhood.

[1] Estimates have been rounded to the nearest thousand to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

7.5.3.3 Imputation-Revised Hispanic/Latino Group (IRHOGRP4)

Imputed Hispanic/Latino group results are summarized below. Table 7.31 shows the percentages of seven Hispanic/Latino groups, where Mexican constitutes the highest percentage at 60.0 percent and Puerto Rican is the second largest group at 11.5 percent. Looking at the imputed cases only, Table 7.32 shows that the unweighted distribution of IRHOGRP4 from Method 5 mirrored the unweighted distribution of completed Hispanic/Latino group respondents more than the current PMN method. The Mexican group consists of 64.8 percent of all imputed Hispanic/Latino group cases for Method 5 and 70.7 percent for the current PMN method. The result for Method 5 is closer to the 60.0 percent distribution among all responding Hispanic/Latino group respondents from the interview data (Table 7.31). This could imply that the repeated imputation makes the imputed results closer to the completed case distributions. However, when comparing all cases in Tables 7.33 and 7.34, both the unweighted and weighted distributions show very little noticeable effect by the imputation method change.

Table 7.31 IRHOGRP4, Completed Cases Only (Unweighted)

Level	Frequency[1]	Unweighted Percentage
Puerto Rican	1,300	11.5
Mexican	6,500	60.0
Cuban	400	4.0
Other	< 50	0.2
Central/South American	1,700	15.8
Dominican	400	3.9
Spanish (from Spain)	500	4.5
Total	10,900	100.0

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

Table 7.32 IRHOGRP4, Imputed Cases Only (Unweighted)

Level	Current PMN		Method 5	
	Frequency	Unweighted Percentage	Frequency	Unweighted Percentage
Puerto Rican	6	10.3	6	11.1
Mexican	41	70.7	35	64.8
Cuban	0	0.0	1	1.9
Other	0	0.0	0	0.0
Central/South American	5	8.6	5	9.3
Dominican	3	5.2	4	7.4
Spanish (from Spain)	3	5.2	3	5.6
Total*	58	100.0	54	100.0

PMN = predictive mean neighborhood.

*Total counts are different because some cases were imputed with different IRHOIND values.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

Table 7.33 IRHOGRP4, All Cases (Unweighted)

Level	Current PMN		Method 5	
	Frequency[1]	Unweighted Percentage	Frequency[1]	Unweighted Percentage
Puerto Rican	1,300	1.8	1,300	1.8
Mexican	6,600	9.6	6,600	9.6
Cuban	400	0.6	400	0.6
Other	< 50	0.0	< 50	0.0
Central/South American	1,700	2.5	1,700	2.5
Dominican	400	0.6	400	0.6
Spanish (from Spain)	500	0.7	500	0.7
Non-Hispanic/Latino	57,600	84.0	57,600	84.0
Total	68,500	100.0	68,500	100.0

PMN = predictive mean neighborhood.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

Table 7.34 IRHOGRP4, All Cases (Weighted)

Level	Current PMN		Method 5	
	Estimate (Thousands)[1]	Weighted Percentage	Estimate (Thousands)[1]	Weighted Percentage
Puerto Rican	3,899	1.5	3,895	1.5
Mexican	23,363	9.2	23,323	9.2
Cuban	1,631	0.6	1,642	0.6
Other	57	0.0	57	0.0
Central/South American	5,511	2.2	5,550	2.2
Dominican	762	0.3	763	0.3
Spanish (from Spain)	1,547	0.6	1,548	0.6
Non-Hispanic/Latino	216,850	85.5	216,842	85.5
Total	253,619	100.0	253,619	100.0

PMN = predictive mean neighborhood.

[1] Estimates have been rounded to the nearest thousand to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010.

7.6 Age Imputation

This section discusses the consistency between the age reported by the screener respondent as compared with the age reported by the interview respondent. As discussed in Section 7.2.2, screener age was already used in the current data processing to create the final AGE interview variable. Interview respondents had multiple opportunities to change their age in response to consistency checks throughout the interview questionnaire. Therefore, it was possible for the age reported by a respondent at the beginning of the interview (CALCAGE) to be different from the age reported at the end of the interview (NEWAGE). The final age variable (AGE) was determined using these two variables and three other sources: the age calculated from the final edited interview date and the raw birth date, the age corresponding to the "self" in the

interview household roster (if it existed), and the pre-interview screener age.[43] In the 2009 NSDUH, nearly 100.0 percent of the respondents had consistent CALCAGE and NEWAGE, and only fewer than 10 cases used the screener age or household roster age to determine the value for the final AGE variable.[44]

Table 7.35 summarizes the percentage of all cases from the 2009 NSDUH where the age reported by the interview respondent did not match the age reported by the screener respondent. The self-reported age at the end of the interview (NEWAGE) was compared with the age reported by the screener respondent.

Table 7.35 Comparison of Screener and Interview Age, 2009 NSDUH

	Total		Dual Respondents		Non-Dual Respondents	
Interview Respondent Age	Frequency[1]	Unweighted Percentage[2]	Frequency[1]	Unweighted Percentage[2]	Frequency[1]	Unweighted Percentage[2]
Consistent with Screener Age	61,700	89.8	20,200	92.5	41,500	88.5
Not Consistent with Screener Age (±1 Year)	5,800	8.5	1,400	6.5	4,400	9.4
Not Consistent with Screener Age (±2 Years)	600	0.9	100	0.5	500	1.1
Not Consistent with Screener Age (±3 Years)	200	0.3	< 50	0.1	200	0.4
Not Consistent with Screener Age (±4 Years)	100	0.2	< 50	0.1	100	0.2
Not Consistent with Screener Age (±5 Years)	100	0.1	< 50	0.1	< 50	0.1
Not Consistent with Screener Age (> ±5 Years)	200	0.2	< 50	0.2	100	0.3
Screener Age is Categorical[3]	100	0.1	< 50	0.1	100	0.1
Total	68,700	100.0	21,800	100.0	46,900	100.0

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Percentages have been rounded to the nearest tenth to ensure respondent confidentiality.

[3] The screener age is 199, 299, 399, 499, or 599, indicating the age categories of 12 to 17, 18 to 25, 26 to 34, 35 to 49, or 50 or older.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

In 2009, 89.8 percent of total complete interview cases had screener age consistent with the age reported at the end of interview, and 8.5 percent of cases had screener age not consistent with interview age by 1 year. Among the dual respondents (i.e., where the same person responded to the screening and interview), 92.5 percent of cases had consistent screener and interview age, and 6.5 percent of cases had screener age not consistent with interview age by 1 year. Compared with the dual respondents, the consistency rate among the non-dual respondents decreased to 88.5 percent, and the rate of screener age not consistent with interview age by 1 year increased to 9.4 percent. Because the screening process is completed before the interview date, this category covered cases where the respondent had a birthday after the screening process was completed but before the interview was completed. However, screener age and interview

[43] Refer to the 2009 MRB imputation report for details on creation of the AGE interview variable (Ault et al., 2011).

[44] Refer to Section 4.2.2.1, Table 4.2, of the 2009 MRB imputation report for details on the AGE editing summary (Ault et al., 2011).

age differ for quite a few cases by 2 or more years (about 1.8 percent, or approximately 1,200 total cases).

7.7 Summary and Options

Overall, switching the imputation order of Hispanic/Latino origin and race does not make a meaningful difference for race as demonstrated by the outcomes for Method 1 in Tables 7.21, 7.25, and 7.28. Including screener race levels as covariates does influence imputed race results shown for Methods 2 to 5 in Tables 7.21, 7.25, 7.28, and 7.32. However, the change is to shift the imputed distribution away from interview data to be more in tune with screener data. Moreover, because of multicollinearity that existed between screener race and interview race, the only screener race level that usually remained in the model was white. This skews the distribution of imputed cases toward more whites. In the model-based approach, imputed values are more sensitive to the covariates used in the model compared with PMN, as shown by larger differences between the completed cases distribution and the imputed cases distribution produced by Methods 3 to 5. Cycling makes the imputed cases distribution more like the completed cases distribution.

With the exception of the Hispanic/Latino group variable, values imputed by the current PMN method and Method 1 are most like the nonmissing values. Including screener covariates in the model shifted the imputed distribution to be more like the screener data, but the addition of screener covariates in PMN changed the distribution less, compared with the addition of these covariates in the model-based approach, as shown by comparing Method 2 with Methods 3, 4, and 5. Method 3 had the most distribution changes, and the addition of cycling in Methods 4 and 5 shifted the distribution back toward the interview data somewhat.

To keep the imputed results more aligned with the interview data, it appears to be more optimal not to use the screener data in the model. PMN has proven to be more robust than the model-based approach from testing. The order of imputing Hispanic/Latino origin and race using PMN does not make too much difference in the final imputed Hispanic/Latino origin and race variables. Adding cycling steps would improve the likeliness of the imputed results matching the interview data, but the additional gain may not offset the extra effort required to carry out this operation.

Because age can be determined using the self-reported age at the beginning (CALCAGE) or at the end (NEWAGE) of the questionnaire for the vast majority of interview respondents, screener age data can be useful in rare occasions when CALCAGE and NEWAGE are not consistent. This practice is already in place during the regular AGE editing processing.

8. Imputation Using the Responses from the Other Pair Member

In each household selected for the National Survey on Drug Use and Health (NSDUH), zero, one, or two household members are selected for interviewing. When two members of the same household are selected and both complete an interview, a "responding pair" is formed. In the 2009 NSDUH, 58.8 percent of the unit respondents[45] were members of a responding pair. The pair relationship can be parent-child, sibling-sibling, spouse-spouse, or some other relationship.[46] The Substance Abuse and Mental Health Services Administration (SAMHSA) was interested in exploring whether certain variables for which imputation is performed have high positive correlation between pair respondent members and therefore could be used in the imputation process. Currently, the information about the other pair member is used only in editing of variables related to the household roster but not in the imputation process. It is possible that assigning the value of the other pair member would be a better imputation method than the current method, predictive mean neighborhood (PMN), or this approach could be incorporated into the current PMN procedures. For example, in an attempt to find a donor using PMN, the set of item respondents could be checked for the other pair member. If the other pair member is a respondent, the value of the other pair member would be assigned. If this attempt does not work, PMN would proceed as normal through the current sets of likeness constraints. Another possibility is to use the estimated probability of agreement for respondent pairs to decide whether to assign the value of the other pair member or to randomly assign a value through PMN.

The goal of this exercise is to assess using information from one pair member to assist with imputation of missing data for the other pair member. This includes choosing candidate variables for which this method would be appropriate and determining whether the benefits of using this method outweigh the costs of development and implementation.

8.1 Choosing Candidate Variables

Certain variables are more suited for this method than others. Some of the questions in the NSDUH ask for household-level information, such as the household roster. The responses the pair members give to these questions should almost always agree. Other than measurement error, the only reason for disagreement would be a change in the household composition between the times when the questions are answered. Some NSDUH questions, such as the majority of the ones in the income section, ask for information about the family in the household. If the pair members are members of the same family, then the responses are more likely to agree than if the pair members are not members of the same family.[47]

---

[45] A case is defined as a unit respondent if data were provided on lifetime use of cigarettes and at least nine other substances.

[46] See Table L.7 for a listing of all pair relationships where the pair members are members of the same family.

[47] Perhaps these questions about the family in the household, or the household as a whole, should only be asked once if a pair is selected. This issue is beyond the scope of this report.

All variables were initially considered as potential candidates for this method. This allowed a contrast between good and bad candidates. Moreover, once the methodology was set up for some variables, it was straightforward to implement for the other variables: the marginal cost of assessing the candidacy for all variables was low. In this initial step, tables for examining missingness were created. The tables in Appendix I list all variables that were imputed[48] in the 2009 NSDUH. The columns in Tables I.1 through I.6 are defined as follows:

- **Variable:** The description of the variable.
- **Number of Respondents in Domain:** The number of unit respondents who are in the domain. The domain for a variable is defined as the set of unit respondents who received a value other than a skip code for the imputation-revised variable of interest. In other words, a domain is the subset of respondents for whom the variable of interest is relevant or applicable.[49]
- **Number of Responding Pairs in Domain:** The number of times both members of the pair are in the domain and both members of the pair are item respondents for the variable.
- **Percentage of Pair Agreement:** Of the number of respondent pairs in the domain (in the preceding column), the percentage of respondents whose values for the variable are equal. This is expected to be very high for family-level variables such as those in the income and health insurance sections.
- **Number Missing in Domain:** The number of missing values for the variable among unit respondents in the domain.
- **Number of Nonrespondents Paired with Respondents:** Of the missing values in the domain (in the preceding column), the number of item nonrespondents who are eligible for the proposed method. In order to be eligible for the proposed method, the item nonrespondent must be (1) in the domain, (2) a member of a pair, and (3) paired with an item respondent for the variable who is also in the domain.
- **Percentage of Nonrespondents Eligible for Edit:** Of the missing values in the domain, the percentage of nonrespondents who are eligible for this proposed method.
- **Number of Nonrespondents Paired with Respondents that Agree after PMN:** Of the nonrespondents eligible for this proposed method, the number whose PMN-imputed values agree with the other pair member's actual response. If the pair members frequently disagree after imputation, but frequently agree when both respond, the proposed method may perform better than PMN.
- **Percentage of Eligible Nonrespondents with Pair Agreement after PMN:** Of the nonrespondents eligible for this proposed method, the percentage whose PMN-imputed values agree with the other pair member's actual response.

The interpretation of these tables is best described with an example. In the 2009 NSDUH, there were 68,700 unit respondents. Because every unit respondent was asked how many people live in his or her household, the domain for household size includes all 68,700 respondents (see Table I.6). In contrast, respondents were asked about their marital status only if they were aged

[48] Appendix A of the 2011 imputation report of the NSDUH methodological resource book (MRB; Frechtel et al., 2013) presents similar tables for variables that were imputed in the 2011 NSDUH.

[49] The domain definitions for each variable are specified in Appendix A of the 2011 MRB imputation report (Frechtel et al., 2013).

15 years or older. Therefore, the domain for marital status includes only the 57,817 respondents aged 15 or older (see Table I.1).

In the 2009 NSDUH, 40,384 of the 68,700 unit respondents were pair members, creating a total of 20,192 pairs of respondents. Of these, there were 20,181 pairs in which both pair members gave valid responses to the question on household size. Within these 20,181 respondent pairs, there were 19,113 (94.71 percent) pairs in which both pair members reported the same number of household members. This left 1,068 (5.29 percent) respondent pairs where the members of the pair reported a different number of household members.

As shown in Table I.6, of the 68,700 survey respondents, 35 either did not answer the question about household size or gave answers that were coded as "bad data" during editing. Eleven of these item nonrespondents were members of a pair in which the other pair member gave a valid response to the household size question. Currently, data from the respondent pair member are only used in consistency checks in the editing process and in creating bounds for the imputation of household roster and pair variables. In fact, if the respondent pair member's household size had been used to impute household size for the other member of the pair, the result would have disagreed with the result obtained from PMN in 8 of the 11 cases.

In general, good candidate variables for this approach meet the following conditions:

- When both pair members are item respondents for the variable, their values almost always agree.
- There are enough missing values where the other pair member responded to justify the cost of developing and implementing the extra imputation steps.
- For item nonrespondents paired with item respondents, the values imputed by PMN often differ from the other pair member's response, and it is logical to assume that the responses would be the same (i.e., a better imputation method than PMN appears readily available for these item nonrespondents).

Table 8.1 below provides a summary of the tables in Appendix I by variable groups. The demographic, household roster, income, and health insurance groups have an average agreement of 75 percent or higher between pair responses. Within the health insurance and household roster groups, an average of almost 50 percent of nonrespondents is eligible for the proposed method. Using PMN results in imputing a value that disagrees with the pair member for approximately 40 percent of eligible nonrespondents in the demographic and income groups and for almost 50 percent in the household roster group.

Table 8.1 Summary of Agreement Rates Presented in Appendix I

Variable Group	Average Percentage of Pair Agreement	Average Percentage of Nonrespondents Eligible for Edit	Average Percentage of Pair Agreement after PMN
Demographics	75.7	26.8	61.5
Drugs	40.5	14.7	20.8
Health Insurance	87.2	46.5	68.3
Income	82.4	38.4	58.7
Pair	100.0	0.0	100.0
Roster	94.6	45.4	52.3

PMN = predictive mean neighborhood.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

A more detailed look at Table I.1 shows that the level of agreement is high (more than 90 percent) for most of the demographic variables. The exceptions are education level, employment status, marital status, and immigrant age of entry, variables that would not be expected to match among a majority of pairs. Similarly, the family-level income variables in Table I.4 have much higher percentages of pair agreement in comparison with the respondent-level income variables, which are not expected to match between pair members. Most of the demographic, household roster, income, and health insurance variables are likely to be good candidates for this method because of their high pair agreement rates and relatively high percentages of eligible nonrespondents or pair disagreement after PMN.

Though the overall pair agreement rate across all drugs is low, many of the lifetime drug use variables show a high level of matching between pair members. This is due to the very low prevalence of these drugs in general (i.e., both pair members usually respond that they do not use the drug) as opposed to a strong correlation between pair members. The pair group consists of variables that give information about the relationship between the pair of household members selected for the survey. Because they are defined at the pair level as opposed to the respondent level, by definition, pair variables always have 100 percent pair agreement. Therefore, these variables are not eligible for this method. They are included in the table for illustration only.

8.2 Taking a Closer Look at Good Candidate Variables

As expected, the demographic, household roster, income, and health insurance variables appear to be the best candidates. Of these groups, the income and health insurance variables were selected for closer examination for the following reasons:

- The income and health insurance variables have a significant amount of missing data, unlike most of the demographic variables and all of the household roster variables.
- All of the demographic variables are at the person level, not the household level or the family-in-household level, so it may be less appropriate to directly assign the value of the other pair member. It makes more sense to use the other pair member as an influence on the final imputed value, not as the sole determinant of the final imputed value. Perhaps there is a way to use the other pair member's value in the prediction model, for example. However, this approach would complicate matters rather than simplify them.

- All of the questions in the health insurance section ask about the respondent, not the family in the household. Still, because many health insurance plans provide coverage at the family level, the pair members are expected to agree more often when they are members of the same family. Because these questions ask about current coverage by health insurance, the number of days between the responses may be a factor: coverage status may change between the times the responses were given.
- All of the questions in the income section ask about the preceding calendar year (versus a rolling past 12-month reference period), so the number of days between the responses of the pair members theoretically should not be a factor. Most of the questions in the income section ask about the family in the household, so as long as the pair members are members of the same family, the only theoretical source of disagreement is measurement error.[50]

When taking a closer look at the income and health insurance variables, three factors were considered:

- **Family Pair Indicator:** As stated above, most of the income questions were asked at the family-in-household level, and although the health insurance questions were asked at the respondent level, family members might be expected to agree much of the time in their responses. There is an imputation-revised variable called IRPRREL that identifies the pair type. This variable was collapsed into a dichotomous variable for further analysis: either the pair members were clearly in the same family (parent-child, sibling-sibling, spouse-spouse, or grandparent-grandchild) or they were not. In the 2009 NSDUH, 85.6 percent of the responding pairs were clearly members of the same family according to IRPRREL.
- **Number of Days between Responses:** Date stamps are available for each section for each respondent, so the number of days between the responses can easily be calculated. As stated above, for the income questions, the number of days between responses theoretically is not important. However, it may increase the likelihood of measurement error. For the health insurance questions, the number of days between responses may be important because the likelihood that there is a change in current health insurance coverage presumably would increase as the number of days between responses increases. In 2009, the responses were entered on the same day 65.5 percent of the time, within 7 days 86.2 percent of the time, within 14 days 92.2 percent of the time, and within 30 days 96.7 percent of the time.
- **Whether the Same Person Answered Both Questions:** For many respondents, the income and health insurance questions were answered by a proxy. This proxy had to be a member of the respondent's family and had to be at least 18 years old. Sometimes, the proxy was the same person as the other pair member. In these cases, the pair members might be expected to agree practically all of the time. An assessment was done of (1) how frequently the responses agree when the proxy and

[50] Perhaps the differences between responses can be used as a simple assessment of measurement error for these items. The differences can be used directly for income, and for the household roster, the differences can be used if the responses were given on the same day (or within a reasonable number of days).

the other pair member are the same person, and (2) how frequently one response is missing and the other is not. This assessment led to the following conclusions:

- In many cases, it is not easy to determine whether the proxy and the other pair member are the same person. A careful review of the household roster for each pair member is required. An algorithm was designed to identify the most obvious cases. For example, if the pair type was parent-child, the father answered the questions for the child, the parent reported being male, and the parent answered the questions himself, then it was assumed that the father answered both questions.
- When the same person gives both responses, the responses agree practically all of the time, and it is very rare that one response is missing and the other is not.
- Because of the difficulty of determining whether the same person gave both responses, and because of the limited number of cases where one response is missing and the other is not, this factor was dropped from further consideration.

In order to determine whether pair type and number of days between responses are related to the extent of agreement between the variables of interest (income and insurance variables), a series of logistic regression models were run where the dependent variable was an indicator of agreement between pair members, and the two independent variables were the first two factors mentioned above: the pair type and the number of days between responses. Whether the two factors were significantly related to the agreement within the variables of interest was examined, as well as the predicted probability of extent of agreement.

If the predicted probability of agreement is close to 1, then the proposed imputation method (which forces agreement between pair members) is a reasonable choice. If the predicted probability of agreement is not close to 1, then perhaps it makes more sense to use PMN with its stochastic component. The two sections below describe the results of logistic regression models involving the income and health insurance variables.

8.2.1 Income

There are nine family-level edited income variables. Separate logistic regression models were fit for each of the nine variables. Agreement was defined as having the exact same value, even for the continuous variables. As expected, for all nine models, the family pair indicator was a statistically significant covariate ($\alpha = 0.05$) and the predicted probabilities of agreement were higher when the pair members were in the same family. Also as expected, the predicted probability of agreement decreased with the increase in the number of days between the two pair interviews for all nine variables. For seven of the nine variables, the regression coefficient associated with the number of intervening days was significantly different from zero ($\alpha = 0.05$).[51] Table 8.2 lists the income variables used in the models, the percentages of family pairs and other pairs whose responses agree, and whether the regression coefficients were statistically significant.

[51] To check whether the results were affected by a few pairs with a large number of intervening days that disagreed, the models were refit using an ordinal version of this dependent variable (0-7, 8-30, 31 or more). Results were similar.

Table 8.2 Summary of Logistic Regression Results, Income Variables

Variable	Number of Pairs Used in Analysis	Actual Percentage of Pairs that Agree		Statistical Significance of Covariates	
		Family	Other	Family Pair Indicator	Number of Days Between Responses
Social Security (yes/no)	19,855	96.05	93.08	Yes	Yes
Supplemental Security Income (yes/no)	19,718	96.86	94.81	Yes	Yes
Welfare Payments (yes/no)	19,924	97.96	96.44	Yes	No
Welfare Services (yes/no)	19,986	96.81	95.41	Yes	Yes
Wages (yes/no)	20,101	94.96	85.05	Yes	No
Food Stamps (yes/no)	20,056	96.50	90.82	Yes	Yes
Welfare Months (continuous)	854	80.13	66.67	Yes	Yes
Total Family Income (dichotomous)	19,049	95.05	78.17	Yes	Yes
Total Family Income (finer categories)	17,355	68.12	18.67	Yes	Yes

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

The proportion of pairs that agree is lower for welfare months and for total family income (finer categories). This is mostly because the other variables are dichotomous, whereas these two are ordinal with several levels. The indicator of agreement does not account for the magnitude of the disagreement. For example, the months-on-welfare variable has values from 1 to 12: if one pair member reports 10 months on welfare in the prior year and the other reports 9, the pair is still in disagreement, just as if one pair member reports 1 month on welfare and the other reports 12 months on welfare. The finer-categories-of-income variable has 29 levels.

Although the regression coefficient on the number of days between responses was statistically significant for most income variables and negative for all income variables, statistical significance is easy to achieve when the sample size is so large. For all income variables except welfare months, the sample size includes several thousand pairs. It is possible that the results are statistically significant but not important in a practical sense: the regression coefficient may still be close to zero, causing the predicted probability of agreement to decline slowly as the number of intervening days increases. To assess this for the family pairs, predicted probabilities of agreement were calculated for fixed values of the number of intervening days. These results are displayed in Table 8.3.

Table 8.3 Predicted Probabilities of Agreement for Family Pairs, as a Function of the Number of Intervening Days, Income Variables

Variable	Predicted Probability of Agreement for Various Values of the Number of Intervening Days (%)						
	0	5	10	20	30	50	70
Social Security	96.23	96.00	95.75	95.20	94.60	93.16	91.37
Supplemental Security Income	96.99	96.82	96.63	96.22	95.77	94.70	93.39
Welfare Payments	98.03	97.94	97.85	97.67	97.46	97.01	96.47
Welfare Services	96.86	96.79	96.71	96.56	96.40	96.05	95.68
Wages	94.99	94.95	94.90	94.81	94.71	94.52	94.32
Food Stamps	96.75	96.45	96.11	95.36	94.46	92.16	89.02

Table 8.3 Predicted Probabilities of Agreement for Family Pairs, as a Function of the Number of Intervening Days, Income Variables (continued)

Variable	Predicted Probability of Agreement for Various Values of the Number of Intervening Days (%)						
	0	5	10	20	30	50	70
Welfare Months	81.44	78.40	75.01	67.26	58.43	39.69	23.56
Total Family Income (dichotomous)	95.48	94.97	94.40	93.08	91.48	87.25	81.34
Total Family Income (finer categories)	70.95	66.62	62.00	52.15	42.13	24.52	12.66

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

As shown in Table 8.3, for most variables, the predicted probability of agreement decreases slowly with the number of intervening days. In 2009, the number of intervening days was less than or equal to 30 for about 97 percent of the pairs. When fixing the number of intervening days at 30, the predicted probability of agreement is still greater than 90 percent for all variables except welfare months and finer categories income.

Given that (1) the income questions ask about the prior year so that the response is not theoretically dependent on the exact date in the current year when the response was given, (2) the income variables discussed in this section store data at the level of the family in the household, and (3) for most of the variables discussed in this section, the predicted probability of agreement decreases slowly as the number of intervening days increases, it appears that the use of the other pair member's value in imputation for these nine income variables would be an improvement over the current PMN method as long as the pair members are in the same family. It does not appear to be necessary to consider the number of days between responses. As shown in Table 8.4, following this recommendation would have reduced the amount of PMN imputation required by up to 40 percent for these variables in 2009.

Table 8.4 Proportion of Item Nonrespondents that Could Be Imputed Using the Other Pair Member Method, Income Variables, 2009 NSDUH

Variable	Number of Item Nonrespondents				Percentage Handled Using Other Pair Member
	Total	In Pairs	(and) Paired with an Item Respondent	(and) In Family Pair	
Social Security	660	364	310	255	38.64
Supplemental Security Income	935	524	424	352	37.65
Welfare Payments	492	293	243	190	38.62
Welfare Services	371	222	190	150	40.43
Wages	193	101	81	65	33.68
Food Stamps	262	147	125	97	37.02
Welfare Months	218	71	45	37	16.97
Total Family Income (dichotomous)	2,557	1,430	856	689	26.95
Total Family Income (finer categories)	6,624	3,836	1,828	1,511	22.81

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

8.2.2 Health Insurance

There are eight edited health insurance variables that undergo imputation. All eight of the edited health insurance variables are dichotomous (yes/no) variables. Logistic regression models were fit for these eight variables using the same methodology that was used for the income variables. Table 8.5 summarizes the results of the models. The family pair indicator was a significant covariate for seven of the eight models, and the number of intervening days was a significant covariate for five of the eight models. The predicted probabilities of agreement were higher when the pair members were in the same family for eight of the nine variables. As for income, the predicted probability of agreement decreased with the number of intervening days for all variables.

Table 8.5 Summary of Logistic Regression Results, Health Insurance Variables

Variable	Number of Pairs Used in Analysis	Percentage of Pairs that Agree		Statistical Significance of Covariates	
		Family	Other	Family Pair Indicator	Number of Intervening Days
Overall Health Insurance, 1999 Method	19,894	83.08	69.18	Yes	Yes
Overall Health Insurance, 2001 Method	19,882	84.85	72.75	Yes	Yes
Private Health Insurance, Consistent with Pre-1999 Surveys	19,932	84.80	70.24	Yes	Yes
Medicaid/CHIP	19,899	86.92	84.15	Yes	No
Medicare	20,052	96.41	96.83	No	No
Military Health Care (CHAMPUS, TRICARE, CHAMPVA, VA)	20,070	97.88	95.79	Yes	No
Private Health Insurance, as Defined by Constituent Variables Method	19,932	84.80	70.24	Yes	Yes
Other Health Insurance	2,069	92.38	87.18	Yes	Yes

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

Table 8.6 shows the predicted probabilities as a function of the number of intervening days for family pairs only. For all variables, the predicted probability of agreement decreases slowly as the number of intervening days increases. This suggests that coverage status does not tend to change very often: assuming no measurement error, it tends to stay constant over the short term.

Table 8.6 Predicted Probabilities of Agreement for Family Pairs, as a Function of the Number of Intervening Days, Health Insurance Variables

Variable	Predicted Probability of Agreement for Various Values of the Number of Intervening Days (%)						
	0	5	10	20	30	50	70
Overall Health Insurance, 1999 Method	83.43	82.93	82.41	81.34	80.21	77.82	75.22
Overall Health Insurance, 2001 Method	85.20	84.71	84.21	83.16	82.06	79.70	77.12

Table 8.6 Predicted Probabilities of Agreement for Family Pairs, as a Function of the Number of Intervening Days, Health Insurance Variables (continued)

Variable	Predicted Probability of Agreement for Various Values of the Number of Intervening Days (%)						
	0	5	10	20	30	50	70
Private Health Insurance, Consistent with Pre-1999 Surveys	85.21	84.64	84.05	82.82	81.51	78.66	75.51
Medicaid/CHIP	87.06	86.85	86.64	86.20	85.75	84.82	83.84
Medicare	96.46	96.39	96.32	96.18	96.04	95.73	95.39
Military Health Care (CHAMPUS, TRICARE, CHAMPVA, VA)	97.88	97.87	97.87	97.85	97.83	97.80	97.76
Private Health Insurance, as Defined by Constituent Variables Method	85.21	84.64	84.05	82.82	81.51	78.66	75.51
Other Health Insurance	92.78	92.32	91.82	90.75	89.55	86.74	83.31

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

As stated above, the health insurance questions in the NSDUH differ from the income questions in two ways: (1) they are asked at the respondent level, not the family level; and (2) they ask about current coverage, not coverage over some fixed time interval. Because of this, the use of the other pair member's value in imputation is harder to justify theoretically, even when the responses were given at the same time. Even when the responses were given on the same day, there is a nontrivial proportion of disagreeing responses. For family pairs, the predicted probability of agreement is below 90 percent for five of the eight variables, even when the number of intervening days is zero. Just to verify that the model's predictions were reasonable when the number of intervening days is zero, the actual proportions of agreement when there were no intervening days was compared with the predicted probabilities. The two measures were similar, as shown in Table 8.7.

Table 8.7 Comparison of Proportion of Agreement to Predicted Probability of Agreement for Family Pairs, No Intervening Days, Health Insurance Variables

Variable	Respondent Pairs with No Intervening Days			Predicted Probability of Agreement, No Intervening Days
	Total	Number of Pairs that Agree	Percentage of Pairs that Agree	
Overall Health Insurance, 1999 Method	15,082	12,618	83.66	83.43
Overall Health Insurance, 2001 Method	15,076	12,868	85.35	85.20
Private Health Insurance, Consistent with Pre-1999 Surveys	15,115	12,900	85.35	85.21
Medicaid/CHIP	15,082	13,143	87.14	87.06
Medicare	15,191	14,660	96.50	96.46
Military Health Care (CHAMPUS, TRICARE, CHAMPVA, VA)	15,202	14,884	97.91	97.88
Private Health Insurance, as Defined by Constituent Variables Method	15,115	12,900	85.35	85.21
Other Health Insurance	1,436	1,330	92.62	92.78

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

Because there are two clear reasons why pair members can disagree, even if they are members of the same family, a deterministic imputation method is not recommended for the health insurance variables. Perhaps a more detailed investigation of the frequency of agreement between family members for certain types of health insurance would lead to the conclusion that the other-pair-member approach could be justified under specific conditions.

8.3 Summary and Options

The purpose of this chapter was to assess the feasibility of an alternative approach to imputation: a simple assignment of the value of the other pair member. In certain situations, this approach seems preferable to PMN because it is both simpler and more accurate.

A simple assessment of feasibility applied to all variables that undergo imputation suggested that the income and health insurance variables were the best candidates for this method. There are nontrivial numbers of missing values for most of these variables ranging from 183 to 6,359 cases, and there are nontrivial numbers of cases to which the other-pair-member method would apply ranging from 31 to 1,736 cases: item nonrespondents that were paired with item respondents. For these cases, the proportion of pairs whose responses agreed after PMN imputation was usually considerably lower than the proportion of responding pairs whose responses agreed, suggesting that the other-pair-member approach was a better imputation method. Most of the income questions ask about the family in the household, and when both pair members respond, they very often agree. Although the health insurance questions ask about the respondent, not the respondent's family in the household, pair members often agree when they are members of the same family.

For the income and health insurance variables, a more refined analysis was done to identify exact conditions under which the use of the other pair member was preferable to PMN. Two factors were considered: (1) the type of pair (definitely in the same family, or not), and (2) the number of days between the presentation of the questions to the pair members. After consideration of these factors using logistic regression models, the following conclusions may be drawn:

- For the nine family-level income variables, the other pair member's value could be used in imputation as long as the pair members are in the same family. This would reduce the amount of PMN imputation required for these variables by more than 30 percent.
- For the eight health insurance variables that undergo imputation, do not use the other pair member's value in imputation, because the deterministic nature of this method is inappropriate in the presence of obvious reasons for disagreement between pair members.

Some reasonable next steps include the following:

- **Take a closer look at the household roster variables.** The household roster variables that undergo imputation store data at the household and family-in-household level. That alone makes them good candidates for the other-pair-member approach. Like the health insurance questions, the household roster questions ask about the

current situation instead of the situation at some fixed period like the income questions; thus, the responses are somewhat dependent on the date the questions were administered. Perhaps an approach similar to that suggested for the health insurance variables would be reasonable. For the questions about the household that are not dependent on familial relationships, perhaps the pair type would not be an important factor. The main reason the household roster variables were not considered in this chapter was that the level of missingness is low.

- **For the health insurance variables, search for specific situations where using the other pair member's response in imputation is appropriate.** Some of the types of health insurance items included in the NSDUH may be at the family level, even though the questions are asked only of the respondent. It might be useful to consult with a subject matter expert and to complete more refined data analyses to determine when the other-pair-member approach is best.
- **Consider using the other pair member's response in the prediction models.** For many variables, including the health insurance and demographic variables, the other pair member's response may be useful as supporting information but not as the sole determinant of the imputed value. Methods for integrating this information into the prediction models could be explored.

9. Imputation Methods for Mental Health Variables

This chapter examines the extent and nature of item nonresponse for the National Survey on Drug Use and Health (NSDUH) mental health variables used in the estimation of serious mental illness (SMI) and any mental illness (AMI) to determine whether the current imputation method is performing adequately or if alternative methods would substantially improve the quality of the estimates.

Section 9.1 briefly describes the SMI model (from which both SMI and AMI estimates are derived). Section 9.2 describes the predictor variables of the SMI model in greater detail and presents an examination of how missing values were dealt with in each case. Section 9.3 describes the construction of new versions of SMI predictor variables in which cases where missing values that had previously been imputed as zero were converted to cases where the missing values were explicitly recorded as missing. Section 9.4 presents a summary of item nonresponse rates of these new versions of the SMI predictor variables (i.e., versions in which missing values were explicitly recorded as missing). Section 9.5 discusses item nonresponse rates of variables related to past month scores from the Kessler-6 (K6) scale obtained from other surveys. Section 9.6 shows the results of a sensitivity analysis that compares the effects of different imputation methods on estimates based on the mental health variables in question; that is, the imputation methods included (1) the current method of imputing all missing values as zero (resulting in a lower bound estimate), (2) imputing all missing values as the maximum value possible (resulting in an upper bound estimate), and (3) dropping all cases with missing values. Section 9.7 summarizes the results of performing imputation using weighted sequential hot-deck imputation, and Section 9.8 provides a summary as well as recommendations based on these analyses.

9.1 SMI Model

The model developed to estimate SMI was fit with data from 4,912 clinical interview adult respondents (i.e., 18 years or older) recruited from the NSDUH main interview from 2008 through 2012. Clinical interview respondents were diagnosed as having or not having SMI.

For modeling purposes, the response variable Y equaled 1 when an SMI diagnosis was positive based on the clinical interview; otherwise, Y was 0. Letting $\mathbf{X}$ be a vector of characteristics attached to a NSDUH respondent and letting the probability that this respondent had SMI be $\pi = \Pr(Y = 1 \mid \mathbf{X})$, the model was

$$\begin{aligned}\log\left[\text{SMIPP\_U} / \left(1 - \text{SMIPP\_U}\right)\right] &= -5.972664 + 0.0873416 * \text{WSPDSC2} \\ &+ 0.3385193 * \text{WHODASC3} + 1.9552664 * \text{MHSUTK\_U} \\ &+ 1.1267330 * \text{AMDEY2\_U} + 0.1059137 * \text{AGE1830},\end{aligned}$$

where SMIPP_U is the predicted probability an adult had SMI (i.e., SMIPP_U is an estimate of π). The five covariates in the model (WSPDSC2, WHODASC3, MHSUTK_U, AMDEY2_U, and AGE1830) come directly from the main NSDUH interview data. Detailed descriptions of these terms are provided in Section 9.2.

Respondents for whom SMIPP_U was greater than or equal to the SMI cut point (0.260573529) were predicted to SMI positive; otherwise, they were predicted to be SMI negative. And respondents for whom SMIPP_U was greater than or equal to the AMI cut point (0.0192519810) were predicted to AMI positive; otherwise, they were predicted to be AMI negative. For further details about the SMI model, refer to the 2012 Mental Health Surveillance Study design and estimation report of the NSDUH methodological resource book (Center for Behavioral Health Statistics and Quality, 2014).

9.2 Predictor Variables of the SMI Model

This section describes the analysis used to determine how missing values were treated in the variable creation chain of the five predictor variables in the SMI model (WSPDSC2, WHODASC3, MHSUTK_U, AMDEY2_U, and AGE1830), as a first step toward developing an understanding of the impact this might have had on resulting estimates.

9.2.1 WSPDSC2

The K6 screening instrument for nonspecific psychological distress (Furukawa, Kessler, Slade, & Andrews, 2003; Kessler et al., 2003) forms the basis of the SMI predictor variable WSPDSC2.

This instrument consists of two 6-item K6 scales that gather information regarding how frequently a respondent experienced symptoms of psychological distress during the past 30 days and during a month in the past 12 months when he or she felt more depressed, anxious, or emotionally stressed than in the past 30 days, respectively. Only respondents who indicated that there was a worse month than the past 30 days (DSTWORST = 1) were asked about the worst month in the past year other than the past 30 days.

The questions comprising the two K6 scales and the screener question for the worst month scale are provided below with their associated edited variable names from the mental health section as well as the response categories for each question:

DSTNRV30 During the past 30 days, how often did you feel nervous?

1 All of the time
2 Most of the time
3 Some of the time
4 A little of the time
5 None of the time
DK/REF

Response categories are the same for the five remaining past month K6 questions:

DSTHOP30 During the past 30 days, how often did you feel hopeless?

DSTRST30 During the past 30 days, how often did you feel restless or fidgety?

DSTCHR30 During the past 30 days, how often did you feel so sad or depressed that nothing could cheer you up?

DSTEFF30 During the past 30 days, how often did you feel that everything was an effort?

DSTNGD30 During the past 30 days, how often did you feel down on yourself, no good, or worthless?

DSTWORST The last questions asked about how you have been feeling during the past 30 days. Now think about **the past 12 months.** Was there a month in the past 12 months when you felt more depressed, anxious, or emotionally stressed than you felt during the past 30 days?

1 Yes
2 No

Response categories for the following K6 questions are identical to those for the corresponding past month K6 questions:

DSTNRV12 Think of one month in the past 12 months when you were the most depressed, anxious, or emotionally stressed. During that same month when you were at your worst emotionally . . .
how often did you feel nervous?

DSTHOP12 During that same month when you were at your worst emotionally . . .
how often did you feel hopeless?

DSTRST12 During that same month when you were at your worst emotionally . . .
how often did you feel restless or fidgety?

DSTCHR12 During that same month when you were at your worst emotionally . . .
how often did you feel so sad or depressed that nothing could cheer you up?

DSTEFF12 During that same month when you were at your worst emotionally . . .
how often did you feel that everything was an effort?

DSTNGD12 During that same month when you were at your worst emotionally . . .
how often did you feel down on yourself, no good, or worthless?

Each K6 scale item shown above was transformed so that "All of the time" was coded 4, "Most of the time" was coded 3, "Some of the time" was coded 2, "A little of the time" was

coded 1, and "None of the time" was coded 0; *all responses matching "Don't know," refusals, bad data, blanks, and legitimate skips were also coded 0*.

A past month distress score (K6SCMON) was calculated by summing these transformed values across the six past 30-day variables (DSTNRV30, DSTHOP30, DSTRST30, DSTCHR30, DSTEFF30, and DSTNGD30) to arrive at a value ranging between 0 and 24. Likewise, a worst month in the past year distress score (K6SCYR) was calculated by summing the transformed values across the six worst month in the past year variables (DSTNRV12, DSTHOP12, DSTRST12, DSTCHR12, DSTEFF12, and DSTNGD12) to arrive at a value ranging between 0 and 24. *The worst month in the past year distress score (K6SCYR) has nonmissing values only for adult respondents who indicated that there was a month in the past year that was worse than the past 30 days (DSTWORST = 1); for all other cases, K6SCYR was designated as missing*. A worst total score (K6SCMAX) was then created that takes on the maximum value of the past month distress score (K6SCMON) and the worst month in the past year distress score (K6SCYR) in order to represent the worst distress score during the past year, regardless of whether this contradicts the response to DSTWORST; in cases where K6SCYR was designated as missing (i.e., when DSTWORST ne 1), then K6SCMAX is simply equal to K6SCMON. *For all the K6 score variables, youths aged 12 to 17 were designated as missing because the mental health section was not administered to youths.*

An alternative worst month in the past year total score variable (WSPDSC2) was created to indicate the worst distress score during the past year. Using the worst month total score (K6SCMAX), the alternative worst month total score (WSPDSC2) is coded 0 when K6SCMAX has a value from 0 to 7, and WSPDSC2 is assigned a value of 1 to 17 when K6SCMAX has a corresponding value of 8 to 24.

For respondents aged 18 or older, WSPDSC2 has no missing values because either all missing values in the variables used to create it were imputed as zero or variables that were assigned missing values did not transfer those missing values further up the chain (e.g., K6SCYR).

9.2.2 WHODASC3

The World Health Organization Disability Assessment Schedule (WHODAS) is a scale used to measure functional impairment that consists of a series of items that are used for assessing disturbances in social adjustment and behavior (i.e., functional impairment). A reduced set of 13 WHODAS items (Novak, Colpe, Barker, & Gfroerer, 2010; Rehm et al., 1999) are included in the NSDUH.

The computer assisted interviewing (CAI) variable DISTRESS was created to determine which NSDUH respondents would be directed to or skipped out of the WHODAS questions. If a respondent recorded a positive past month or past year K6 item score, then DISTRESS = 1 and that respondent was directed to answer the WHODAS questions; *otherwise, DISTRESS = 2 and the respondent was skipped out of those questions and all WHODAS item scores were recorded as zero*.

Responses to the WHODAS impairment scale were used to create eight variables that were transformed and summed to define the WHODAS total score used in the development of the SMI prediction model. The questions comprising the abbreviated WHODAS are provided below, with their associated edited variable names from the mental health section as well as the response categories for each question:

> The next questions are about how much your emotions, nerves, or mental health caused you to have **difficulties in daily activities**.
>
> In answering, think of the **one month** in the past 12 months when your emotions, nerves, or mental health interfered **most** with your daily activities.
>
> **IMPREMEM** During that one month when your emotions, nerves or mental health interfered **most** with your daily activities . . .
> how much difficulty did you have **remembering to do things you needed to do**?
>
> 1 No difficulty
> 2 Mild difficulty
> 3 Moderate difficulty
> 4 Severe difficulty
> DK/REF
>
> *Response categories for IMPCONCN are identical to those for IMPREMEM:*
>
> **IMPCONCN** how much difficulty did you have concentrating on doing something important when other things were going on around you?
>
> *The first four response categories for IMPGOUT, IMPPEOP, IMPSOC, IMPHHLD, and IMPRESP are identical to those for IMPREMEM; but these five items also have an additional fifth "not applicable" category, such as "you didn't go to work or school":*
>
> **IMPGOUT** how much difficulty did you have **going out of the house and getting around on your own?**
>
> **IMPPEOP** how much difficulty did you have **dealing with people you did not know well?**
>
> **IMPSOC** how much difficulty did you have **participating in social activities, like visiting friends or going to parties?**
>
> **IMPHHLD** how much difficulty did you have **taking care of household responsibilities?**
>
> **IMPRESP** how much difficulty did you have **taking care of your daily responsibilities at work or school?**

The next five questions were triggered if the "not applicable" category was selected in any of the related questions above:

IMPGOUTM [IF IMPGOUT = 5] Did problems with your emotions, nerves, or mental health keep you from leaving the house on your own?

1 Yes
2 No
DK/REF

Response categories for IMPPEOPM, IMPSOCM, IMPHHLDM, and IMPRESPM are identical to those for IMPGOUTM:

IMPPEOPM [IF IMPPEOP = 5] Did problems with your emotions, nerves, or mental health keep you from dealing with people you did not know well?

IMPSOCM [IF IMPSOC = 5] Did problems with your emotions, nerves, or mental health keep you from participating in social activities?

IMPHHLDM [IF IMPHHLD = 5] Did problems with your emotions, nerves, or mental health keep you from taking care of household responsibilities?

IMPRESPM [IF IMPRESP = 5] Did problems with your emotions, nerves, or mental health keep you from working or going to school?

Response categories for IMPWORK are identical to those for IMPREMEM, but this question was skipped if IMPRESP = 5 (and this item was then scored according to the response to IMPRESPM):

IMPWORK [IF **IMPRESP** NE 5] During that one month when your emotions, nerves or mental health interfered **most** with your daily activities . . .
how much difficulty did you have **getting your daily work done as quickly as needed**?

An original WHODAS total score (WHODASC2) was created to indicate the level of difficulty in performing daily activities due to problems with emotions, nerves, or mental health. Each of the eight variables created from the WHODAS items shown above was transformed into values of 0 to 3 so that a response of "severe difficulty" was coded 3, "moderate difficulty" was coded 2, "mild difficulty" was coded 1, and "no difficulty" was coded 0; *all responses matching "Don't know," refusals, bad data, blanks, and legitimate skips were also coded 0*.

Some items had a fifth category to deal with "not applicable" responses. For example, the question about difficulties regarding taking care of daily responsibilities at work or school (IMPRESP) had a fifth category, "you didn't work or go to school." If this category was selected, then another question was asked as to whether respondents' emotions, nerves, or mental health caused them to be unable to work or go to school (IMPRESPM). A "yes" response to the follow-up question (IMPRESPM = 1) was coded 3 and a "no" response (IMPRESPM = 2) was coded 0; *all responses to IMPRESPM matching "Don't know," refusals, bad data, blanks, and legitimate skips were also coded 0*. One exception to this coding was the last WHODAS recode on how

much difficulty the respondents had in getting their daily work done as quickly as needed (IMPWORK). This item was asked of the respondents only if in the previous question they responded that they worked or went to school (IMPRESP = 1 to 4). In the case that they responded that they did not work or go to school (IMPRESP = 5), their response to the follow-up question referred to above (IMPRESPM) determined the final item score for IMPWORK; otherwise, IMPWORK was recoded similar to the other items.

The transformed scale values were summed across the eight variables created from the WHODAS items (remembering, concentrating, going out of the house on your own, dealing with people you don't know well, participating in social activities, taking care of household responsibilities, taking care of daily work/school responsibilities, and getting your daily work done as quickly as needed) to arrive at a value ranging between 0 and 24. *For both the WHODAS total scores, youths aged 12 to 17 were designated as missing because the mental health section was not administered to youths*.

An alternative WHODAS total score (WHODASC3) was created to indicate the number of daily activities in which a respondent had moderate or severe difficulty performing or did not perform due to problems with emotions, nerves, or mental health. Each of the eight variables created from WHODAS items shown above was transformed into values of 0 or 1 so that responses indicating "moderate difficulty" or "severe difficulty" were coded 1 and responses indicating "mild difficulty" or "no difficulty" were coded 0. The transformed scale values were summed across the eight WHODAS activities to arrive at a value ranging between 0 and 8.

For respondents aged 18 or older, WHODASC3 has no missing values because all missing values in the variables used to create it were imputed as zero.

9.2.3 MHSUTK_U

The recoded suicidal thoughts variable (MHSUITHK) originates from the suicidal thoughts question (SUICTHNK) asked of all adult respondents: "Did you seriously think about killing yourself in the past 12 months?" MHSUITHK was coded as 1 for a "yes" response to the SUICTHNK question, 0 for a "no" response, *and designated as missing for all other responses (i.e., "Don't know," refusals, blanks, and legitimate skips)*.

MHSUTK_U was coded as 1 if MHSUITHK was coded as 1, and coded as 0 in all other cases except when the respondent was aged 12 to 17; that is, *missing MHSUITHK values for respondents aged 18 or older were imputed as zero for MHSUTK_U*.

For respondents aged 18 or older, MHSUTK_U has no missing values because all missing values in the variables used to create it were imputed as zero.

9.2.4 AMDEY2_U

The recoded adult major depressive episode (MDE) variable (AMDEYR) originates from a complex set of parent variables that in combination determine whether the respondent has met the criteria for a positive indication of past year MDE; see the 2010 Mental Health Surveillance Study codebook (Center for Behavioral Health Statistics and Quality, 2011) for further details. AMDEYR was coded as 1 for respondents determined to have a positive indication of past year

MDE, coded as 2 for respondents who did not meet the requirements for a positive indication, *and designated as missing for respondents aged 12 to 17, or for respondents aged 18 or older for whom neither a positive nor negative indication could be determined due to missing values in one or more of the parent variables*.

AMDEY2_U was coded as 1 if AMDEYR was coded as 1, and coded as 0 in all other cases except when the respondent was aged 12 to 17; *that is, missing AMDEYR values for respondents aged 18 or older were imputed as zero for AMDEY2_U*.

For respondents aged 18 or older, AMDEY2_U has no missing values because all missing values in the variables used to create it were imputed as zero.

9.2.5 AGE1830

The final variable in the model, AGE1830, was created from a continuous age variable that has no missing values, so this variable can be used directly in the SMI model to assess item nonresponse among the other predictor variables.

9.3 Creation of Versions of SMI Predictor Variables with Explicit Category for Missing Values

As noted in Section 9.2, four of the five SMI predictor variables (i.e., WSPDSC2, WHODASC3, MHSUTK_U, and AMDEY2_U) had all applicable missing values in the variable creation chain imputed as zero. Therefore, versions of these variables where the applicable missing values are explicitly categorized as missing values need to be used to assess the impact of missing values on all estimates related to these variables.

In the case of MHSUTK_U and AMDEY2_U, their respective parent variables, MHSUITHK and AMDEYR, both of which contain an explicit category of applicable missing values, can be used.

But in the case of WSPDSC2 and WHODASC3, new versions of these variables need to be created with an explicit category for missing values; these new variables are called WSPDSC2_M and WHODASC3_M, respectively. A description of the construction of these new variables follows.

9.3.1 WSPDSC2_M

The creation of WSPDSC2_M proceeds as follows. To simplify the description of the process, all respondents aged 12 to 17 are excluded from the process. These cases would automatically receive a legitimate skip code, and since the age variable has no missing values, there is no ambiguity involved.

Let XXX represent an element from the set {NRV, HOP, RST, CHR, EFF, NGD}.

- Define new variables XXX30_M as follows:
  - If DSTXXX30 in (1, 2, 3, 4, 5) then XXX30_M = 5 – DSTXXX30.
  - Else XXX30_M = SAS missing (.).
- Define new variable K6SCMON_M as follows:
  - If each XXX30_M variable ne SAS missing (.) then K6SCMON_M = sum of all XXX30_M variables.
  - Else K6SCMON_M = SAS missing (.).
- Define new variables XXX12_M as follows:
  - If DSTXXX12 in (1, 2, 3, 4, 5) then XXX12_M = 5 – DSTXXX30.
  - Else if DSTWORST = 2 then XXX12_M = 99 (this is to separate legitimate skips from other missing values).
  - Else XXX12_M = SAS missing (.).
- Define new variable K6SCYR_M as follows:
  - If each XXX12_M variable not in (SAS missing (.), 99) then K6SCYR_M = sum of all XXX12_M variables.
  - Else K6SCYR_M = SAS missing (.).
- Define new variable K6SCMAX_M as follows:
  - If DSTWORST = 2 then K6SCMAX_M = K6SCMON_M.
  - Else if DSTWORST = 1 and K6SCMON_M ne SAS missing (.) and K6SCYR_M ne SAS missing (.) then K6SCMAX_M = max(K6SCMON_M, K6SCYR_M).
  - Else K6SCMAX_M = SAS missing (.).
- Define new variable WSPDSC2_M as follows:
  - If K6SCMAX_M ne SAS missing (.) then WSPDSC2_M = max(0, K6SCMAX_M – 7).
  - Else WSPDSC2_M = SAS missing (.).

9.3.2 WHODASC3_M

The creation of WHODASC3_M proceeds as follows. To simplify the description of the process, all respondents aged 12 to 17 are excluded from the process (these cases would automatically receive a legitimate skip code).

Let XXX represent an element from the set {NRV, HOP, RST, CHR, EFF, NGD}; let YYY represent an element from the set {REMEM, CONCN}; and let ZZZ represent an element from the set {GOUT, PEOP, SOC, HHLD, RESP}.

- Note that the raw (CAI) variable DISTRESS is defined as follows:
  - If (raw versions of) any of (DSTXXX30, DSTXXX12) in (1, 2, 3, 4) then DISTRESS = 1 (i.e., this occurs if any K6 item records a positive score).
  - Else DISTRESS = 2.
- Define new variable DISTRESS_M as follows:
  - If any of (XXX30_M, XXX12_M) > 0 then DISTRESS_M = 1 (i.e., this coincides exactly with DISTRESS = 1)
  - Else if K6SCMAX_M = 0 then DISTRESS_M = 0.
  - Else DISTRESS_M = SAS missing (.).
- Define new variables YYY_M as follows:
  - If IMPYYY = 1 or DISTRESS_M = 0 then YYY_M = 0.
  - Else if IMPYYY in (2, 3, 4) then YYY_M = IMPYYY – 1.
  - Else YYY_M = SAS missing (.).
- Define new variables ZZZ_M as follows:
  - If IMPZZZ = 1 or DISTRESS_M = 0 or (IMPZZZ = 5 and IMPZZZM = 2) then ZZZ_M = 0.
  - Else if IMPZZZ in (2, 3, 4) then ZZZ_M = IMPZZZ – 1.
  - Else if IMPZZZ = 5 and IMPZZZM = 1 then ZZZ_M = 3.
  - Else ZZZ_M = SAS missing (.).
- Define new variable WORK_M as follows:
  - If IMPWORK = 1 or DISTRESS_M = 0 or (IMPRESP = 5 and IMPRESPM = 2) then WORK_M = 0.
  - Else if IMPWORK in (2, 3, 4) then WORK_M = IMPWORK – 1.
  - Else if IMPRESP = 5 and IMPRESPM = 1 then WORK_M = 3.
  - Else WORK_M = SAS missing (.).
- Define new variable WHODASC2_M as follows:
  - If each of (YYY_M, ZZZ_M, WORK_M) ne SAS missing (.) then WHODASC2_M = sum of all (YYY_M, ZZZ_M, WORK_M) variables.
  - Else WHODASC2_M = SAS missing (.).
- Define new variables (YYY2_M, ZZZ2_M, WORK2_M) as follows:
  - If (YYY_M, ZZZ_M, WORK_M) in (0, 1) then (YYY2_M, ZZZ2_M, WORK2_M) = 0.
  - Else if (YYY_M, ZZZ_M, WORK_M) in (2, 3) then (YYY2_M, ZZZ2_M, WORK2_M) = 1.
  - Else (YYY2_M, ZZZ2_M, WORK2_M) = SAS missing (.).

- Define new variable WHODASC3_M as follows:
  - If each of (YYY2_M, ZZZ2_M, WORK2_M) ne SAS missing (.) then WHODASC3_M = sum of all (YYY2_M, ZZZ2_M, WORK2_M) variables.
  - Else WHODASC3_M = SAS missing (.).

9.4 Item Nonresponse Rates of SMI Predictor Variables

Because the SMI predictor variables actually used in the model contain no missing values for adult respondents, item nonresponse rates are based on versions of these variables that do contain an explicit category of missing values; for simplicity, versions with an explicit missingness category are referred to as "m-versions." Item nonresponse counts and rates are calculated with respect to 45,844 adult respondents in the 2010 NSDUH. Item nonresponse counts are presented as unweighted counts, and item nonresponse rates are presented as weighted rates (using the NSDUH analysis weight, ANALWT).

Weighted item nonresponse rates of the m-versions of various K6 variables are shown in Table 9.1. Nonresponse rates for m-versions of all K6 item scores are consistently low (less than 1 percent), and the nonresponse rate for WSPDSC2_M is also low at 1.44 percent.

Table 9.1 Item Nonresponse Counts and Rates for M-Versions of K6 Variables, 2010 NSDUH

K6 Variable	Nonresponse Count	Weighted Nonresponse Rate (Percent)
NRV30_M	163	0.32
HOP30_M	185	0.39
RST30_M	221	0.47
CHR30_M	163	0.36
EFF30_M	387	0.67
NGD30_M	177	0.36
DSTWORST	246	0.60
NRV12_M[1]	287	0.69
HOP12_M[1]	280	0.67
RST12_M[1]	295	0.70
CHR12_M[1]	277	0.64
EFF12_M[1]	329	0.70
NGD12_M[1]	281	0.66
K6SCMON_M	525	1.00
K6SCYR_M[1]	370	0.83
K6SCMAX_M	698	1.44
WSPDSC2_M	698	1.44

K6 = Kessler-6, a 6-item psychological distress scale; m-versions = versions of variables that contain an explicit missingness category.

[1] Note that there were 29,936 (65.30 percent) adult cases that were legitimately skipped (i.e., DSTWORST = 2); these cases were included in the denominator to determine nonresponse rates.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010 (Revised March 2012).

Item nonresponse counts and rates of the edited past month K6 variables (DSTNRV30, DSTHOP30, DSTRST30, DSTCHR30, DSTEFF30, and DSTNGD30) are displayed in Table J.1

of Appendix J, and item nonresponse rates of the edited worst month in past year K6 variables (DSTNRV12, DSTHOP12, DSTRST12, DSTCHR12, DSTEFF12, and DSTNGD12) are displayed in Table J.2. All unweighted and weighted nonresponse rates displayed in both tables are consistently less than 1 percent.

Weighted item nonresponse rates of the m-versions of various WHODAS variables are shown in Table 9.2. Nonresponse rates for m-versions of WHODAS item scores are consistently low (nearly all less than 1 percent), and the nonresponse rate for WHODASC3_M is also low at 1.60 percent.

Table 9.2 Item Nonresponse Counts and Rates for M-Versions of WHODAS Variables, 2010 NSDUH

WHODAS Variable	Nonresponse Count	Weighted Nonresponse Rate (Percent)
DISTRESS_M	212	0.47
REMEM_M, REMEM2_M	431	0.90
CONCN_M, CONCN2_M	415	0.91
GOUT_M, GOUT2_M	387	0.76
PEOP_M, PEOP2_M	412	0.83
SOC_M, SOC2_M	414	0.87
HHLD_M, HHLD2_M	391	0.83
RESP_M, RESP2_M	427	1.04
WORK_M, WORK2_M	397	0.88
WHODASC2_M	688	1.60
WHODASC3_M	688	1.60

M-versions = versions of variables that contain an explicit missingness category; WHODAS = World Health Organization Disability Assessment Schedule.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010 (Revised March 2012).

Item nonresponse counts and rates of the edited past month WHODAS variables (IMPREMEM, IMPCONCN, IMPGOUT, IMPGOUT2, IMPPEOP, IMPPEOP2, IMPSOC, IMPSOC2, IMPHHLD, IMPHHLD2, IMPRESP, IMPRESP2, and IMPWORK) are displayed in Table J.3 of Appendix J. All unweighted and weighted nonresponse rates displayed in Table J.3 are consistently less than 1 percent.

Weighted item nonresponse rates for MHSUITHK and AMDEYR are shown in Table 9.3, which indicates that the rates are both low (less than 1 percent).

Table 9.3 Item Nonresponse Counts and Rates for MHSUITHK and AMDEYR Variables, 2010 NSDUH

WHODAS Variable	Nonresponse Count	Weighted Nonresponse Rate (Percent)
MHSUITHK	163	0.32
AMDEYR	385	0.73

WHODAS = World Health Organization Disability Assessment Schedule.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010 (Revised March 2012).

9.5 Item Nonresponse Rates for Mental Health Variables Available in Other Surveys

Four other surveys were examined to see if they also collected data corresponding to any of the variables associated with the NSDUH SMI prediction model. These four surveys included the 2009 Behavioral Risk Factor Surveillance System (BRFSS), the 2009 Medical Expenditure Panel Survey (MEPS), the 2001-2003 National Comorbidity Survey Replication (NCS-R), and the 2010 National Health Interview Survey (NHIS).

The BRFSS, MEPS, and NHIS all collected data related to past month K6 variables, and the NCS-R collected data related to worst month in past year K6 variables, past year suicidal thoughts, and past year MDE. Weighted nonresponse rates associated with these variables were compared with those of the corresponding NSDUH variables.

Weighted item nonresponse rates for the past month K6 variables obtained from the BRFSS, MEPS, and NHIS and from the 2010 NSDUH are displayed in Table 9.4. Item nonresponse rates for the past month K6 variables differ markedly among some of the four surveys. For example, in the BRFSS the nonresponse rates are all above 8 percent; although the section containing the K6 questions was applied only to eight states (Georgia, Hawaii, Mississippi, Missouri, Nevada, South Carolina, Vermont, and Wyoming), this geographic restriction does not seem to be sufficient to explain the much higher nonresponse rates. In the MEPS, the nonresponse rates are all above 2 percent, which are slightly but consistently higher than those of the remaining two surveys (NHIS and NSDUH), whose nonresponse rates are consistently below 1 percent.

Table 9.4 Item Nonresponse Rates for Past Month K6 Variables in BRFSS, MEPS, NHIS, and NSDUH

Past Month K6 Variable	BRFSS	MEPS	NHIS	NSDUH
Nervous	8.43	2.02	0.68	0.32
Hopeless	8.45	2.02	0.73	0.39
Restless	8.53	2.14	0.67	0.47
Sad/Depressed	8.41	2.02	0.69	0.36
Effort	9.42	2.18	0.75	0.67
Worthless	8.55	2.19	0.71	0.36

K6 = Kessler-6, a 6-item psychological distress scale.

Note: The item nonresponse rates presented are weighted rates using the respective survey weight. The rate is the total number of logically assigned, missing, and blank cases divided by the number of applicable cases.

Sources: 2009 BRFSS, https://www.cdc.gov/brfss/annual_data/annual_2009.htm
2009 MEPS, https://meps.ahrq.gov/mepsweb/data_stats/download_data_files.jsp
2010 NHIS, https://www.cdc.gov/nchs/nhis/nhis_2010_data_release.htm
2010 NSDUH Final Analytic Data File, SAMHSA, Center for Behavioral Health Statistics and Quality (Revised March 2012)

Table 9.5 indicates that item nonresponse rates for K6 variables most closely associated with worst month in past year scores, past year suicidal thoughts, and past year MDE are very low (almost all less than 1 percent) for the NCS-R and NSDUH.

To summarize, although it is difficult to make direct comparisons in nonresponse rates across different surveys due to their different objectives, designs, and questionnaire contexts within which the applicable mental health questions have been embedded, among other things, Tables 9.4 and 9.5 do indicate that nonresponse rates for mental health variables in the 2010 NSDUH compare very favorably with those available from the other four surveys.

Table 9.5 Item Nonresponse Rates for Other Mental Health Variables in NCS-R and NSDUH

Mental Health Variable	NCS-R	NSDUH
Worst Month in Past Year K6 Variable[1]		
Nervous	1.33	0.69
Hopeless	0.41	0.67
Restless	0.44	0.70
Sad/Depressed	0.37	0.64
Effort	0.56	0.70
Worthless	0.40	0.66
Past Year Suicidal Thoughts	0.84[2]	0.32
Past Year MDE	0.04[2]	0.73

K6 = Kessler-6, a 6-item psychological distress scale; MDE = major depressive episode.

Note: The item nonresponse rates presented are weighted rates using the respective survey weight. The rate is the total number of logically assigned, missing, and blank cases divided by the number of applicable cases.

[1] In NCS-R, worst month in past year K6 item variables are specifically defined as such. In NSDUH, no such item variables exist; the closest equivalent item variables are past year K6 item variables asked of respondents if they had indicated that the past month was not the worst month in the past year.

[2] No missing values were recorded for the final NCS-R past year suicidal thoughts or past year MDE variables; however, missing values were observed in screener variables used to determine if a respondent was administered those questions, and these missing values were used to calculate the nonresponse rate.

Sources: 2001-2003 NCS-R, https://www.hcp.med.harvard.edu/ncs/ncs_data.php
2010 NSDUH Final Analytic Data File, SAMHSA, Center for Behavioral Health Statistics and Quality (Revised March 2012)

9.6 Evaluating the Need for Imputation of the Mental Health Variables

To determine how missing data affect prevalence estimates of each of the predictor variables of the SMI model and the resulting SMI and AMI prevalence estimates themselves, a sensitivity analysis was performed to assess the range of possible values for the prevalence estimates of these variables after different methods of imputation were applied. The following imputation methods were compared in the analysis:

- *Complete Case (CC) Analysis:* Only the complete cases of the m-versions of variables used to create the SMI predictor variables were used. A surprising consequence of this sensitivity analysis was the discovery that many SMI and AMI responses could be "reclaimed" from respondents for whom one or more of the m-versions of predictor variables had missing values (see Section 9.6.1 for more details). The denominator of the prevalence estimate for each variable was calculated as the weighted number of respondents corresponding to the set of complete cases, respectively.
- *Impute to Zero (I_0) Method/Current Method (CM):* Missing values for each of the m-versions of the variables used to create the SMI predictor variables was assigned a value of zero, thus making the overall estimate as low as possible. The numerator was

the same as with the CC analysis, but the denominator was the weighted number of adult respondents in the entire NSDUH sample. Note that the I_0 method is identical to the current method (CM) because the actual SMI predictor variables in current use are identical to their m-versions after application of the I_0 method.

- *Impute to Maximum (I_{Max}) Method:* Missing values for each of the m-versions of the variables used to create the SMI predictor variables was assigned the maximum possible value, thus making the overall estimate as high as possible. The denominator of the prevalence estimate for each variable was calculated as the weighted number of adult respondents in the entire NSDUH sample.

The I_0 method provides the lower bound of prevalence estimates of each of SMI, AMI, and the SMI predictor variables, and the I_{Max} method provides the upper bound. Therefore, if the I_{Max} point estimate lies outside the I_0/CM confidence interval, then that would suggest some evidence of bias due to nonresponse for that variable, and consequently an alternative method of imputation method might reduce that bias.

9.6.1 Reclaimed SMI and AMI Responses

The lower bound and upper bound properties of the I_0 and I_{Max} imputation methods, respectively, can be used to "reclaim" certain SMI and AMI prevalence estimates even when some of the m-versions of the SMI predictor variables have missing values.

Consider the subset of adult respondents for whom at least one of the m-versions of the SMI predictor variables has missing values. Let $SMIPP_0$ refer to the SMI predicted probability under I_0 and let $SMIPP_{Max}$ refer to it under I_{Max}, noting that because all predictor variables in the SMI model have positive coefficients, $SMIPP_0 \leq SMIPP_{Max}$ in all cases. Let C_S refer to the SMI cut point and let C_A refer to the AMI cut point. The following six scenarios can be used to reclaim SMI and AMI estimates:

1. $SMIPP_{Max} < C_A$: This implies that both SMI and AMI will be predicted to be negative regardless of what is used to impute missing value(s) in the predictor variables. Therefore, both SMI and AMI estimates can be reclaimed (i.e., they can be established without any ambiguity).
2. $SMIPP_0 < C_A \leq SMIPP_{Max} < C_S$: This implies that only SMI will be predicted to be negative regardless of what is used to impute missing value(s) in the predictor variables. Therefore, only SMI estimates can be reclaimed.
3. $SMIPP_0 < C_A, C_S \leq SMIPP_{Max}$: This implies that neither SMI nor AMI can be predicted without ambiguity depending on the imputation method. Therefore, neither SMI nor AMI estimates can be reclaimed.
4. $C_A \leq SMIPP_0$, $SMIPP_{Max} < C_S$: This implies that SMI will be predicted to be negative and AMI will be predicted to be positive regardless of what is used to impute missing value(s) in the predictor variables. Therefore, both SMI and AMI estimates can be reclaimed.
5. $C_A \leq SMIPP_0 < C_S \leq SMIPP_{Max}$: This implies that only AMI will be predicted to be positive regardless of what is used to impute missing value(s) in the predictor variables. Therefore, only AMI estimates can be reclaimed.

6. $C_S \leq SMIPP_0$: This implies that both SMI and AMI will be predicted to be positive regardless of what is used to impute missing value(s) in the predictor variables. Therefore, both SMI and AMI estimates can be reclaimed.

The extent to which SMI and AMI prevalence estimates were reclaimed for each of the six scenarios described above can be seen in Table 9.6. This table shows that there were 1,232 (2.69 percent) adult respondents for whom at least one of the m-versions of the SMI predictor variables had missing values and that of these cases, 999 (81.1 percent) SMI cases and 643 (52.2 percent) AMI cases were reclaimed.

Table 9.6 Reclaimed Counts and Percentages for SMI and AMI, 2010 NSDUH

Scenario	Total Missing Count	SMI Reclaimed Count	AMI Reclaimed Count	SMI Reclaimed Percentage	AMI Reclaimed Percentage
1. $SMIPP_{Max} < C_A$	398	398	398	32.3	32.3
2. $SMIPP_0 < C_A \leq SMIPP_{Max} < C_S$	413	413	0	33.5	0.0
3. $SMIPP_0 < C_A, C_S \leq SMIPP_{Max}$	176	0	0	0.0	0.0
4. $C_A \leq SMIPP_0, SMIPP_{Max} < C_S$	150	150	150	12.2	12.2
5. $C_A \leq SMIPP_0 < C_S \leq SMIPP_{Max}$	57	0	57	0.0	4.6
6. $C_S \leq SMIPP_0$	38	38	38	3.1	3.1
Total	1,232	999	643	81.1	52.2

AMI = any mental illness; C_A = AMI cut point; C_S = SMI cut point; SMI = serious mental illness; $SMIPP_0$ = SMI predicted probability under imputation method I_0; $SMIPP_{Max}$ = SMI predicted probability under imputation method I_{Max}.
Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010 (Revised March 2012).

Table 9.7 presents the effect of SMI and AMI reclaimed counts on their respective item nonresponse counts. This table shows that the SMI nonresponse count was reduced to 233 (0.51 percent) and the AMI nonresponse count to 589 (1.28 percent). Thus, the actual SMI and AMI weighted nonresponse rates were both 2.65 percent, but after reclamation the effective weighted nonresponse rates were reduced to 0.55 percent for SMI and 1.41 percent for AMI.

Table 9.7 Nonresponse Counts for SMI Predictor Variables, SMI, and AMI, 2010 NSDUH

WSPDSC2_M	WHODASC3_M	MHSUITHK	AMDEYR	Missing Count	SMI Missing Count	AMI Missing Count
X	X	X	.	155	13	59
X	X	.	X	31	8	9
X	.	X	X	283	4	92
.	X	X	X	314	3	55
X	X	.	.	20	7	14
X	.	X	.	25	4	16
.	X	X	.	19	4	9
X	.	.	X	9	2	6
.	X	.	X	2	0	1
.	.	X	X	204	34	170
X	.	.	.	11	11	9
.	X	.	.	3	2	2
.	.	X	.	69	54	61
.	.	.	X	4	4	4
.	.	.	.	83	83	82
			Total	**1,232**	**233**	**589**

AMI = any mental illness; SMI = serious mental illness.
Note: X indicates responses not missing for the variable in question; a period indicates nonresponse.
Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010 (Revised March 2012).

9.6.2 Sensitivity Analysis Results

The sensitivity analysis results related to the various imputation methods are presented in Table 9.8. For each of the three methods, CC, I_0, and I_{Max} (recalling that I_0 is identical to the current method [CM]), the point estimates and the 95 percent confidence intervals are shown for the four SMI predictor variables with missing values, SMI, and AMI. For a particular variable, if the I_{Max} point estimate lies outside the I_0/CM confidence interval, then imputation would be recommended because this would imply the potential for imputation to move the estimate out of the range that is considered reasonable in the national estimates. For all the variables in question, the I_{Max} point estimate did exceed the upper bound of the I_0/CM confidence interval. Therefore, an investigation of an alternative imputation method for all these variables is recommended.

Table 9.8 Imputation Results for SMI Predictor Variables, SMI, and AMI, 2010 NSDUH

Variable	Method	Estimate[1]	Lower Confidence Limit	Upper Confidence Limit
SMI	CC	4.1	3.8	4.4
	I_0/CM	4.1	3.8	4.4
	I_{Max}	4.6	4.3	5.0
AMI	CC	18.3	17.7	18.9
	I_0/CM	18.1	17.5	18.7
	I_{Max}	19.5	18.9	20.1
WSPDSC2_M	CC	1.5	1.4	1.5
	I_0/CM	1.4	1.4	1.5
	I_{Max}	1.6	1.5	1.6
WHODASC3_M	CC	0.9	0.8	0.9
	I_0/CM	0.9	0.8	0.9
	I_{Max}	0.9	0.9	1.0
MHSUITHK	CC	3.8	3.6	4.1
	I_0/CM	3.8	3.5	4.1
	I_{Max}	4.1	3.8	4.4
AMDEYR	CC	6.8	6.5	7.2
	I_0/CM	6.8	6.4	7.2
	I_{Max}	7.5	7.1	7.9

AMI = any mental illness; CC = complete case; CM = current method; I_0 = impute to zero; I_{Max} = impute to maximum value; K6 = Kessler-6, a 6-item psychological distress scale; SMI = serious mental illness; WHODAS = World Health Organization Disability Assessment Schedule.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

[1] SMI, AMI, MHSUITHK, and AMDEYR estimates are prevalence estimates expressed as percentages; WSPDSC2_M and WHODASC3_M estimates are means of K6 and WHODAS total scores, respectively.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010 (Revised March 2012).

9.7 Alternative Imputation Method

The current method for treating missing values in the mental health variables used to derive SMI and AMI estimates is simply to impute them as zero. Clearly, this is a conservative approach that downwardly biases the estimates because zero is the lower bound of all the item scores. The question is how substantive is this bias considering the low nonresponse rate and the

potential values of the item scores under an alternative imputation method (i.e., how different from zero are the imputed values likely to be?). The results from the sensitivity analysis suggest that an imputation method should be tested because the point estimates for all mental health variables showed substantial differences between I_0/CM and I_{Max} methods and the I_{Max} estimates were outside the range of the I_0/CM confidence intervals.

To gauge the level of downward bias resulting from the current practice of imputing missing item responses as zero, a weighted sequential hot-deck (WSHD) imputation method (as described in Chapter 3 of this report) was implemented on the 2010 NSDUH data. The imputation classes were defined by using a nonparametric classification tree analysis (Breiman, Friedman, Olshen, & Stone, 1984) similar to Chi-square Automatic Interaction Detection (CHAID) analysis (Section 3.1.1). An implicit assumption is that the WSHD imputation method results in unbiased estimates, so differences in paired estimates based on the CM versus WSHD imputation methods should provide an estimate of the bias associated with the CM imputation method.

A cyclic approach for the WSHD imputation was implemented similar to the cyclical methods discussed in Chapter 4 for IVEware and Chapter 5 for modPMN-MI method. The cyclic hot-deck imputation is discussed in Marker, Judkins, and Winglee (2002). The first cycle uses the complete responses and any previously imputed variables to develop imputation classes and perform imputations. For each subsequent imputation cycle, all variables on the dataset are available for the tree-based methodology to create imputation classes. Cycling is advantageous because it determines the best relationships among variables and minimizes the relationships among variables that may be caused by the sequential nature of the imputation process. For item-level mental health variables, only two cycles were performed to allow each of the item-level variables to be used as a predictor variable.

The set of predictor variables for the CHAID analysis included the following: age, gender, race/Hispanicity recode (7 levels), poverty level (3 levels), education level (4 levels), and employment status (4 levels). It also always includes all other variables in the same imputation set. Tables J.4 through J.9 in Appendix J show the imputation classes that resulted from the CHAID analysis for all mental health variables used in the imputation process.

Table 9.9 shows the paired estimates of the SMI predictor variables, SMI, and AMI based on the CM versus WSHD imputation methods. For each of the six variables, the point estimate based on WSHD always lies within the 95 percent confidence interval around the corresponding point estimate based on CM*, indicating little evidence of substantive differences in estimates between the two imputation methods*.

In fact, Table 9.9 indicates that the paired SMI estimates are very similar (4.1 percent for CM versus 4.1 percent for WSHD), but the paired AMI estimates are somewhat less similar (18.1 percent for CM versus 18.2 percent for WSHD). It is interesting to note that the AMI estimate based on the WSHD method is less than that based on the CC method (Table 9.8 indicates 18.3 percent for CC).

Table 9.9 Comparison of Estimates of SMI Predictor Variables, SMI, and AMI Based on Current Method and WSHD Imputation, 2010 NSDUH

Variable	Method	Estimate[1]	Lower Confidence Limit	Upper Confidence Limit
SMI	I_0/CM	4.1	3.8	4.4
	WSHD	4.1	3.8	4.4
AMI	I_0/CM	18.1	17.5	18.7
	WSHD	18.2	17.6	18.8
WSPDSC2_M	I_0/CM	1.4	1.4	1.5
	WSHD	1.5	1.4	1.5
WHODASC3_M	I_0/CM	0.9	0.8	0.9
	WSHD	0.9	0.8	0.9
MHSUITHK	I_0/CM	3.8	3.5	4.1
	WSHD	3.8	3.6	4.1
AMDEYR	I_0/CM	6.8	6.4	7.2
	WSHD	6.9	6.5	7.2

AMI = any mental illness; CM = current method; I_0 = impute to zero; K6 = Kessler-6, a 6-item psychological distress scale; SMI = serious mental illness; WHODAS = World Health Organization Disability Assessment Schedule; WSHD = weighted sequential hot deck.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

[1] SMI, AMI, MHSUITHK, and AMDEYR estimates are prevalence estimates expressed as percentages; WSPDSC2_M and WHODASC3_M estimates are means of K6 and WHODAS total scores, respectively.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010 (Revised March 2012).

Although the paired AMI estimates do not appear to be substantively different (because the WSHD point estimate lies within the 95 percent confidence interval of the respective CM point estimate), it might be useful to examine what could be driving this somewhat larger difference.

Paired differences within WSPDSC2_M, WHODASC3_M, and MHSUITHK are extremely small (0.01 percent in each case), so none of these variables are likely to be contributing to the observed difference in paired AMI estimates. However, the difference between the pair of AMDEYR estimates is a little larger (6.8 percent for CM versus 6.9 for WSHD), so this variable appears to be the only likely candidate driving the difference in paired AMI estimates. And a possible reason why paired AMDEYR differences might contribute to a larger difference in paired AMI estimates but not in paired SMI estimates might be due to the higher data reclamation rate for SMI (see Section 9.5.1), resulting in an effective rate of missing data for SMI that is considerably smaller (0.55 percent) than that for AMI (1.41 percent).

A further examination of these paired differences across several imputation classes (age, gender, and race/Hispanicity) is displayed in Tables J.10 and J.11 of Appendix J, and it appears that the differences are carried across the imputation classes fairly consistently (i.e., there appears to be little evidence of interaction effects).

9.8 Summary

The amount of missing data for the item-level mental health variables used to create the SMI predictor variables is low (around 1 percent), missingness rates for the predictor variables themselves were also low (less than 2 percent), and the effective missingness rate for SMI and AMI was even lower due to the fact that many values could be "reclaimed" even though one or more predictor variables had missing values.

The sensitivity analysis showed that the estimates of the SMI predictor variables, SMI, and AMI may be affected by imputation because the upper bound estimate (I_{Max}) was not in the I_0/CM confidence interval, indicating that the estimates may be underestimated under the current imputation method. Although all the mental health variables in question have high response rates, the differences between the estimate under I_0/CM and the estimate under I_{Max} are substantive and an alternative imputation method might correct at least some of the negative bias in the I_0/CM method.

None of the pairs of estimates of the mental health variables based on the CM versus the WSHD methods appear to be substantively different, because in each case the WSHD point estimate lies within the 95 percent confidence interval of the respective CM point estimate. In fact, paired estimates for SMI and the three predictor variables, WSPDSC2_M, WHODASC3_M, and MHSUITHK, are extremely small, but the paired differences for AMI and AMDEYR are somewhat larger.

These results regarding SMI are encouraging in that they indicate that an alternative imputation method is likely to have almost no impact on the estimation of SMI (because the negative bias resulting from use of the current method is negligible), *and hence an alternative imputation method would not be required to estimate SMI*.

These results regarding AMI are slightly less encouraging because, although no substantive difference between AMI estimates based on the two imputation methods is apparent, the observed difference does not appear to be entirely negligible. However, if a recommendation to use an alternative imputation method is to be based on the observation of substantive differences between current and alternative imputation methods, *then the results indicate that the current method of imputation is sufficient for providing AMI estimates*.

10. Imputation Methods for Substance Dependence and Abuse Variables

This chapter examines the extent and nature of item nonresponse for the National Survey on Drug Use and Health (NSDUH) substance dependence and abuse variables to determine whether imputation would substantially improve the quality of the data. The current method used for handling missing values in these variables produces estimates with a negative bias because missing values are effectively replaced with values of zero (indicating the absence of the symptom). This chapter assesses the amount of negative bias induced by the current method and suggests whether imputation is needed for these variables. Nonresponse was evaluated at the item level (i.e., individual variables that are used to create criteria for dependence and abuse), the criterion level (i.e., combination of item-level variables based on the *Diagnostic and Statistical Manual of Mental Disorders* [DSM-IV], 4th edition [American Psychiatric Association, 1994] criteria), and the overall dependence/abuse indicator level. Section 10.1 presents a brief introduction and definitions for substance dependence and abuse. Section 10.2 presents information on the different routes through which missingness can be introduced in the substance abuse and dependence criteria. Section 10.3 discusses the item nonresponse rates for variables that are used to create the criteria for abuse and dependence (presented in Appendix K). Section 10.4 presents details on the item nonresponse patterns at the criterion level where dependence or abuse could be affected based on missing data. Section 10.5 shows the results of a sensitivity analysis that assesses the range of possible estimates for dependence and abuse from different imputation methods. Section 10.6 summarizes the results of performing imputation using weighted sequential hot-deck imputation, and Section 10.7 provides a summary as well as recommendations based on these analyses.

10.1 Description of the Substance Dependence and Abuse Variables

This section contains a brief description of the dependence and abuse variables that were assessed for missingness. The NSDUH questions are designed to measure dependence and abuse of illicit drugs, alcohol, and dependence on nicotine (cigarettes).

For nicotine (cigarettes), dependence is based on criteria from the Nicotine Dependence Syndrome Scale (NDSS) and a simplified version of the Fagerstrom Test of Nicotine Dependence (FTND). A respondent is defined as being dependent if he or she met either the NDSS or the FTND criteria for dependence. The NDSS score, NDSSANSP, is calculated as the average score over 17 questions pertaining to five aspects of dependence. Based on the NDSS score, a respondent is defined as having nicotine dependence, NDSSDNSP = 1, if their average score is greater than or equal to 2.75. The FTND measure is defined by assessing how soon after waking a respondent had his or her first cigarette (CIGWAKE). Based on the FTND scale, a respondent is defined as having nicotine dependence, FTNDDNSP = 1, if the first cigarette was smoked within 30 minutes of waking up on the days he or she smoked (CIGWAKE = 1 or 2) and the respondent reported smoking cigarettes in the past month. Based on the NDSS and the FTND, a respondent who reported smoking cigarettes is defined as having nicotine dependence

in the past month, DNICNSP = 1, if he or she met either the NDSS or the FTND criteria for dependence.

For respondents who answered 16 of the 17 NDSS nicotine (cigarette) dependence questions used in the NDSS algorithm, imputation was implemented using the 16 of the NDSS item-level variables as covariates in a weighted least squares regression model (Frechtel et al., 2013). For substances other than nicotine (cigarettes), which include alcohol, cocaine, heroin, pain relievers, sedatives, marijuana, tranquilizers, stimulants, hallucinogens, and inhalants, dependence and abuse are based on the DSM-IV criteria (American Psychiatric Association, 1994). The Recoded Substance Dependence and Abuse Variable Documentation Appendix of the 2011 Analytic Codebook (RTI International, 2012) describes the criteria used for defining each of the individual substance dependence and abuse variables in detail.

A respondent is defined as having marijuana, inhalant, hallucinogen, or tranquilizer dependence (DEPNDMRJ, DEPNDINH, DEPNDHAL, DEPNDTRN) if the respondent reported a positive response to three or more of the six dependence criteria[52] listed below:

1. Spent a great deal of time over a period of a month getting, using, or getting over the effects of the substance (xxxLOTTM = Yes or xxxGTOVR = Yes).
2. Unable to keep set limits on substance use or used more often than intended (xxxKPLMT = Yes).
3. Needed to use substance more than before to get desired effects or noticed that using the same amount had less effect than before (xxxNDMOR = Yes or xxxLSEFX = Yes).
4. Able to cut down or stop using the substance every time he or she tried or wanted to (xxxCUTEV = No).
5. Continued to use substance even though it was causing emotional, nervous, mental health, or physical problems (xxxEMCTD = Yes or xxxPHCTD = Yes).
6. Reduced or gave up participation in important activities due to substance use (xxxLSACT = Yes).

An additional question pertaining to withdrawal symptoms was asked for the following six substances: alcohol, pain relievers, cocaine, heroin, sedatives, and stimulants. The withdrawal question asked the respondent if he or she had experienced substance-specific withdrawal symptoms at one time that lasted for longer than a day after he or she cut back or stopped using. The specific number and type of listed withdrawal symptoms varied by substance. A respondent was defined as having alcohol, pain reliever, cocaine, heroin, sedative, or stimulant dependence (DEPNDALC, DEPNDANL, DEPNDCOC, DEPNDHER, DEPNDSED, DEPNDSTM) if the respondent reported a positive response to three or more of the seven dependence criteria (including the six standard criteria listed above, plus a seventh withdrawal symptom criteria).

A respondent is defined as having alcohol, marijuana, cocaine, heroin, hallucinogen, inhalant, pain reliever, tranquilizer, stimulant, or sedative abuse (ABUSEALC, ABUSEMRJ, ABUSECOC, ABUSEHER, ABUSEHAL, ABUSEINH, ABUSEANL, ABUSETRN,

[52] The abbreviations used in the substances dependence variables are as follows: marijuana (MRJ), inhalants (INH), hallucinogens (HAL), tranquilizers (TRN), alcohol (ALC), prescription pain relievers (ANL), cocaine (COC), heroin (HER), sedatives (SED), and stimulants (STM).

ABUSESTM, ABUSESED), if a respondent did not have dependence for that substance and reported a positive response to one or more of the four abuse criteria listed below:

1. Respondent reported having serious problems due to substance use at home, work, or school (xxxSERPB = Yes).
2. Respondent reported using substance regularly and then did something where substance use might have put him or her in physical danger (xxxPDANG = Yes).
3. Respondent reported substance use was causing actions that repeatedly got him or her in trouble with the law (xxxLAWTR = Yes).
4. Respondent reported having problems caused by substance use with family or friends (xxxFMFPB = 1) and continued to use substance even though it was thought to be causing problems with family and friends (xxxFMCTD = Yes).

For the purposes of this report, the following definitions are used to assist with discussion of the item nonresponse rates and missing data patterns. An item-level variable is a questionnaire variable that is used to create criterion-level variables. For example, ALCLOTTM is an item-level variable for alcohol dependence. A criterion-level variable is a variable that combines two or more item-level variables to check a particular condition for dependence or abuse. For example, the first criterion-level variable for alcohol is based on item-level variables ALCLOTTM and ALCGTOVR. Finally, an indicator-level variable is a variable that combines two or more criterion-level variables and determines respondent's final dependence or abuse status.

There is no formal imputation procedure used for the item-level substance dependence and abuse variables. When an item-level variable is used in the computation for the dependence or abuse criteria, missing values are considered as not meeting the criteria for dependence or abuse (i.e., they are treated the same as a "No" response) because the computation is only counting the number of positive responses to determine if the respondent meets the criteria.

10.2 Inconsistencies between Domain Variables and Variables for Nicotine Dependence, Alcohol Dependence, and Alcohol Abuse

This section documents the inconsistencies between core and noncore questionnaire variables for nicotine dependence and alcohol abuse or dependence and discusses whether any of these inconsistencies affect the respondent's classification for substance use disorder. The core variables are used to define the "domain" (i.e., the set of respondents who are eligible for the dependence and abuse questions). The inconsistencies occur when either the imputation-revised core variables suggest that the respondent belongs in the domain, but the respondent never saw the dependence and abuse questions, or the imputation-revised core variables suggest that the respondent does not belong in the domain, but the respondent did see the dependence and abuse questions. The Recoded Substance Dependence and Abuse Variable Documentation Appendix of the 2011 Analytic Codebook (RTI International, 2012) includes the following specific notes about inconsistency issues for each measure. The first quote is only applicable to substances other than nicotine because it refers to past year use, and the second quote refers to all substances.

A respondent might have provided ambiguous information about past year use of any individual substance, in which case these respondents were not asked the dependence and abuse questions for that substance. Subsequently, these respondents could be imputed to be past year users of the respective substance. In this situation, the dependence and abuse data were unknown. Thus, these respondents were classified as not dependent on or abusing the respective substance, without ever having been asked the dependence and abuse questions.

Responses from core substance use, frequency of substance use, and noncore substance use questions were used as criteria to determine whether a respondent was asked the alcohol and/or illicit drug dependence and abuse questions. Unknown responses to the core substance use and frequency of substance use questions were imputed. However, the imputation process did not take into account data reported in the noncore CAI sections. Therefore, responses to the dependence and abuse questions that were inconsistent with responses to the imputed substance use or frequency of use questions could exist. These inconsistent responses remained in the edited data and were not excluded in the creation of the dependence and abuse recodes. Since different sets of criteria were used as skip logic for each substance, different types of inconsistencies between the dependence and abuse variables and the imputed substance use and frequency of substance use variables could occur by substance (RTI International, 2012).

Before inconsistencies can be discussed, it is important to define the domain for each substance dependence and abuse indicator-level variable. The domain is defined as the set of respondents for which questions are applicable. Most of the domains are based on the imputation-revised domains (i.e., the set of respondents who should have been asked the questions based on the imputation-revised values of all variables related to the skip logic). For most of the dependence and abuse measures, the imputation-revised domain is completely defined; that is, the imputation-revised variables unambiguously determine whether the respondent belongs in the domain. However, for cocaine, heroin, and stimulants, there are some cases for which domain status is uncertain because one or more of the variables that determines the domain does not undergo imputation. The domain definitions for each substance are described below and summarized in Table 10.1.

- For nicotine dependence, domain members are those who used cigarettes in the past month according to the imputation-revised variable IRCIGRC (or, equivalently, the dichotomous recode CIGMON).
- For dependence and abuse of alcohol (xxx = ALC) and marijuana (xxx = MJ), domain members are those who used the given substance on at least 6 days in the past 12 months according to the imputation-revised variable IRxxxFY.
- For dependence and abuse of hallucinogens (xxx = HAL), inhalants (xxx = INH), pain relievers (xxx = ANL), tranquilizers (xxx = TRN), and sedatives (xxx = SED), domain members are those who used the given substance in the past year according to the imputation-revised variable IRxxxRC (or, equivalently, the dichotomous recode xxxYR).
- For dependence and abuse of cocaine, domain members are those who used the given substance in the past year according to the imputation-revised variable IRCOCRC (or,

equivalently, the dichotomous recode COCYR), or used a needle to inject cocaine in the past year according to the item-level variable CONDLREC. CONDLREC has some missing values that were not imputed as the standard NSDUH imputation process. Therefore, the domain status is uncertain for respondents who are not past year users of cocaine according to IRCOCRC but have a missing value for CONDLREC.[53]

- For dependence and abuse of heroin, domain members are those who used the given substance in the past year according to the imputation-revised variable IRHERRC (or, equivalently, the dichotomous recode HERYR); smoked heroin in the past year according to the item-level variable HRSMKREC; sniffed heroin in the past year according to the item-level variable HRSNFREC; or used a needle to inject heroin in the past year according to the item-level variable HRNDLREC. Except for IRHERRC, the other variables that determined whether a respondent would get routed into the heroin dependence and abuse section were not imputed. Therefore, the domain status is uncertain for respondents who are not past year users of heroin according to IRHERRC; have at least one missing response among HRSMKREC, HRSNFREC, and HRNDLREC; and have no past year responses among HRSMKREC, HRSNFREC, and HRNDLREC.
- For dependence and abuse of stimulants, domain members are those who used the given substance in the past year according to the imputation-revised variable IRSTMRC (or, equivalently, the dichotomous recode STMYR); used a needle to inject methamphetamine in the past year according to the item-level variable MTNDLREC; or used a needle to inject some other stimulant in the past year according to the item-level variable OSTNLREC. Except for IRSTMRC, the other variables that determined whether a respondent would get routed into the stimulant dependence and abuse section were not imputed. Therefore, the domain status is uncertain for respondents who are not past year users of stimulants according to IRSTMRC; have at least one missing response between MTNDLREC and OSTNLREC; and have no past year responses between MTNDLREC and OSTNLREC.[54]

[53] The question associated with CONDLREC is only asked of lifetime users of cocaine. Those who were imputed to lifetime nonuse of cocaine were considered to be not in the domain for cocaine dependence and abuse. Similar ideas apply to the domains for dependence and abuse for heroin and stimulants.

[54] There are a handful of cases with one of the following two patterns: (1) known to be a lifetime user of methamphetamine, not a past year user of methamphetamine according to the imputation-revised variable IRMTHRC, known not to be a past year user of methamphetamine according to MTNDLREC, and imputed to lifetime nonuse of other stimulants; (2) known to be a lifetime user of other stimulants, not a past year user of other stimulants according to the imputation-revised variable IRMTHRC, known not to be a past year user of other stimulants according to OSTNLREC, and imputed to lifetime nonuse of methamphetamine. These cases were treated as domain nonmembers. The reasoning is that lifetime nonusers are not in the domain for the needle use variables MTNDLREC and OSTNLREC.

Table 10.1 Domain Definitions for Substance Dependence and Abuse Indicator-Level Variables

Drug	Domain for Dependence and Abuse	Dependence	Abuse
Cigarettes	Past month users (core only)	NDSS score > 2.75 or FTND	N/A
Alcohol	Past year users with 12-month frequency ≥ 6 (core only)	Met at least 3 of 7 criteria	No dependence and met at least 1 of 4 criteria
Marijuana	Past year users with 12-month frequency ≥ 6 (core only)	Met at least 3 of 6 criteria	No dependence and met at least 1 of 4 criteria
Cocaine	Past year users (core or noncore)	Met at least 3 of 7 criteria	No dependence and met at least 1 of 4 criteria
Heroin	Past year users (core or noncore)	Met at least 3 of 7 criteria	No dependence and met at least 1 of 4 criteria
Hallucinogens	Past year users (core only)	Met at least 3 of 6 criteria	No dependence and met at least 1 of 4 criteria
Inhalants	Past year users (core only)	Met at least 3 of 6 criteria	No dependence and met at least 1 of 4 criteria
Pain Relievers	Past year users (core only)	Met at least 3 of 7 criteria	No dependence and met at least 1 of 4 criteria
Tranquilizers	Past year users (core only)	Met at least 3 of 7 criteria	No dependence and met at least 1 of 4 criteria
Stimulants	Past year users (core or noncore)	Met at least 3 of 7 criteria	No dependence and met at least 1 of 4 criteria
Sedatives	Past year users (core only)	Met at least 3 of 7 criteria	No dependence and met at least 1 of 4 criteria

FTND = Fagerstrom Test of Nicotine Dependence; N/A = not applicable; NDSS = Nicotine Dependence Syndrome Scale.

Using the 2011 NSDUH data, membership in the domain was cross-classified with an indicator of whether the respondent actually saw the questions, for both nicotine dependence and alcohol dependence and abuse. Tables were created that look like Table 10.2 for nicotine dependence and alcohol dependence and abuse, and the cases off the main diagonal of the table (i.e., cells (1,2) and (2,1)) were examined.

Table 10.2 Impact of Editing and Imputation on Domain for Dependence and Abuse

		Was the respondent asked the dependence/abuse questions?	
		Yes	No
Is the respondent in the imputation-revised domain for dependence/abuse?	**Yes**	(1,1)	(1,2)
	No	(2,1)	(2,2)

For nicotine dependence, the imputation-revised domain included only those who used cigarettes in the past month according to the imputation-revised variable IRCIGRC. The conclusions for nicotine dependence are listed below:

- No respondents appeared in cell (1,2). This is because the editing procedures never edit a respondent out of past month use of cigarettes. If the respondent is a past month

user according to the skip logic imposed by the computer-assisted interviewing (CAI) instrument, then the respondent remains a past month user throughout the editing and imputation procedures.
- A few (fewer than 20) respondents appeared in cell (2,1). These cases were either edited or imputed to past month use of cigarettes.

For alcohol dependence and abuse, the imputation-revised domain included only those who had used alcohol in the past 12 months according to the imputation-revised variable IRALCRC and those who had used alcohol on at least 6 days in the past 12 months according to the imputation-revised variable IRALCFY. The conclusions for alcohol dependence and abuse are listed below:

- About 200 respondents appeared in cell (1,2). All of these cases had missing values for recency, 12-month frequency, or both and were edited or imputed into the domain.
- About 300 respondents appeared in cell (2,1). Most of these respondents were edited or imputed out of the domain.[55] Using the current procedures to create the substance dependence and abuse indicator-level variables, these respondents have a chance to be classified as having substance dependence or abuse, even though they are not in the imputation-revised domain, because the current procedures do not account for the recency and frequency; they only account for the responses to the dependence and abuse questions. If the imputation-revised domain was used to determine substance dependence or abuse, then responses for these respondents would not be used in analyses of substance dependence and abuse.

This section presents only results from the investigation of inconsistencies for nicotine dependence and alcohol dependence and abuse. These examples helped to gain an understanding of the issues associated with defining the domain for all substance dependence and abuse variables. However, for each substance, a thorough review of the questionnaire skip patterns and variable editing and imputation procedures was completed to determine the domain for each variable and to allow for calculation of item nonresponse rates and patterns presented in the next section. For the other drugs, the review was simpler than for alcohol because the 12-month frequency was not involved in the determination of the imputation-revised domain. For these drugs, the proportion of cases in cells (1,2) and (2,1) tended to be lower than for alcohol.

10.3 Item Nonresponse Rates for Substance Dependence and Abuse Variables

Using the 2011 NSDUH, item nonresponse rates were computed for each item-level variable used to create the criterion-level variables used for determining the substance dependence and abuse indicator-level variables. Tables K.1 through K.21 in Appendix K show weighted and unweighted item nonresponse rates for each of the variables used to derive the criterion-level and ultimately the indicator-level variables for substance dependence and abuse listed in the section above. Each table distinguishes the following different response categories:

[55] The skip logic implemented by the CAI instrument is such that those respondents reporting past year use of alcohol who do not respond to the 12-month frequency questions are presented with the alcohol dependence and abuse questions.

- never used or not used in past month or year,
- used in past month or year,
- not missing,
- logically assigned, and
- missing (don't know, refused, blank).

The item nonresponse rates are computed by using the imputation-revised domain (as defined in Section 10.2). The following discussion of Table K.1 is presented to assist in the interpretation of the data reported in the Appendix K tables. In row 1, among the 70,109 respondents to the 2011 NSDUH, 54,132 never used nicotine or had not used nicotine in the past month and 15,977 had used nicotine in the past month. Of the 15,977 past month users of nicotine, 15,908 responded to the question regarding "needing to smoke to feel less irritable." Among past month nicotine users, two individuals were logically assigned to past month recency and 67 respondents did not provide an answer to this question. The unweighted item response rate is 0.42 computed as the number of missing cases (67) divided by the number of total past month users (15,908 + 67 = 15,975) number of past month users (excluding those logically assigned) multiplied by 100. Similarly, the weighted item nonresponse rate of 0.58 is computed as the number of missing values (excluding the number of logically assigned) divided by the number of past month users using the final analytic weight.

Tables K.2 through K.21 present the item nonresponse for the remaining substances. For three substances—cocaine, heroin, and stimulants—the domain contained some uncertain cases (because one or more of the variables that determines the domain does not undergo imputation). These uncertain cases are discussed in Section 10.2, and the footnotes in Tables K.4, K.6, and K.10 present the definitions for these uncertain domain cases.

Tables 10.3 and 10.4 present a summary of the item nonresponse rates for all substance dependence and abuse variables. As shown in Table 10.3, of the 116 item-level substance dependence variables, 35 (or 30 percent) of the variables have weighted item nonresponse rates of less than one percent, 62 (or 54 percent) of the variables have rates between 1 and 5 percent and 19 (or 16 percent) variables have item nonresponse rates greater than 5 percent. The substances that have variables with item nonresponse rates greater than 5 percent include inhalants and sedatives. The sample size for these substances are small (n = 1,125 for inhalants and n = 223 for sedatives). Table 10.4 summarizes the item nonresponse rates for the item-level substance abuse variables. Similar to the dependence item-level variables, most had weighted item nonresponse rates less than 5 percent (40 out of 50 or 80 percent).

Table 10.3 Weighted Item Nonresponse Rates for Item-Level Substance Dependence Variables, 2011 NSDUH

Substance	Number of Item-Level Dependence Variables	Item Nonresponse Rates		
		Less than 1 Percent	1 to 5 Percent	More than 5 Percent
All Substances	116	35	62	19
Nicotine	18	17	1	0
Alcohol	10	10	0	0
Marijuana	9	0	9	0
Cocaine	11	0	11	0
Heroin	10	8	2	0
Hallucinogens	9	0	9	0
Inhalants	9	0	0	9
Pain Relievers	10	0	10	0
Tranquilizers	9	0	9	0
Stimulants	11	0	11	0
Sedatives	10	0	0	10

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2011.

Table 10.4 Weighted Item Nonresponse Rates for Item-Level Substance Abuse Variables, 2011 NSDUH

Substance	Number of Item-Level Abuse Variables	Item Nonresponse Rates		
		Less than 1 Percent	1 to 5 Percent	More than 5 Percent
All Substances	50	6	34	10
Alcohol	5	5	0	0
Marijuana	5	0	5	0
Cocaine	5	0	5	0
Heroin	5	1	4	0
Hallucinogens	5	0	5	0
Inhalants	5	0	0	5
Pain Relievers	5	0	5	0
Tranquilizers	5+	0	5	0
Stimulants	5	0	5	0
Sedatives	5	0	0	5

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2011.

The item nonresponse rates for these item-level substance dependence and abuse variables are similar to other NSDUH variables for which imputation is performed. The 2011 imputation report of the NSDUH methodological resource book (Frechtel et al., 2013) provides a summary of the item nonresponse rates for the variables that are imputed on the NSDUH main study.

10.4 Item Nonresponse Patterns for Substance Dependence and Abuse

This section describes the missing data patterns for the criterion-level substance dependence and abuse variables described in the prior section. Some patterns of missing data

may not have any effect on the final value for substance dependence or abuse because the dependence or abuse status can be logically inferred from the responses to the nonmissing items. Tables K.22 through K.42 in Appendix K present the item nonresponse patterns for the criterion-level substance dependence and abuse variables. Each of the tables shows patterns where dependence (Tables K.22 through K.32) or abuse (Tables K.33 through K.42) could be affected based on missing data. The patterns presented in each table are based on the criteria used to create the substance dependence or abuse variable and include the following:

- Missing no criteria: respondents with no missing data.
- Dependence or abuse regardless of missing data: respondents coded as having dependence or abuse regardless of any missing data. For example, if a respondent reported positive responses to three of the six marijuana dependence criteria and left the other three missing, he or she is defined as having marijuana dependence despite the missing data.
- No dependence or abuse regardless of missing data: respondents coded as having no dependence or abuse regardless of any missing data. For example, if a respondent reported negative responses to four of the six marijuana dependence criteria and left the other two missing, he or she is defined as not having marijuana dependence despite the missing data, because at least three positive responses are required.
- Item nonresponse patterns that affect dependence or abuse: respondents where missing data would affect substance dependence or abuse. For example, marijuana dependence is unknown for respondents with missing values for all six criteria.

In Appendix K, Table K.22 shows the item nonresponse patterns based on the nicotine dependence syndrome scale score (NDSSANSP) and the Fagerstrom Test of Nicotine Dependence scale score (FTNNDNSP) among past month users because this dependence is defined for respondents who reported smoking cigarettes in the past month. The item-level variable NDSSANSP is calculated as the average score over 17 item-level variables pertaining to five aspects of dependence. The item-level variable FTNNDNSP records whether past month users smoke within 30 minutes after waking up. Nicotine dependence status is true if NDSSANSP average score is greater than or equal to 2.75 or the FTNNDNSP variable is true. Among those past month users of nicotine, 96.2 percent (15,203 cases) had no missing data for the 18 item-level variables used to create nicotine dependence.

For the respondents missing *only* one of the 17 questions for NDSSANSP (151 cases), imputation is used to fill in the values for the missing variable. There were 47 cases (0.36 percent) where dependence could be determined regardless of missing data (27 cases being identified as having dependence and 20 cases not having dependence). These three categories account for 1.3 percent of past month users whose nicotine dependence status can be determined using current imputation procedures regardless of missing data for all past month users. The rows in Table K.22 (in table section described as "Item Nonresponse Patterns that Affect Determination of Dependence Status") labeled "Missing in FTND measure and 'No' in NDSS measure" and "Missing *x* variables in NDSS measure" show 576 cases (2.7 percent) where dependence status is unknown due to missing data. Most of these 576 cases (466) were missing the FTND measure or the variable CIGWAKE, which is not imputed and has an item nonresponse rate of 2.6 percent as shown in Table K.1. Table K.22 also shows the distribution of the number of item-level variables missing in the computation of the NDSS measure (ranging

from missing 2 to 17 variables). Of the 27 cases missing all 17 item-level variables, 14 cases were imputed into the domain because the respondent was not asked the substance dependence and abuse questions (as described in Table 10.2), and thus these variables would have missing values logically assigned. The 2.7 percent of cases missing data can be interpreted as the item nonresponse rate for the nicotine dependence indicator-level variable because it combines all of the variables used in its computation.

In Appendix K, Table K.23 shows the item nonresponse patterns based on the criterion-level variables used to create alcohol dependence.

- The row labeled "Used in the past year and used on at least 6 days in past year" describe the imputation-revised domain for alcohol dependence.
- The row labeled "Missing none of the 7 criteria" describes those respondents with no missing data (i.e., none of the 10 item-level variables used to derive the seven criterion-level variables have any missing data). For example, the first criterion-level variable for alcohol is based on ALCLOTTM = 1 or ALCGTOVR = 1, and then the number of missing values for the combinations of ALCOTMM or ALCGTOVR was computed. Missing data is defined as values other than 1 (Yes), 2 (No), or 99 (Legitimate Skip) for all item-level variables (ALCLOTTM, ALCGTOVR, ALCKPLMT, ALCNDMOR, ALCLSEFX, ALCCUTEV, ALCMCTD, ALCPHCTD, ALCLSACT, and ALCWDSMT).
- The row labeled "Dependence regardless of missing data" describes those respondents who meet the criteria for alcohol dependence (positive responses for three or more of the seven criterion-level variables) regardless of missing any of the 10 variables used to compute the alcohol dependence indicator-level variable.
- The row labeled "No dependence regardless of missing data" describes those respondents who do not meet the criteria (the number of positive criterion-level variables plus the number of missing criterion-level variables is less than three) and where imputation would not change their alcohol dependence.
- The row labeled "Item Nonresponse Patterns that Affect Determination of Dependence Status" describes the respondents for whom the dependence status could not be determined due to missing data. The rows labeled "Number of criteria true" and "Number of criteria missing" show the different possible combinations of true criteria (i.e., the dependence criteria being met) and missing data (summarized in the previous row) that would affect their dependence status.

Among the 33,703 past year alcohol users, 98.3 percent answered all questions for seven criterion-level variables. Alcohol dependence status can be determined for 1.1 percent of users, even though responses were missing on some item-level variables. Only 0.7 percent of past year users (344 cases) had an uncertain alcohol dependence status. The main item nonresponse pattern is when none of the dependence criterion-level variables have positive responses but three or more criterion-level variables were missing (305 cases). The approximately 200 cases mentioned in Section 10.2 (where the respondent was not asked the substance dependence and abuse questions because he or she was edited or imputed into the domain) are a subset of the 305 cases where three or more criterion-level variables were missing. Similar to the interpretation of the nicotine dependence indicator-level variable, the item nonresponse rate for the alcohol dependence indicator-level variable is 0.65 percent.

Table 10.5 presents a summary of the item nonresponse patterns for criterion-level substance dependence variables, excluding nicotine dependence. As shown in Table 10.5, the highest percent of uncertain cases (7.9 percent) was in sedative users (22 of 223 past year users), followed by 6.7 percent in inhalant users (108 of 1,125 past year users). The lowest percent of missing (1.3 percent and 0.7 percent) were in marijuana and alcohol users (164 among 9,553 and 344 among 33,073). The percentages of missing dependence status in users of pain relievers, hallucinogens, stimulants, tranquilizers, cocaine, and heroin were 3.9 percent, 3.6 percent, 3.1 percent, 2.8 percent, 2.2 percent, and 1.7 percent, respectively. The average percent of cases that are affected by missing data across all substances is 3.4 percent.

Table 10.5 Percentage of Respondents Where Substance Dependence Status Is Affected by Missing Data, 2011 NSDUH

Substance	Number of Past Year Users	Item Nonresponse Patterns that Affect Dependence				
		Past Year Users with Missing Criteria that Affect Dependence		No Criterion-Level Variables Positive and Missing 3 or More Criterion-Level Variables	One Criterion-Level Variable Positive and Missing 2 or More Criterion-Level Variables	Two Criterion-Level Variables Positive and Missing 1 or More Criterion-Level Variables
		Number	Weighted Percentage			
Average			3.39	3.15	0.17	0.07
Sedatives	223	22	7.91	7.83	0.00	0.08
Inhalants	1,125	108	6.66	6.34	0.22	0.11
Pain Relievers	4,684	240	3.88	3.76	0.04	0.08
Hallucinogens	2,303	88	3.63	3.62	0.00	0.01
Stimulants	1,211	43	3.09	2.68	0.29	0.12
Tranquilizers	1,915	66	2.82	2.79	0.03	0.00
Cocaine	1,581	47	2.17	2.15	0.02	0.00
Heroin	267	7	1.73	0.68	1.05	0.00
Marijuana	9,553	164	1.34	1.15	0.02	0.17
Alcohol	33,073	344	0.65	0.53	0.03	0.10

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2011.

Tables K.22 through K.32 also present the number and percentage of cases not being affected by missing data. The number of cases not affected by missing data range from 1 case (0.21 percent) for heroin to 317 cases (1.07 percent) for alcohol. Excluding these two substances, the majority of cases not affected by missing data average less than 1 percent (or 30 cases).

In Appendix K, Table K.33 shows the item nonresponse patterns based on the criterion-level variables used to create alcohol abuse.

- The row labeled "Missing none of the 4 criteria" describes those respondents with no missing data for the four criterion-level variables. Missing data is defined as values other than 1 (Yes), 2 (No), and 99 (Legitimate Skip) for all item-level variables (ALCSERPB, ALCPDANG, ALCLAWTR, ALCFMFPB, and ALCFMCTD).
- The row labeled "Abuse regardless of missing data" describes those respondents who meet the criteria for alcohol abuse (positive responses for one or more of the four

criterion-level variables) regardless of missing any of the five item-level variables used to compute the alcohol abuse indicator-level variable.

- The row labeled "No abuse regardless of missing data" describes those respondents who do not meet the criteria (negative response for one or more of the four criterion-level variables) and where imputation would not change their alcohol abuse value.
- The row labeled "One or more criteria missing when no criteria true" describes those where missing data would affect their alcohol abuse value.

As shown in Table K.33, there were 99.6 percent alcohol users who answered all questions related to four criteria in the past year alcohol users (33,703). Missing data would not affect alcohol abuse status for only ten cases (three for abuse and seven for no abuse). Less than 1 percent (0.42 percent) of past year users (298) had missing data whose alcohol abuse status was affected by missing data.

Table 10.6 presents a summary of the item nonresponse patterns for criterion-level substance abuse variables. Similar to dependence status, the highest percentages of uncertain cases (6.7 percent and 6.5 percent) were in inhalant users (1,125) and sedative users (223). The lowest percentages of missing cases (0.4 percent and 1.3 percent) were in alcohol and marijuana users (33,073 and 9,553). The percentages of missing abuse status in pain reliever, hallucinogen, stimulants, tranquilizer, cocaine, and heroin users were 3.9 percent, 3.6 percent, 3.0 percent, 2.9 percent, 2.4 percent, and 1.9 percent, respectively. Unlike dependence status, the number of cases whose abuse status can be determined regardless of missing data is very small (10 or fewer cases) for all substances.

Table 10.6 Percentage of Respondents Where Substance Abuse Would Be Affected by Missing Data, 2011 NSDUH

Substance	Number of Past Year Users	Past Year Users with Missing Criterion-Level Variables that Affect Abuse	
		Number	Weighted Percentage
Average			3.26
Inhalants	1,125	107	6.69
Sedatives	223	20	6.49
Pain Relievers	4,684	233	3.88
Hallucinogens	2,303	88	3.62
Stimulants	1,211	39	3.00
Tranquilizers	1,915	65	2.93
Cocaine	1,581	48	2.35
Heroin	267	7	1.94
Marijuana	9,553	153	1.25
Alcohol	33,073	298	0.42

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2011.

10.5 Evaluating the Need for Imputation of the Substance Dependence and Abuse Variables

To determine how missing data affects the indicator-level substance dependence and abuse variables, a sensitivity analysis was performed to assess the range of possible values for

the dependence and abuse variables after imputation. The following methods were compared in the analysis:

- *Complete Case (CC) Analysis:* Only the complete cases were used to create the indicator-level substance dependence or abuse variables. Cases with known dependence or abuse status were considered complete even if some item-level variables or criterion-level variables were missing. For example, the proportion of dependence was the weighted number of respondents known to be both (1) in the domain and (2) dependent based on responses to the dependence questions, divided by the weighted number of respondents whose dependence status is known based on both the domain and the dependence criteria.
- *Impute to Zero (I_0) Method:* Each respondent with a missing value for the dependence and/or abuse indicator was assigned a value of zero, thus making the overall estimate as low as possible. The numerator was the same as with the CC analysis, but the denominator was the weighted number of respondents in the entire NSDUH sample.
- *Impute to One (I_1) Method:* Each respondent with a missing value for the dependence and/or abuse indicator was assigned a value of one, thus making the overall estimate as high as possible. The denominator was the same as with the I_0 method, but the numerator was the weighted number of respondents who (1) either were known to be in the domain or could have been imputed into the domain, and (2) either were known to be dependent or could have been imputed to be dependent.
- *Current Method (CM):* When an item-level variable is used to compute a dependence or abuse criterion-level variable, missing values are considered as not meeting the criteria for dependence or abuse (i.e., they are treated the same as a "No" response). This current method is different than the I_0 method because it ignores domain issues and examines the missingness at the item level.

The I_0 method represents the lower endpoint of the prevalence of dependence or abuse, and the I_1 method represents the upper endpoint. If there is not much difference between the I_0 and I_1 methods, then that would suggest a simple imputation method would be sufficient or that imputation may not be needed. The complete case method could not possibly be biased by very much, and the negative bias associated with the current method would be minor.

10.5.1 Implementation Procedures of Methods by Substance

To implement the CC, I_0, and I_1 methods, the cases with missing dependence or abuse indicator-level variables were first identified. Under CC, these cases were dropped from the analysis. Under I_0, these missing cases were assigned as not having dependence or abuse. Under I_1, these missing cases were assigned as having dependence or abuse. Cases with uncertain domain status were set to missing for the dependence and abuse indicator-level variables for both I_0, and I_1 methods. Only three substances—cocaine, heroin, and stimulants—have cases with uncertain imputation-revised domains because one or more of the variables that determines the domain does not undergo imputation. For example, the imputation-revised domain for cocaine dependence includes four cases with unknown values. These cases have a recency value of not used in past year (IRCOCRC = 3) and time since last used needle to inject cocaine (CONDLREC) values of refused (97) or blank (98) since CONDLREC is not imputed.

Once cases with uncertain domain status were set to missing for the indicator-level dependence and abuse variables, cases that were definitely in the domain were examined further for missingness with respect to dependence and abuse. The method by which cases definitely in the domain are identified as missing is described below.

10.5.1.1 Nicotine Dependence

The criterion-level variable NDSSDNSP is created using the mean of the responses to 17 item-level variables for nicotine dependence, and the criterion-level variable FTNDDNSP is created using the item-level variable CIGWAKE. The indicator-level nicotine dependence variable, DNICNSP, is set to 1 (yes) if either NDSSDNSP = 1 (yes) or FTNDDNSP = 1 (yes). Missing cases are shown in italics in Table 10.7.

Table 10.7 Value of Indicator-Level Nicotine Dependence (DNICNSP) as Derived from Criterion-Level Variables NDSSDNSP and FTNDDNSP

		Criterion-Level Variable FTNDDNSP		
		Yes	**No**	**Missing**
Criterion-Level Variable NDSSDNSP	**Yes**	Yes	Yes	Yes
	No	Yes	No	*Missing*
	Missing	Yes	*Missing*	*Missing*

For the purposes of this methods study, NDSSNSP is defined as missing if $46.75 - 5m < S < 46.75 - m$, for $m > 2$, where m is the number of missing responses (out of 17 item-level variables) and S is the sum of the $17 - m$ nonmissing responses. This inequality can be derived from the following:

- NDSSNSP is 1 if and only if the mean of the 17 item-level variables is 2.75 or higher. Equivalently, NDSSNSP is 1 if and only if the sum of the 17 item-level variables is 46.75 or higher.
- Each of the 17 item-level variables range from 1 (least dependent) to 5 (most dependent). Therefore, $5m$ represents the highest value and m represents the lowest value the missing items could have contributed to the sum of the 17 item-level variables.
- If only one of the 17 item-level variables is missing, the variable with missing values is filled in as part of standard processing (Frechtel et al., 2013, Chapter 6).

Respondents whose nonmissing variables sum to less than $46.75 - 5m$ are defined as not having dependence according to the NDSS, regardless of imputation results. Respondents whose nonmissing variables sum to at least $46.75 - m$ are defined as having dependence according to the NDSS, regardless of imputation results. The missingness of FTNDNSP is easily assessed because it is based on the item-level variable CIGWAKE. The missingness of DNICNSP is computed using the combination of cells from Table 10.7.

10.5.1.2 Dependence for Alcohol, Cocaine, Heroin, Pain Relievers, Stimulants, and Sedatives

The indicator-level dependence variables for these six substances are created using seven criterion-level variables. If a respondent meets at least three of the criteria, they are defined as having dependence. It follows that the dependence status of respondents cannot be established (i.e., is missing) if $3 - m \le y < 3$, where y is the number of criteria that are met and m is the number of missing criteria. Respondents with $y \ge 3$ are defined as having dependence regardless of imputation, and respondents with $y + m < 3$ are defined as not having dependence regardless of imputation. The indicator-level dependence variables depend on imputation for all other respondents (i.e., those with $y < 4$ and $y \ge 4 - m$).

10.5.1.3 Dependence for Marijuana, Hallucinogens, Inhalants, and Tranquilizers

The indicator-level dependence variables for these four substances are created using six criterion-level variables instead of seven, but the principles are still the same. If a respondent meets at least three of the criteria, they are defined as having dependence. The dependence status of respondents that cannot be established is $3 - m \le y < 3$, just as it is for the substances discussed in the preceding section.

10.5.1.4 Abuse for All Substances

To be defined as having abuse, a respondent must not have dependence and must meet at least one of four criteria. In the absence of dependence, respondents who met one or more criteria were defined as having abuse; respondents who met none of the criteria were defined as not having abuse; and all other respondents in the imputation-revised domain (i.e., those who had at least one missing criteria and did not meet any of the nonmissing criteria) were classified as missing for abuse of that particular substance.

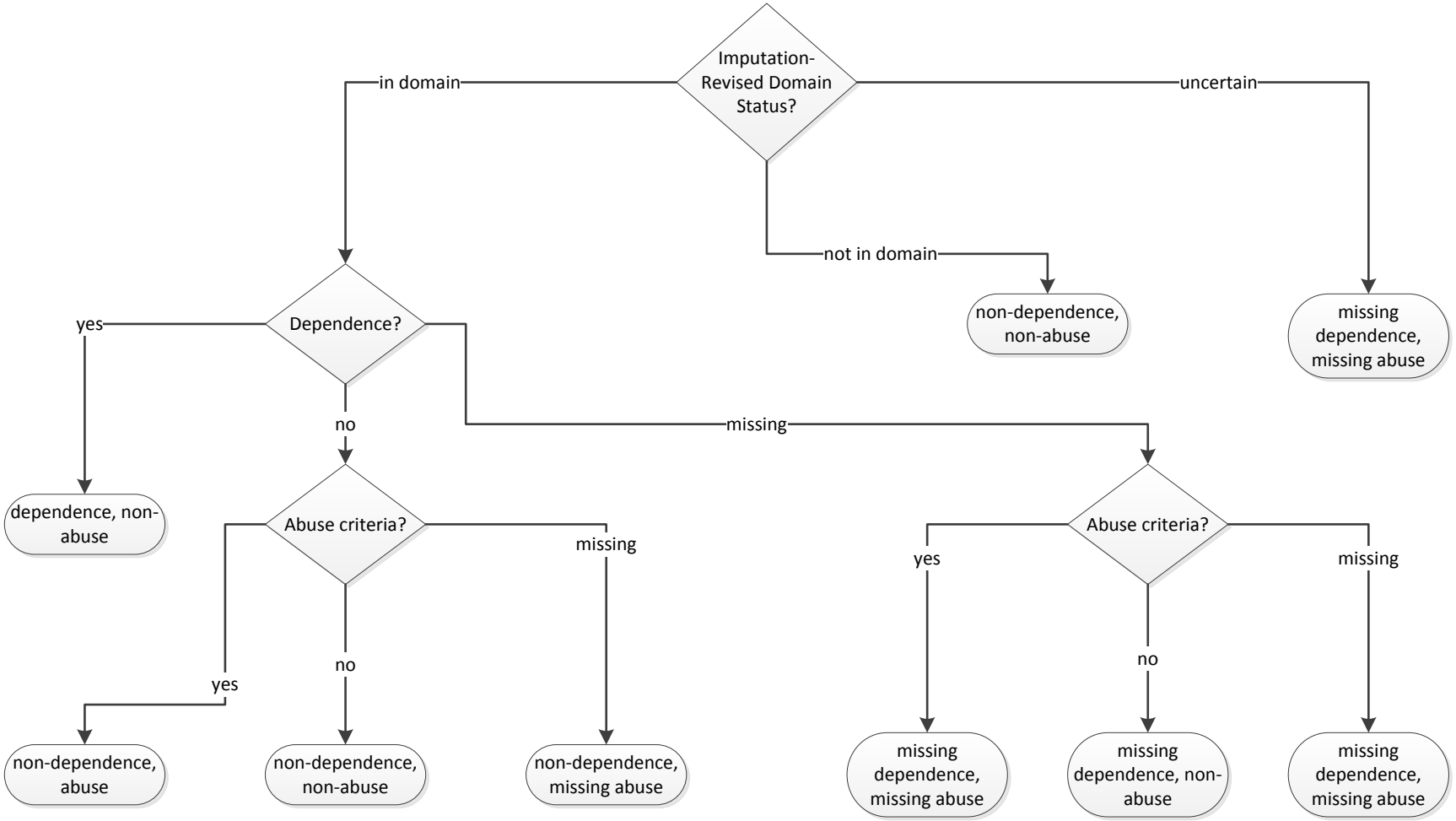
When dependence is considered, matters become more complicated. Those respondents defined as having dependence are automatically defined as not having abuse regardless of the values of the four criterion-level abuse variables. Those who are missing dependence status and meet at least one of the abuse criteria are still classified as missing for abuse status because their classification for abuse depends on the dependence status. Missing cases for abuse are shown in italics in Table 10.8.

Table 10.8 Value of Abuse as Derived from Dependence and the Abuse Criteria

		Dependence		
		Yes	**No**	**Missing**
Abuse Criteria	**Met at least one**	No	Yes	*Missing*
	Met none, and none were missing	No	No	No
	Met none, and at least one was missing	No	*Missing*	*Missing*

A decision tree illustrating the creation of the dependence and abuse indicator-level variables for the purposes of this methods study is provided in Figure 10.1.

Figure 10.1 Decision Tree for Creating Indicator-Level Substance Dependence and Abuse Variables



10.5.2 Sensitivity Analysis Results

The sensitivity analysis results are presented in Tables 10.9 and 10.10 for dependence and abuse respectively. For each of the four methods, CC, I_0, I_1, and CM, the point estimate and the 95 percent confidence interval are shown. The confidence interval represents the possible range for the substance dependence and abuse prevalence estimates. If the I_1 point estimate is outside the I_0 or CM confidence interval, then imputation would be recommended because this would imply that imputation could move the estimate out of the range that is considered reasonable in the national estimates. Of the 11 dependence measures, the I_1 point estimate is not in I_0 or CM confidence intervals for eight substances. Of the 10 abuse measures, the I_1 point estimate is not in I_0 or CM confidence intervals for eight substances. These measures appear in boldface in the tables. The noteworthy results include the following:

- For all substances except alcohol and marijuana, the dependence estimates for CM and I_0 are exactly the same. Under CM, all respondents who were defined as having dependence based on the responses to the dependence questions were defined as having dependence, and all other respondents were defined as not having dependence. Under I_0, no respondents were edited or imputed out of the domain, and all domain members who were missing for dependence were imputed as not having dependence. The key point is that no one was edited or imputed out of the domain. The editing rules are such that no past year users are ever edited out of past year use, and any respondents whose recency was missing never saw the dependence questions (and therefore were imputed as not having dependence).
- For alcohol and marijuana dependence, I_0 is slightly less than CM. A few respondents were defined as having dependence based on the responses to the dependence questions, but their frequency was imputed to be less than 6. Under I_0, these respondents were not part of the domain and were defined as not having dependence.
- For nicotine dependence, I_0 is less than CM because of the differential handling of a single respondent. (This is not noticeable out to the two decimal places displayed in the tables.) Under CM, respondents missing 2 or more of the 17 item-level variables are automatically defined as not having dependence. Under I_0, I_1, and CC, respondents whose nonmissing responses are such that they would be defined as having dependence regardless of imputation were defined as having dependence. For example, a respondent whose nonmissing item-level variables sum to more than 46.75 should be defined as having dependence, even if several responses are missing.
- For all substances except nicotine, the dependence results for CC are similar to the results for I_0 and CM. In the numerator, CC is equivalent to I_0. In the denominator, CC is smaller than I_0, but for most of these substances, the vast majority of respondents are not in the imputation-revised domain. All of these respondents who are not in the imputation-revised domain are treated the same way under CC and I_0 (with the exception of the alcohol and marijuana cases imputed out of the domain).
- For nicotine dependence, CC is noticeably larger than I_0 and CM. As said in the previous bullet, for the other dependence measures, the vast majority of respondents are not in the imputation-revised domain. But for nicotine dependence, the denominator for the percentages shown in Table 10.9 includes only past month users

of cigarettes (this is consistent with what is shown in the standard set of NSDUH detailed tables). The impact is more noticeable because it is not dampened by the vast majority of respondents who are not in the imputation-revised domain.

- For abuse, the results for CC and I_0 are occasionally quite a bit smaller than the results for CM. This is likely due to the differential handling of respondents whose dependence was missing and who met at least one of the abuse criteria. Under CM, these respondents were defined as not having dependence and therefore were assigned to abuse. Under I_0, these respondents were defined as not having abuse.
- For many of the substances, the dependence questions have low nonresponse, causing the I_1 estimate to be not substantially greater than the I_0 estimate. Also, for some of the rare substances, the vast majority of cases are not in the domain, and these cases are defined as not having abuse and as not having dependence for both I_0 and I_1.

Table 10.9 Imputation Results for Substance Dependence, 2011 NSDUH

Substance	Method	Estimate	Lower Confidence Limit	Upper Confidence Limit
Nicotine	CM	56.5	55.1	58.0
	CC	58.1	56.6	59.5
	I_0	56.5	55.1	58.0
	I_1	59.2	57.8	60.6
Alcohol	CM	3.0	2.8	3.2
	CC	3.0	2.8	3.2
	I_0	3.0	2.8	3.2
	I_1	3.4	3.2	3.6
Marijuana	CM	1.0	0.9	1.1
	CC	1.0	0.9	1.1
	I_0	1.0	0.9	1.1
	I_1	1.1	1.0	1.2
Cocaine	CM	0.2	0.2	0.3
	CC	0.2	0.2	0.3
	I_0	0.2	0.2	0.3
	I_1	0.3	0.2	0.3
Heroin	CM	0.1	0.1	0.2
	CC	0.1	0.1	0.2
	I_0	0.1	0.1	0.2
	I_1	0.1	0.1	0.2
Hallucinogens	CM	0.1	0.0	0.1
	CC	0.1	0.0	0.1
	I_0	0.1	0.0	0.1
	I_1	0.1	0.1	0.1
Inhalants	CM	0.0	0.0	0.0
	CC	0.0	0.0	0.0
	I_0	0.0	0.0	0.0
	I_1	0.1	0.1	0.1
Pain Relievers	CM	0.5	0.5	0.6
	CC	0.5	0.5	0.6
	I0	0.5	0.5	0.6
	I1	0.7	0.6	0.8
Tranquilizers	CM	0.1	0.0	0.1
	CC	0.1	0.0	0.1
	I0	0.1	0.0	0.1
	I1	0.1	0.1	0.2

Table 10.9 Imputation Results for Substance Dependence, 2011 NSDUH (continued)

Substance	Method	Estimate	Lower Confidence Limit	Upper Confidence Limit
Stimulants	CM	0.1	0.1	0.1
	CC	0.1	0.1	0.1
	I_0	0.1	0.1	0.1
	I_1	0.2	0.1	0.2
Sedatives	CM	0.0	0.0	0.1
	CC	0.0	0.0	0.1
	I_0	0.0	0.0	0.1
	I_1	0.0	0.0	0.1

CC = complete case; CM = current method; I_0 = impute to zero; I_1 = impute to one.

Note: Substances in which the I_1 point estimate is greater than the upper confidence limits for I_0 and CM are shown in boldface.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2011.

Table 10.10 Imputation Results for Substance Abuse, 2011 NSDUH

Substance	Method	Estimate	Lower Confidence Limit	Upper Confidence Limit
Alcohol	CM	3.5	3.2	3.7
	CC	3.4	3.2	3.7
	I_0	3.4	3.2	3.7
	I_1	3.7	3.5	3.9
Marijuana	CM	0.6	0.5	0.7
	CC	0.6	0.5	0.6
	I_0	0.6	0.5	0.6
	I_1	0.7	0.6	0.8
Cocaine	CM	0.1	0.1	0.1
	CC	0.1	0.1	0.1
	I_0	0.1	0.1	0.1
	I_1	0.1	0.1	0.2
Heroin	CM	0.0	0.0	0.0
	CC	0.0	0.0	0.0
	I_0	0.0	0.0	0.0
	I_1	0.0	0.0	0.0
Hallucinogens	CM	0.1	0.0	0.1
	CC	0.1	0.0	0.1
	I_0	0.1	0.0	0.1
	I_1	0.1	0.1	0.2
Inhalants	CM	0.0	0.0	0.0
	CC	0.0	0.0	0.0
	I_0	0.0	0.0	0.0
	I_1	0.1	0.1	0.1
Pain Relievers	CM	0.2	0.1	0.2
	CC	0.2	0.1	0.2
	I0	0.2	0.1	0.2
	I1	0.3	0.3	0.4
Tranquilizers	CM	0.1	0.1	0.1
	CC	0.1	0.1	0.1
	I0	0.1	0.1	0.1
	I1	0.1	0.1	0.2

Table 10.10 Imputation Results for Substance Abuse, 2011 NSDUH (continued)

Substance	Method	Estimate	Lower Confidence Limit	Upper Confidence Limit
Stimulants	CM	0.0	0.0	0.0
	CC	0.0	0.0	0.0
	I_0	0.0	0.0	0.0
	I_1	0.1	0.0	0.1
Sedatives	CM	0.0	0.0	0.0
	CC	0.0	0.0	0.0
	I_0	0.0	0.0	0.0
	I_1	0.0	0.0	0.0

CC = complete case; CM = current method; I_0 = impute to zero; I_1 = impute to one.

Note: Substances in which the I_1 point estimate is greater than the upper confidence limits for I_0 and CM are shown in boldface.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2011.

10.6 Alternative Imputation Method

The current method for computing the substance dependence and abuse estimates is a conservative approach because respondents with missing values are considered as not meeting the criteria for dependence or abuse, which potentially undercounts those who meet the criteria for substance dependence or abuse. The results from the sensitivity analysis suggest that an imputation method should be tested because the substance dependence and abuse point estimates for several substances showed substantial differences between the I_0 and I_1 methods and the I_1 estimates were outside the range of the CM and I_0 confidence intervals.

To determine whether the current practice of effectively categorizing item nonrespondents as "no" responses during the determination of dependence or abuse status, a weighted sequential hot-deck (WSHD) imputation method (as described in Chapter 3 of this report) was implemented on the 2011 NSDUH data. The imputation classes were defined by using a nonparametric classification tree analysis (Breiman, Friedman, Olshen, & Stone, 1984) similar to Chi-square Automatic Interaction Detection (CHAID) analysis (Section 3.1.1).

A cyclic approach for the WSHD imputation was implemented similar to the cyclical methods discussed in Chapter 4 for IVEware and Chapter 5 for modPMN-MI method. The cyclic hot-deck imputation is discussed in Marker, Judkins, and Winglee (2002). The first cycle uses the complete responses and any previously imputed variables to develop imputation classes and perform imputations. For each subsequent imputation cycle, all variables on the dataset are available for the tree-based methodology to create imputation classes. Cycling is advantageous because it determines the best relationships among variables and minimizes the relationships among variables that may be caused by the sequential nature of the imputation process. For item-level substance dependence and abuse variables, only two cycles were performed to allow each of the item-level variables to be used as a predictor variable.

The item-level substance dependence and abuse variables shown in Tables K.2 through K.21 in Appendix K were imputed. Additional item-level variables for three substances (cocaine, heroin, and stimulants) that were needed to determine the domain were also imputed. These additional variables are footnoted in the respective tables shown in Appendix K. The set of

predictor variables for the CHAID analysis included the following: age, gender, recency of use, frequency of use, and age of initiation (or first use). Tables K.43 and K.44 show the imputation classes that resulted from the CHAID analysis for pain relievers and stimulants.

After all item-level substance dependence and abuse variables were imputed, the criterion-level and indicator-level variables were created. Table 10.11 shows the substance dependence, abuse, and disorder estimates based on WSHD imputed data and CM. The differences are hardly noticeable when percentages are rounded to one decimal place. One difference is for illicit drug disorder, where the CM estimate is 2.5 percent as compared with the WSHD estimate of 2.6 percent.

Table 10.11 Comparison of Substance Dependence, Abuse, and Disorder Estimates Based on Current Method and WSHD Imputation, 2011 NSDUH

	Dependence		Abuse		Disorder	
Substance	**CM (95% CI)**	**WSHD (95% CI)**	**CM (95% CI)**	**WSHD (95% CI)**	**CM (95% CI)**	**WSHD (95% CI)**
Nicotine	56.5 (55.1, 58.0)	56.9 (55.5, 58.4)	N/A	N/A	N/A	N/A
Alcohol	3.0 (2.8, 3.2)	3.0 (2.8, 3.2)	3.5 (3.2, 3.7)	3.5 (3.2, 3.7)	6.5 (6.2, 6.8)	6.5 (6.2, 6.8)
Marijuana	1.0 (0.9, 1.1)	1.0 (0.9, 1.1)	0.6 (0.5, 0.7)	0.6 (0.5, 0.7)	1.6 (1.5, 1.7)	1.6 (1.5, 1.7)
Cocaine	0.2 (0.2, 0.3)	0.2 (0.2, 0.3)	0.1 (0.1, 0.1)	0.1 (0.1, 0.1)	0.3 (0.3, 0.4)	0.3 (0.3, 0.4)
Heroin	0.1 (0.1, 0.2)	0.1 (0.1, 0.2)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.2 (0.1, 0.2)	0.2 (0.1, 0.2)
Hallucinogens	0.1 (0.0, 0.1)	0.1 (0.0, 0.1)	0.1 (0.0, 0.1)	0.1 (0.1, 0.1)	0.1 (0.1, 0.2)	0.1 (0.1, 0.2)
Inhalants	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.1 (0.0, 0.1)	0.1 (0.0, 0.1)
Pain Relievers	0.5 (0.5, 0.6)	0.5 (0.5, 0.6)	0.2 (0.1, 0.2)	0.2 (0.1, 0.2)	0.7 (0.6, 0.8)	0.7 (0.6, 0.8)
Tranquilizers	0.1 (0.0, 0.1)	0.1 (0.0, 0.1)	0.1 (0.1, 0.1)	0.1 (0.1, 0.1)	0.2 (0.1, 0.2)	0.2 (0.1, 0.2)
Stimulants	0.1 (0.1, 0.1)	0.1 (0.1, 0.1)	0.0 (0.0, 0.0)	0.1 (0.0, 0.1)	0.1 (0.1, 0.2)	0.2 (0.1, 0.2)
Sedatives	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.1)	0.0 (0.0, 0.1)
Illicit Drug	1.8 (1.6, 1.9)	1.8 (1.7, 1.9)	0.8 (0.7, 0.8)	0.8 (0.7, 0.9)	2.5 (2.4, 2.7)	2.6 (2.5, 2.8)
Illicit Drug Other than Marijuana	0.9 (0.8, 1.0)	0.9 (0.8, 1.0)	0.3 (0.3, 0.4)	0.4 (0.3, 0.4)	1.2 (1.1, 1.4)	1.3 (1.2, 1.4)
Illicit Drug Excluding Marijuana	0.8 (0.7, 0.8)	0.8 (0.7, 0.9)	0.3 (0.3, 0.4)	0.4 (0.3, 0.4)	N/A	N/A
Psychotherapeutic	0.6 (0.5, 0.7)	0.7 (0.6, 0.7)	0.2 (0.2, 0.2)	0.2 (0.2, 0.3)	0.8 (0.7, 0.9)	0.9 (0.8, 1.0)
Illicit Drug or Alcohol	4.3 (4.1, 4.6)	4.4 (4.1, 4.6)	4.0 (3.8, 4.3)	4.1 (3.9, 4.3)	8.0 (7.7, 8.3)	8.1 (7.8, 8.4)
Illicit Drug and Alcohol	0.4 (0.4, 0.5)	0.4 (0.4, 0.5)	0.2 (0.2, 0.2)	0.2 (0.2, 0.2)	1.0 (0.9, 1.1)	1.0 (0.9, 1.1)

CI = confidence interval; CM = current method; N/A = not applicable; WSHD = weighted sequential hot deck.
Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.
Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2011.

10.7 Summary

The amount of missing data for the item-level substance dependence and abuse variables used to create the dichotomous indicator-level variables is low and similar to other item nonresponse rates for many other variables on the NSDUH for which imputation is performed. Of the 115 item-level variables used to compute substance dependence, 96 variables (83 percent) have item nonresponse rates of less than 5 percent. Of the 50 variables used to compute substance abuse, 40 variables (80 percent) have item nonresponse rates of less than 5 percent.

The item nonresponse patterns for criterion-level substance dependence and abuse variables indicate that differential treatment of missing data affects the computation of the dichotomous substance dependence or abuse variables. The percentage of substance dependence and abuse cases with missing item-level variables range from a low of 0.7 percent to a high of 7.9 percent, where an average 3 percent of cases could have their dependence or abuse values affected if imputation were performed.

The sensitivity analysis showed that the substance dependence and abuse estimates may be affected by imputation because the upper bound estimate (I_1) was not in I_0 or CM confidence intervals, indicating that the estimates for dependence and abuse may be underestimated if no imputation was performed. Although most of the dependence variables have high response rates, the differences between the estimate under I_0 and the estimate under I_1 are noticeable and any imputation method would correct at least some of the negative bias in the CM method, a method that is similar to I_0.

The substance dependence and abuse estimates from both the CM and WSHD methods are similar, where the differences are small (a high of 0.04 percent for dependence and 0.06 percent for abuse). For most substances, the individual substance dependence and abuse estimates from both the CM and WSHD methods when rounded to the nearest tenth are not different. Although the differences in the CM and WSHD estimates are not substantial when examined for each individual substance, the effects of the WSHD imputation are more visible at the disorder level because the increase is a function of the item nonresponse rates for many variables (at least eight or nine substances). When rounding to the nearest tenth, the WSHD estimates for illicit drug disorders would be at least 0.1 percent higher than the CM, indicating that the CM underestimates the illicit drug disorder estimates. Because illicit drug disorder is a key measure used in many NSDUH analytic studies, it is possible that the increase in disorder rates due to imputation may be considered a substantive improvement in the overall substance use disorder estimates.

This page intentionally left blank

11. Conclusions and Next Steps

This chapter presents conclusions and possible next steps that would lead to potential improvements to the imputation procedures for the National Survey on Drug Use and Health (NSDUH). The main goal of this report was to assess whether simpler, faster, or more cost-effective imputation methods could be applied to the NSDUH data and whether these methods would maintain, minimally reduce, or even improve data quality. This goal was difficult to quantify because there is no gold standard to compare quality against. Instead, this study focused on identifying any statistically significant and substantively meaningful differences in national estimates, cost differences in terms of staff hours, and compliance with post-survey data processing schedules between the current predictive mean neighborhood (PMN) method and the alternative methods explored. Because the comparisons were based on specified versions of each alternative method, it is possible different options or combinations may have resulted in slightly different results, but the goal was to be able to make inferences about the overall implementation of these different methods. This report also examined other aspects of the NSDUH imputation process that could potentially be improved such as using the respondent race and ethnicity screener data in interview race and ethnicity imputation, using data from a pair respondent to impute for item nonresponse for the other pair, and imputing certain mental health and substance dependence and abuse variables. The next steps outlined at the end of this chapter attempt to identify the advantages of the methods that were examined while keeping the need to be able to maintain trends and comparability over time.

11.1 Comparing the PMN Method with Alternative Methods: Summary and Conclusions

Overall across the different methods examined for substance use variables, there was not a great deal of difference based on the methods used. This may be, in part, due to the low missingness rates on the NSDUH. The item nonresponse rates for most variables that undergo imputation are typically small, averaging less than 5 percent. However, rarer drugs, such as cocaine, inhalants, heroin, and hallucinogens, have higher nonresponse rates because these variables have a smaller number of respondents. The results for these variables may be influenced by these smaller sample sizes. However, even though the overall estimates remained comparable, the simpler methods, with fewer consistency checks, result in more inconsistent data. The disadvantage of having inconsistent data is that analysts have to make decisions on how to address these inconsistencies during analysis. This leads not only to a greater effort during the analysis phase but also to more variability between analysts, depending on how they each decide to handle the inconsistencies. Therefore, the time and cost savings associated with simpler methods could result in overall acceptable data quality but still have unacceptable adverse results. The following sections describe briefly the strengths and weaknesses of each method examined and the potential time and cost impact of implementing them.

11.1.1 PMN Imputation

PMN is the imputation methodology currently used on the NSDUH. It was developed for the 1999 survey data when the survey converted from face-to-face paper-and-pencil interviewing

(PAPI) and self-administered questionnaire (SAQ) to audio computer-assisted self-interviewing (ACASI) and computer-assisted personal interviewing (CAPI), with skips programmed into the questionnaire. It is used in conjunction with an eligible case rule and the flag-and-impute editing procedure. The eligible case rule requires responses for at least 10 lifetime drug use questions including cigarettes. The flag-and-impute editing procedure identifies inconsistent responses and allows incorporation of partial data to restrict imputation outcomes. Initially, the combined editing and imputation methodology, including PMN, was applied to basic demographics and all core substance use variables but now is applied to all variables except the imputation of data of birth (random) and nicotine dependence (regression-based). Some of the PMN features considered important for comparison with other methods are the following:

- use of survey weights for all model fitting
- use of item response propensity models to adjust item respondent weights for each imputed variable or variable set
- definition of donor sets based on similar predicted means (termed the delta neighborhood)
- exclusion of donors that do not satisfy logical constraints
- exclusion of donors that do not satisfy likeness constraints (when feasible)
- ability to perform multivariate imputation (simultaneously imputing two or more related variables)
- multivariate modeling of related variables or variable sets to preserve relationships among variables
- use of a final hot-deck step to select a donor from among the eligible donor set
- use of the same three key processes for both categorical and continuous variables

11.1.2 Weighted Sequential Hot-Deck Imputation

One main goal of this evaluation was to identify whether there are advantages to using PMN imputation over simpler hot-deck methods and whether these advantages are sufficient to justify its continued use for the NSDUH. To address these questions, a simple weighted sequential hot-deck (WSHD) method and a complex WSHD method were developed and tested.

The evaluation of the WSHD methods attempted to determine whether using additional predictor variables in a WSHD method for imputing the drug variables improves estimates and whether there is an advantage in imputing the drug variables in a particular sequence. Due to the lack of significant differences between estimates based on WSHD and PMN methods, it can be concluded that there may not be an advantage to including additional predictor variables in the imputation model or imputing the drug variables in a particular sequence for the NSDUH. In simple WSHD where only a few predictor variables were used for developing imputation classes, the estimates did not show many significant differences; therefore, it may be possible to impute the drug variables with only a small subset of predictor variables. In complex WSHD, the drug usage variables were imputed independent of other drug usage variables. In other words, there were no relationships between drugs included in the development of the imputation classes, and it may be possible to conclude that the use of one drug may not necessarily assist with the imputation of another drug.

One main advantage that PMN has over these simpler methods is its ability to maintain high-quality data at the individual record level due to its complex logical and likeness constraints built specifically for the NSDUH data. As shown in Chapter 3, there are not many significant differences between the estimates based on PMN and WSHD (both simple and complex) imputed data, and the differences found to be statistically significant are small. The WSHD methods were not successful in maintaining all logical constraints as defined in PMN and did not test any likeness constraints. These simpler methods do not sacrifice data quality at the national estimate level, but inconsistencies remain at the individual level, and some of these would likely exhibit themselves in the complex conditional tables often required by NSDUH analysis. Additional modifications could potentially be made (by further restricting the donor sets) for each WSHD approach to resolve some of the inconsistencies, which would restore some of the complexities of the current PMN methods.

11.1.3 Sequential Regression Multivariate Imputation Using IVEware

IVEware software was used to determine whether NSDUH data could be imputed with off-the-shelf software because it contains several important features similar to PMN such as use of regression models and the ability to restrict and bound imputed values. In addition, it utilizes a multiple imputation (MI) procedure to measure the amount of variance inflation resulting from imputation. The evaluation of this software resulted in identifying problems with the software's performance. Similar problems to the one encountered in the WSHD methods in terms of maintaining data consistency also were identified. After different approaches were taken, it was apparent that the available version of this off-the-shelf software would not be best suited for imputing complex survey data with many logical constraints. Rather than attempting to modify IVEware, an approach that utilized some PMN features and some IVEware features was developed and evaluated (modPMN-MI, Section 11.1.4).

For comparison with other methods, variables for only three drugs were imputed using IVEware. The results for IVEware showed the most differences when compared with other methods.

Some technical features of the IVEware version tested that are important for comparing methods include the following:

- unweighted prediction model fitting
- no need for response propensity modeling
- sequential model fitting using both respondent variables and imputed variables
- iteration of model fitting so that each variable is modeled as a function of all other variables (could be restricted to subsets of variables)
- iteration stopping when a convergence criterion or a specified iteration count is reached
- augmentation of model parameters to achieve "proper" multiple imputations
- imputed values based on a random draw from the posterior distribution for each augmented model
- no multivariate imputation
- allowance for estimation of the variance contribution for imputation

As part of the investigation of the IVEware, the variance due to imputation was examined. The relative increase in variance due to imputation represents the increase relative to the naive within-imputation variance contribution that results when the imputation variance contribution is ignored. For the variables imputed with the IVEware method, the relative increase in variance due to imputation is typically small (less than 3 percent).

11.1.4 Modified Predictive Mean Neighborhood Multiple Imputation

Modified predictive mean neighborhood multiple imputation (modPMN-MI) was developed to utilize some of the best features of IVEware and PMN, and these features include the following:

- use of survey weights for all model fitting
- use of item response propensity models to adjust for item nonresponse for each imputed variable
- sequential model fitting so that each variable is modeled as a function of both reported and previously imputed variable values
- iteration for two cycles
- augmentation of model parameters to achieve "proper" multiple imputations
- for categorical variables, the final imputed variable based on a random draw from the posterior distribution for each augmented model
- for categorical variables, conformance with logical constraints achieved by draws from a conditional posterior distribution
- for continuous variables, a final hot-deck step to select a donor among the eligible donor set
- donor set restricted by logical constraints and by likeness constraints, when feasible

Initial cost estimates for modPMN-MI indicate an increase in cost for this method. Some of this increase is due to the need to implement and perform quality control checks on a new set of programs. It may be possible to reduce the cost impact through further simplifications. The schedule for implementing modPMN-MI still appears to be feasible and fits with other tasks required for post-survey data processing.

11.2 Race and Hispanicity Imputation

The purpose of this investigation was to examine improvements in the race and Hispanicity imputations, including using the screener data to impute interview race and Hispanicity data, imputing Hispanicity prior to race, and using alternative imputation methods for race and Hispanicity. After examining the race and Hispanicity distributions among the interview and screener data, it was determined that a deterministic imputation based on a proxy report during the screener for race may be good for white, black/African American, and Asian but may be biased for other race categories among non-Hispanics/Latinos. For Hispanic/Latino, the deterministic correlation only demonstrated a high degree of association among white and showed poor correlation for American Indian/Alaska Native.

The testing of five methods for imputing race and Hispanicity indicated that the screener race and Hispanicity variables are useful predictors in addition to the variables currently used in

PMN. The results indicate that including the screener covariates in PMN changed the final imputed distribution less than the addition of these covariates in other approaches. However, including screening variables in the model often resulted in convergence problems. Thus, more variables including some levels of the race screener variable needed to be removed. Using screener race as a predictor for interview race also resulted in an increase of white in the imputed race, which can be explained by the higher percentage of white among screener race respondents.

Race is an important variable in the NSDUH and is used in many analyses and tables. It is also used as a covariate in almost all of the imputation models. The standard NSDUH processing demonstrated that the best auxiliary variable for interview race imputation is the race and Hispanic/Latino origin indicator of the "head of household" according to the screener. Chapter 7 shows that the screener race and interview race distributions are not similar, especially for the white category, where screener race white tends to be identified as non-white in the interview data with "American Indian/Alaska Native" as the most changed category. Chapter 7 also compares the standard PMN method with several alternative model-based methods with screener race as a covariate. The screener race is a strong predictor in imputing interview race, Hispanic/Latino origin indicator, and the Hispanic/Latino group. However, when using the model-based imputation methods, the imputed interview race distribution tends to be more like the screener nonmissing race data. The standard PMN method produces the imputed race results most identical to the nonmissing interview race data.

The problem with the pure model-based methods (modPMN-MI, IVEware, and Methods 3-5 in Section 7.5) is that it is difficult to incorporate all these auxiliary variables in the models without encountering convergence problems. PMN has the advantage of being able to incorporate some variables via the model and others via likeness constraints in the hot-deck step. However, creative approaches to the modeling that allow the use of all the best auxiliary variables would be the best solution. It would permit an accurate imputation while simplifying the procedures by eliminating likeness constraints. It is possible that Method 5 could be tweaked to allow the use of the best covariates. For example, when convergence problems occur, a straightforward stochastic assignment based on the frequency distribution (like what was done for finer Asian categories in Section 7.5) is a reasonable alternative. Also, perhaps separate race models could be fit for some of the more common Hispanic/Latino groups. Further details on possible improvements are included in the last section of this chapter.

11.3 Pair Member Editing and Imputation

The pair member data were examined to determine whether a deterministic imputation method, where if a pair member has missing data, then the missing values would be assigned from the other pair member without missing data, could be applied. The investigation showed this approach performs better than PMN for the family-level income variables. This deterministic imputation method would reduce the amount of PMN imputation required for the income variables by more than 30 percent. However, a closer examination of nonresponse for the total family income variable (documented in Appendix L) reveals that the use of the other pair member's value can sometimes result in internally inconsistent records and should be implemented only if certain conditions are met.

11.4 Mental Health Imputation

The extent and nature of item nonresponse for the mental health variables used in the estimation of serious mental illness (SMI) and any mental illness (AMI) was examined in order to determine whether the current imputation method is performing adequately or whether alternative methods would substantially improve the quality of the estimates. Because the current imputation method replaces missing values with the smallest possible values, the SMI and AMI estimates thus developed have a negative bias. However, because the item response rates are high, the negative bias of the estimates may be minimal.

As a first test of whether the negative bias was noticeable, an alternative imputation method was applied where missing values were replaced with the largest possible values instead of with the smallest possible values. If the estimates derived using this "upper bound" estimate were outside the 95 percent confidence interval using the "lower bound" estimate, then that would suggest some evidence of bias due to nonresponse for that variable, and consequently an alternative imputation method might reduce that bias. This was indeed the case for both AMI and SMI.

Because this first test for bias was positive, a more sensitive test was applied. In this second test, instead of replacing the missing values with the largest possible values, a more complex imputation method was applied. This method was a more sophisticated application of the WSHD method described in Section 11.1.2 and was expected to result in unbiased estimates. The estimates for SMI and AMI using WSHD were well inside the 95 percent confidence interval associated with the lower bound estimate and in fact were barely larger than the lower bound point estimate. This was especially true for SMI.

11.5 Substance Dependence and Abuse Imputation

The extent and nature of item nonresponse for the substance dependence and abuse variables was examined using methods similar to those used for the mental health variables. Similar to the mental health variables, the current imputation method involves replacing missing values with the smallest possible values. And, also similar to the mental health variables, the item response rates for the substance dependence and abuse questions were high. However, some of the substance dependence and abuse variables are recodes of numerous variables, and when multiple variables are used to create a recoded variable, the overall response rate decreases as a function of the number of variables and the number of missing values used in the recode.

When estimates derived using WSHD were compared with estimates derived using the current imputation method, in most cases, they were not found to be substantially different. Even for the substance use disorder (SUD) variables, which are the most complex recodes, the WSHD-derived estimates were never outside the 95 percent confidence interval associated with the lower bound estimates. Still, the point estimates of prevalence (expressed as percentages) were far enough apart for several of the SUD variables to be noticeable out to one decimal place. Some of the SUD variables showing noticeably different point estimates are used in many analytic studies, and a more refined imputation method may be considered a substantive improvement.

11.6 Income Item Nonresponse Patterns

Patterns of item nonresponse for the total income questions were examined, both within a respondent's record and between the records of members of a family pair (following up on Chapter 8). Within a respondent's record, for the key income recodes INCOME5 and POVERTY2, it is possible to directly assign values for many respondents whose total finer categories family income variable is missing. Across the members of a family pair, there is often disagreement about total finer categories family income, suggesting a nontrivial amount of measurement error for these variables, and further suggesting that caution should be used when using the other pair member's value in imputation. It is possible that probe questions in the questionnaire and the use of proxy respondents may reduce both nonresponse bias and measurement error.

11.7 Possible Next Steps

These possible next steps and improvements are based on the premise that logical constraints are necessary to produce consistent data for NSDUH analysis. Possible next steps are grouped into the following three sets:

11.7.1 Improvements That Are Not Expected to Affect Trend Estimates

- Using a reduced set of predictor variables in certain response propensity models and continuing the investigation of developing a reduced set of predictor variables for the predictive mean models.
- Creating alternate versions of both mental health variables and substance dependence and abuse variables where missing values are explicitly recorded as missing.
- Creating imputation indicators for imputation-revised recoded variables like INCOME and POVERTY2 so that data users who do not wish to use imputed values can easily undo the imputation treatment.

11.7.2 Improvements That May Affect Trend Estimates

- Redefining the eligible case rule given a new imputation method.
- Reexamining editing procedures for possible simplification or other revisions given a new imputation method.
- Continuing to use PMN for imputing race and Hispanicity and switching the order of imputing Hispanic/Latino origin and race.
- For family-level income variables, using the other pair member's value in imputation when one pair member does not respond and the other does respond, as long as (1) the pair members are definitely from the same family, (2) the pair members agree that there are other family members in the household, and (3) the other pair member's value does not produce an internally inconsistent record for the item nonrespondent.
- Considering to use the other pair member's value in imputation for some of the demographic, household roster, and health insurance variables and considering to use the other pair member's value in the imputation models instead of in a deterministic manner.

- Considering a more refined imputation method for substance dependence and abuse variables.

11.7.3 Possible Areas for Further Exploration

- Attempting to modify the modPMN-MI for further simplicity and testing the impact on national and selected domain estimates.
- Dropping all or most likeness constraints for PMN.
- Using a single-imputation version of modPMN-MI without the data augmentation step for the primary imputation of variables to appear in the analytic file.
- Developing the capability for MI as a program module for PMN to be used for the following:

  - Measuring the impact of imputation on the variance of estimates for selected NSDUH reports.
  - Implementing MI procedures for variance estimation purposes only for variables based on low item response rates.
  - For continuous variables, considering defining the donor set based on satisfying logical constraints alone and doing a simple hot-deck selection. This is particularly applicable when the number of eligible donors is already small (e.g., imputation of 30-day frequency of use for past month heroin users).

- Assessing the measurement error associated with the total family income variable.
- Considering the use of questionnaire probes and the more liberal use of proxy respondents for total family income.

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Aldworth, J., Chromy, J. R., Foster, M. S., Heller, D. C., Packer, L. E., & Spagnola, K. (2009). Statistical inference report. In *2008 National Survey on Drug Use and Health: Methodological resource book* (Section 14, prepared for the Substance Abuse and Mental Health Services Administration, Contract No. 283-2004-00022, Deliverable No. 39, RTI/0209009.481.002). Research Triangle Park, NC: RTI International. Retrieved from https://www.samhsa.gov/data/

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (DSM-IV) (4th ed.). Washington, DC: American Psychiatric Association.

Ault, K., Aldworth, J., Barnett-Walker, K., Carpenter, L., Copello, E., Frechtel, P., Liu, B., & Martin, P. (2009). Imputation report. In *2007 National Survey on Drug Use and Health: Methodological resource book* (Section 11, prepared for the Substance Abuse and Mental Health Services Administration, Contract No. 283-2004-00022, Deliverable No. 39, RTI/0209009.377.007). Research Triangle Park, NC: RTI International. Retrieved from https://www.samhsa.gov/data/

Ault, K., Barnett-Walker, K., Carpenter, L., Copello, E., Cummiskey, C., Frechtel, P., Laufenberg, J., Liu, B., Martin, P., & Moore, A. (2010). Imputation report. In *2008 National Survey on Drug Use and Health: Methodological resource book* (Section 11, prepared for the Substance Abuse and Mental Health Services Administration, Contract No. 283-2004-00022, Deliverable No. 39, RTI/0209009.477.007). Research Triangle Park, NC: RTI International. Retrieved from https://www.samhsa.gov/data/

Ault, K., Barnett-Walker, K., Carpenter, L., Cummiskey, C., Frechtel, P., Laufenberg, J., Martin, P., Moore, A., & Scott, V. (2011). Imputation report. In *2009 National Survey on Drug Use and Health: Methodological resource book* (Section 11, prepared for the Substance Abuse and Mental Health Services Administration, Contract No. 283-2004-00022, Deliverable No. 39, RTI/0209009.577.007). Research Triangle Park, NC: RTI International. Retrieved from https://www.samhsa.gov/data/

Barnard, J., & Rubin, D. B. (1999). Small sample degrees of freedom with multiple imputation. *Biometrika, 86*, 948-955.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review, 51*, 279-292.

Bollinger, C. R., Hirsch, B. T., Hokayem, C. M., & Ziliak, J. P. (2014). *Trouble in the tails: Earnings non-response and response bias across the distribution.* Unpublished manuscript.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Boca Raton: Chapman & Hall/CRC.

Center for Behavioral Health Statistics and Quality. (2011). *National Survey on Drug Use and Health: 2010 Mental Health Surveillance Study codebook*. Retrieved from https://www.samhsa.gov/data/

Center for Behavioral Health Statistics and Quality. (2014). *2012 National Survey on Drug Use and Health: Methodological resource book (Section 16a, 2012 Mental Health Surveillance Study: Design and estimation report)*. Retrieved from https://www.samhsa.gov/data/

Center for Behavioral Health Statistics and Quality. (2015). *2014 National Survey on Drug Use and Health: Methodological summary and definitions*. Retrieved from https://www.samhsa.gov/data/

Chen, P., Cribb, D., Dai, L., Gordek, H., Laufenberg, J., Sathe, N., & Westlake, M. (2011). Person-level sampling weight calibration. In *2009 National Survey on Drug Use and Health: Methodological resource book* (Section 12, prepared for the Substance Abuse and Mental Health Services Administration, Contract No. 283-2004-00022, Phase V, Deliverable No. 39, RTI/0209009.574.002). Research Triangle Park, NC: RTI International. Retrieved from https://www.samhsa.gov/data/

Cochran, W. G. (1977). *Sampling techniques*. New York: Wiley.

Cox, B. G. (1980). The weighted sequential hot deck imputation procedure. In *Proceedings of the 1980 American Statistical Association, Survey Research Methods Section, Houston, TX* (pp. 721-726). Washington, DC: American Statistical Association. [Available as a PDF at http://www.amstat.org/ASA/Membership/Sections-and-Interest-Groups.aspx]

Frechtel, P., & Copello, E. (2007). Patterns of nonresponse for key questions in NSDUH and implications for imputation. In *Proceedings of the American Statistical Association, Survey Research Methods Section*.

Frechtel, P., Scott, V., Couzens, A., Moore, A., & Bose, J. (2012). Imputation using the other pair member. In *Proceedings of the American Statistical Association, Survey Research Methods Section* (pp. 4248-4258).

Frechtel, P., Archambault, H., Carpenter, L., Cummiskey, C., Edwards, S., Laufenberg, J., Martin, P., Moore, A., & Scott, V. (2012). Imputation report. In *2010 National Survey on Drug Use and Health: Methodological resource book* (Section 11, prepared for the Substance Abuse and Mental Health Services Administration, Contract No. HHSS283200800004C, Deliverable No. 39, RTI/0211838.107.006.007). Research Triangle Park, NC: RTI International. Retrieved from https://www.samhsa.gov/data/

Frechtel, P., Archambault, H., Carpenter, L., Cummiskey, C., Edwards, S., Laufenberg, J., Martin, P., Moore, A., & Scott, V. (2013). Imputation report. In *2011 National Survey on Drug Use and Health: Methodological resource book* (Section 11, prepared for the Substance Abuse and Mental Health Services Administration, Contract No. HHSS283200800004C, Deliverable No. 39, RTI/0211838.207.006.007). Research Triangle Park, NC: RTI International. Retrieved from https://www.samhsa.gov/data/

Furukawa, T. A., Kessler, R. C., Slade, T., & Andrews, G. (2003). The performance of the K6 and K10 screening scales for psychological distress in the Australian National Survey of Mental Health and Well-Being. *Psychological Medicine, 33*, 357-362.

Juster, T., & Smith, J. P. (1997). Improving the quality of economic data: Lessons from the HRS and AHEAD. *Journal of the American Statistical Association, 92*, 1268-1278.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical variables. *Applied Statistics, 29(2)*, 119-127.

Kessler, R. C., Barker, P. R., Colpe, L. J., Epstein, J. F., Gfroerer, J. C., Hiripi, E., Howes, M. J., Normand, S. L., Manderscheid, R. W., Walters, E. E., & Zaslavsky, A. M. (2003). Screening for serious mental illness in the general population. *Archives of General Psychiatry, 60*, 184-189. doi: yoa20567 [pii]

Khare, M., Little, R. J. A., Rubin, D. B., & Schafer, J. L. (1993). Multiple imputation of NHANES III. In *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Kott, P. S., & Folsom, R. E. (2010). Weights, double protection, and multiple imputation. In *Joint Statistical Meetings, Survey Research Methods Section.* Alexandria, VA: American Statistical Association.

Kroutil, L. A., & Handley, W. (2009). General principles and procedures for editing drug use data in the 2007 NSDUH computer-assisted interview. In *2007 National Survey on Drug Use and Health: Methodological resource book* (Section 10, prepared for the Substance Abuse and Mental Health Services Administration, Contract No. 283-2004-00022, Deliverable No. 39, RTI/0209009.373). Research Triangle Park, NC: RTI International. Retrieved from https://www.samhsa.gov/data/

Kroutil, L. A., Handley, W., Felts, B. J., Bradshaw, M. R., & Chien, C. (2009). Procedures for editing supplementary self-administered data in the 2007 NSDUH computer-assisted interview. In *2007 National Survey on Drug Use and Health: Methodological resource book* (Section 10, prepared for the Substance Abuse and Mental Health Services Administration, Contract No. 283-2004-00022, Deliverable No. 39, RTI/0209009.373). Research Triangle Park, NC: RTI International. Retrieved from https://www.samhsa.gov/data/

Little, R., & Raghunathan, T. E. (1997). Should imputation of missing data condition on all observed variables? In *Proceedings of the Section on Survey Research Methods, 1997 JSM*. Anaheim, CA.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Manly, B. F. J. (1986). *Multivariate statistical methods: A primer*. London, England: Chapman and Hall.

Marker, D. A., Judkins, D. R., & Winglee, M. (2002). Large-scale imputation for complex survey. In R. M. Groves, D. A. Dillman, J. L. Eltinge, R. J. Little (Eds.), *Survey nonresponse*. New York: Wiley-InterScience.

Moore, J. C., Stinson, L. L., & Welniak, Jr., E. J. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics, 16*(4), 331-361.

Morton, K. B., Martin, P. C., Chromy, J. R., Foster, M., & Hirsch, E. L. (2010). Sample design report. In *2009 National Survey on Drug Use and Health: Methodological resource book* (Section 2, prepared for the Substance Abuse and Mental Health Services Administration, Contract No. 283-2004-00022, Phase V, Deliverable No. 8, RTI/0209009.530.004). Research Triangle Park, NC: RTI International. Retrieved from https://www.samhsa.gov/data/

Novak, S. P., Colpe, L. J., Barker, P. R., & Gfroerer, J. C. (2010). Development of a brief mental health impairment scale using a nationally representative sample in the USA. *International Journal of Methods in Psychiatric Research*, *19*(Suppl. 1), 49-60. doi:10.1002/mpr.313

Perneger, T. V., & Burnand, B. (2005). A simple imputation algorithm reduced missing data in SF-12 health surveys. *Journal of Clinical Epidemiology, 58*(2), 142-149.

Pleis, J. R., & Dahlhamer, J. M. (2004). Family income response patterns for varying levels of income detail: An analysis of the National Health Interview Survey (NHIS). In *Proceedings of the Joint Statistical Meetings, American Statistical Association, Survey Research Methods Section* (pp. 4200-4207). Alexandria, VA: American Statistical Association.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A Multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*, 75-85.

Raghunathan, T. E., Solenberger, P., & Van Hoewyk, J. (2002, March). IVEware: Imputation and variance estimation software user guide. Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan. [Available at http://www.src.isr.umich.edu/software/]

Rehm, J., Üstün, T. B., Saxena, S., Nelson, C. B., Chatterji, S., Ivis, F., & Adlaf, E. (1999). On the development and psychometric testing of the WHO screening instrument to assess disablement in the general population. *International Journal of Methods in Psychiatric Research, 8*(2), 110-123. doi:10.1002/mpr.61

RTI International. (2012, August 29). *2011 National Survey on Drug Use and Health: Final analytic file codebook* (prepared for the Substance Abuse and Mental Health Services Administration under Contract No. HHSS283200800004C, Deliverable No. 24, RTI 0211838). Research Triangle Park, NC: Author.

RTI International. (2013). *SUDAAN*®, Release 11.0.1 [computer software]. Research Triangle Park, NC: RTI International.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics, 4*, 87-94.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: John Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association, 91*, 473-489.

Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association, 81*, 366-374.

SAS Institute Inc. (2002). Enterprise Miner 4.3 [Computer software]. Cary, NC: SAS Institute Inc.

Schenker, N., Raghunathan, T. E., Chiu, P. L., Makuc, D. M., Zhang, G., & Cohen, A. J. (2006). Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association, 101*, 924-933.

Schenker, N., Raghunathan, T. E., Chiu, P.-L., Makuc, D. M., Zhang, G., & Cohen, A. J. (2007). *Multiple imputation of family income and personal earnings in the National Health Interview Survey: Methods and examples*. University of Michigan: National Center for Health Statistics.

Singh, A. Grau, E., & Folsom, R. (2002). Predictive mean neighborhood imputation for NHSDA substance use data. In *Redesigning an ongoing national household survey: Methodological issues. DHHS Publication No. SMA 03-3768*, edited by J. Gfroerer, J. Eyerman and J. Chromy. Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Strang, G. (1988). *Linear algebra and its applications* (3rd ed.). Brooks/Cole.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859.

Yulei, H., Zaslavsky, A. M., Harrington, D. P., Catalano, P., & Landrum, M. B. (2007). Imputation in a multiformat and multiwave survey of cancer care. In *Proceedings of the Section on Health Policy Statistics, 2007 JSM*.

Zeger, S., & Liang, K. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics, 42*, 121-130.

This page intentionally left blank

Appendix A: Model Summaries

This page intentionally left blank

A.1 Model Summaries for the Predictive Mean Neighborhood Imputation Method

Table A.1 Model Summaries for PMN, Demographics: 15 Years or Older

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model
Marital Status[1]	Census Region; Gender; Population Density; Age Category; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Age Category * Gender	Age; Percentage Black/African American in Segment; Percentage Owner Occupied in Segment; Gender; Age * Gender; Census Region; Population Density; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment

PMN = predictive mean neighborhood.

Note: An asterisk "*" represents an interaction between two variables.

[1] Respondents aged 12 to 14 were assigned a skip code and were not included in any imputation steps for marital status.

Table A.2 Model Summaries for PMN, Demographics: 12 to 17 Years

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model
Race	Census Region; Household Type; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment	Census Region; Household Type; Age; Age Squared; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment
Hispanic/Latino Origin	Census Region; Imputed Race; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment	Census Region; Imputed Race; Household Type; Age; Age Squared; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment
Education Level[1]	No model used: no nonrespondents	Census Region; Imputed Race; Gender; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment

PMN = predictive mean neighborhood.

[1] Respondents aged 12 to 17 were assigned a skip code and were not included in any imputation steps for education level.

Table A.3 Model Summaries for PMN, Demographics: 18 to 25 Years

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model
Race	Census Region; Household Type; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment	Census Region; Household Type; Age; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Imputed Marital Status
Hispanic/Latino Origin	Census Region; Imputed Race; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment	Census Region; Imputed Race; Household Type; Age; Age Squared; Age Cubed; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Imputed Marital Status
Education Level	Census Region; Imputed Race; Gender; Age Category; Age Category * Gender; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment	Census Region; Imputed Race; Gender; Age; Age Squared; Age Cubed; Age * Gender; Age Squared * Gender; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Imputed Marital Status

PMN = predictive mean neighborhood.

Note: An asterisk "*" represents an interaction between two variables.

Table A.4 Model Summaries for PMN, Demographics: 26 Years or Older

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model
Race	Census Region; Household Type; Age Category; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment	Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Census Region
Hispanic/Latino Origin	Census Region; Imputed Race; Age Category; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment	Household Type; Census Region; Imputed Race; Age; Age Squared; Age Cubed; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Imputed Marital Status
Education Level	Census Region; Imputed Race; Gender; Age Category; Age Category * Gender; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment	Census Region; Imputed Race; Gender; Age; Age Squared; Age Cubed; Age * Gender; Age Squared * Gender; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Imputed Marital Status

PMN = predictive mean neighborhood.

Note: An asterisk "*" represents an interaction between two variables.

Table A.5 Model Summaries for PMN, Marijuana: 12 to 17 Years

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model
Lifetime	Gender; Race; Gender * Race; CBSA; Census Region; Cigarette Lifetime Indicator	Cigarette Lifetime Indicator; Intermediate Lifetime Indicators for Snuff, Chewing Tobacco, Cigars, Pipes, Alcohol, and Inhalants; Age; Gender; Race; Age Squared; Age Cubed; Gender * Race; Age * Gender; Age * Race; State Rank; CBSA; Census Region
Recency	Imputed Recencies for Cigarettes and Alcohol; Imputed Lifetime Indicators for Smokeless Tobacco, Cigars, Pipes, Inhalants, Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Gender; Race; Gender * Race; Census Region; CBSA; State Rank	Imputed Recencies for Cigarettes and Alcohol; Imputed Lifetime Indicators for Smokeless Tobacco, Cigars, Pipes, Inhalants, Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Census Region; CBSA; State Rank
12-Month Frequency	Intermediate Past Month Marijuana Indicator; Imputed Recencies for Cigarettes, Smokeless Tobacco, Cigars, Pipes, Alcohol, and Inhalants; Imputed Lifetime Indicators for Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Race; Gender; Census Region; CBSA	Intermediate Past Month Marijuana Indicator; Imputed Recencies for Cigarettes, Smokeless Tobacco, Cigars, Pipes, Alcohol, and Inhalants; Imputed Lifetime Indicators for Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Census Region; CBSA; State Rank
30-Day Frequency	Intermediate Marijuana 12-Month Frequency; Imputed Recencies for Cigarettes, Smokeless Tobacco, Cigars, Pipes, Alcohol, and Inhalants; Imputed Lifetime Indicators for Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Race; Gender; Census Region; CBSA	Intermediate Marijuana 12-Month Frequency; Imputed Recencies for Cigarettes, Smokeless Tobacco, Cigars, Pipes, Alcohol, and Inhalants; Imputed Lifetime Indicators for Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Census Region; Age; Gender; Race; Age Squared; Age Cubed; Age * Race; Gender * Race; Age * Gender; CBSA; State Rank

Table A.5 Model Summaries for PMN, Marijuana: 12 to 17 Years (continued)

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model
Age at First Use	Imputed Recencies for Cigarettes, Smokeless Tobacco, Cigars, Pipes, Alcohol, Inhalants, and Marijuana; Imputed Lifetime Indicators for Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Race; Gender; Census Region; CBSA	Marijuana 30-Day Frequency; Marijuana 12-Month Frequency; Imputed Ages at First Use for Cigarettes, Daily Cigarettes, Smokeless Tobacco, Cigars, Alcohol, and Inhalants; Imputed Recencies for Cigarettes, Smokeless Tobacco, Cigars, Pipes, Alcohol, Inhalants, and Marijuana; Imputed Lifetime Indicators for Daily Cigarettes, Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Age; Gender; Race; State Rank; Age Squared; Age Cubed; Gender * Race; Age * Gender; Age * Race; Age Squared * Gender; Age Squared * Race; Census Region; CBSA

CBSA = core-based statistical area; PMN = predictive mean neighborhood.

Note: An asterisk "*" represents an interaction between two variables.

Table A.6 Model Summaries for PMN, Marijuana: 18 to 25 Years

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model
Lifetime	Gender; Race; Gender * Race; Marital Status; Education Level; Employment Status; CBSA; Census Region; Cigarette Lifetime Indicator	Cigarette Lifetime Indicator; Intermediate Lifetime Indicators for Snuff, Chewing Tobacco, Cigars, Pipes, Alcohol, and Inhalants; Education Level; Age; Gender; Race; Age Squared; Age Cubed; Marital Status; Employment Status; Gender * Race; Age * Gender; Age * Race; State Rank; CBSA; Census Region
Recency	Imputed Recencies for Cigarettes and Alcohol; Imputed Lifetime Indicators for Smokeless Tobacco, Cigars, Pipes, Inhalants, Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Gender; Race; Gender * Race; Marital Status; Education Level; Employment Status; Census Region; CBSA; State Rank	Imputed Recencies for Cigarettes and Alcohol; Imputed Lifetime Indicators for Smokeless Tobacco, Cigars, Pipes, Inhalants, Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Marital Status; Education Level; Employment Status; Census Region; CBSA; State Rank
12-Month Frequency	Intermediate Past Month Marijuana Indicator; Imputed Recencies for Cigarettes, Smokeless Tobacco, Cigars, Pipes, Alcohol, and Inhalants; Imputed Lifetime Indicators for Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Race; Gender; Census Region; CBSA	Intermediate Past Month Marijuana Indicator; Imputed Recencies for Cigarettes, Smokeless Tobacco, Cigars, Pipes, Alcohol, and Inhalants; Imputed Lifetime Indicators for Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Marital Status; Education Level; Employment Status; Census Region; CBSA; State Rank
30-Day Frequency	Intermediate Marijuana 12-Month Frequency; Imputed Recencies for Cigarettes, Smokeless Tobacco, Cigars, Pipes, Alcohol, and Inhalants; Imputed Lifetime Indicators for Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Race; Gender; Census Region; CBSA	Intermediate Marijuana 12-Month Frequency; Imputed Recencies for Cigarettes, Smokeless Tobacco, Cigars, Pipes, Alcohol, and Inhalants; Imputed Lifetime Indicators for Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Census Region; Age; Gender; Race; Age Squared; Age Cubed; Age * Race; Gender * Race; Age * Gender; Marital Status; Education Level; Employment Status; CBSA; State Rank

Table A.6 Model Summaries for PMN, Marijuana: 18 to 25 Years (continued)

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model
Age at First Use	Imputed Recencies for Cigarettes, Smokeless Tobacco, Cigars, Pipes, Alcohol, Inhalants, and Marijuana; Imputed Lifetime Indicators for Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Race; Gender; Census Region; CBSA	Marijuana 30-Day Frequency; Marijuana 12-Month Frequency; Imputed Ages at First Use for Cigarettes, Daily Cigarettes, Smokeless Tobacco, Cigars, Alcohol, and Inhalants; Imputed Recencies for Cigarettes, Smokeless Tobacco, Cigars, Pipes, Alcohol, Inhalants, and Marijuana; Imputed Lifetime Indicators for Daily Cigarettes, Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Age; Gender; Race; State Rank; Age Squared; Age Cubed; Gender * Race; Age * Gender; Age * Race; Age Squared * Gender; Age Squared * Race; Marital Status; Education Level; Employment Status; Census Region; CBSA

CBSA = core-based statistical area; PMN = predictive mean neighborhood.

Note: An asterisk "*" represents an interaction between two variables.

Table A.7 Model Summaries for PMN, Marijuana: 26 Years or Older

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model
Lifetime	Age Category; Gender; Race; Gender * Race; Marital Status; Education Level; Employment Status; CBSA; Census Region; Cigarette Lifetime Indicator	Cigarette Lifetime Indicator; Intermediate Lifetime Indicators for Snuff, Chewing Tobacco, Cigars, Pipes, Alcohol, and Inhalants; Education Level; Age; Gender; Race; Employment Status; Age * Gender; Age * Race
Recency	Imputed Recencies for Cigarettes and Alcohol; Imputed Lifetime Indicators for Smokeless Tobacco, Cigars, Pipes, Inhalants, Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Gender; Age Category; Race; Gender * Race; Marital Status; Education Level; Employment Status; Census Region; CBSA; State Rank	Imputed Recencies for Cigarettes and Alcohol; Imputed Lifetime Indicators for Smokeless Tobacco, Cigars, Pipes, Inhalants, Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Marital Status; Education Level; Employment Status; Census Region; CBSA; State Rank
12-Month Frequency	Intermediate Past Month Marijuana Indicator; Imputed Recencies for Cigarettes, Smokeless Tobacco, Cigars, Pipes, Alcohol, and Inhalants; Imputed Lifetime Indicators for Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Age Category; Race; Gender; Census Region; CBSA	Intermediate Past Month Marijuana Indicator; Imputed Recencies for Cigarettes, Smokeless Tobacco, Cigars, Pipes, Alcohol, and Inhalants; Imputed Lifetime Indicators for Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Marital Status; Education Level; Employment Status; Census Region; CBSA; State Rank
30-Day Frequency	Intermediate Marijuana 12-Month Frequency; Imputed Recencies for Cigarettes, Smokeless Tobacco, Cigars, Pipes, Alcohol, and Inhalants; Imputed Lifetime Indicators for Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Age Category; Race; Gender; Census Region; CBSA	Intermediate Marijuana 12-Month Frequency; Imputed Recencies for Cigarettes, Smokeless Tobacco, Cigars, Pipes, Alcohol, and Inhalants; Imputed Lifetime Indicators for Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Census Region; Age; Gender; Race; Age Squared; Age Cubed; Age * Race; Gender * Race; Age * Gender; Marital Status; Education Level; Employment Status; CBSA; State Rank

Table A.7 Model Summaries for PMN, Marijuana: 26 Years or Older (continued)

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model
Age at First Use	Imputed Recencies for Cigarettes, Smokeless Tobacco, Cigars, Pipes, Alcohol, Inhalants, and Marijuana; Imputed Lifetime Indicators for Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Age Category; Race; Gender; Census Region; CBSA	Marijuana 30-Day Frequency; Marijuana 12-Month Frequency; Imputed Ages at First Use for Cigarettes, Daily Cigarettes, Smokeless Tobacco, Cigars, Alcohol, and Inhalants; Imputed Recencies for Cigarettes, Smokeless Tobacco, Cigars, Pipes, Alcohol, Inhalants, and Marijuana; Imputed Lifetime Indicators for Daily Cigarettes, Hallucinogens, Pain Relievers, Tranquilizers, Stimulants, Sedatives, Cocaine, Crack, and Heroin; Age; Gender; Race; State Rank; Age Squared; Age Cubed; Gender * Race; Age * Gender; Age * Race; Age Squared * Gender; Age Squared * Race; Marital Status; Education Level; Employment Status; Census Region; CBSA

CBSA = core-based statistical area; PMN = predictive mean neighborhood.

Note: An asterisk "*" represents an interaction between two variables.

A.2 Chi-square Automatic Interaction Detection Results for the Simple and Complex Weighted Sequential Hot-Deck Imputation Methods

Table A.8 CHAID Results for Simple WSHD, Demographics: 15 Years or Older

Variable	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
Marital Status[1]	Census Region; Gender; Population Density; Age; Percentage Black/African American in Segment; Percentage Owner Occupied in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment	Age; Gender; Percentage Black/African American in Segment; Percentage Owner Occupied in Segment

CHAID = Chi-square Automatic Interaction Detection; WSHD = weighted sequential hot deck.

[1] Respondents aged 12 to 14 were assigned a skip code and were not included in any imputation steps for marital status.

Table A.9 CHAID Results for Simple WSHD, Demographics: 12 Years or Older

Variable	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
Hispanic/Latino Origin	Census Region; Household Type; Age; Percentage Black/African American in Segment; Percentage Owner Occupied in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Imputed Marital Status; Imputed Race	Household Type; Percentage Hispanic/Latino in Segment; Imputed Race; Census Region; Percentage Black/African American in Segment
Education Level[1]	Census Region; Gender; Age; Percentage Black/African American in Segment; Percentage Owner Occupied in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Imputed Marital Status; Imputed Race	Age; Percentage Hispanic/Latino in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Black/African American in Segment; Imputed Race; Imputed Marital Status; Gender; Percentage Owner Occupied in Segment; Census Region

CHAID = Chi-square Automatic Interaction Detection; WSHD = weighted sequential hot deck.

[1] Respondents aged 12 to 17 were assigned a skip code and were not included in any imputation steps for education level.

Table A.10 CHAID Results for Simple WSHD, Race: 12 to 17 Years

Variable	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
Race	Census Region; Household Type; Age; Percentage Black/African American in Segment; Percentage Owner Occupied in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Imputed Marital Status	Percentage Black/African American in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment

CHAID = Chi-square Automatic Interaction Detection; WSHD = weighted sequential hot deck.

Table A.11 CHAID Results for Simple WSHD, Race: 18 to 25 Years

Variable	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
Race	Census Region; Household Type; Age; Percentage Black/African American in Segment; Percentage Owner Occupied in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Imputed Marital Status	Percentage Black/African American in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment

CHAID = Chi-square Automatic Interaction Detection; WSHD = weighted sequential hot deck.

Table A.12 CHAID Results for Simple WSHD, Race: 26 Years or Older

Variable	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
Race	Census Region; Household Type; Age; Percentage Black/African American in Segment; Percentage Owner Occupied in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Imputed Marital Status	Percentage Black/African American in Segment

CHAID = Chi-square Automatic Interaction Detection; WSHD = weighted sequential hot deck.

Table A.13 CHAID Results for Simple WSHD, Marijuana: 12 to 17 Years

Variable	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
Lifetime	Census Region; Gender; Age; Imputed Marital Status; Imputed Race; Imputed Hispanic/Latino Origin; Imputed Education Level; Population Density; Cigarettes Lifetime Use	Cigarettes Lifetime Use; Age; Census Region; Population Density; Imputed Race

CHAID = Chi-square Automatic Interaction Detection; WSHD = weighted sequential hot deck.

Table A.14 CHAID Results for Simple WSHD, Marijuana: 18 to 25 Years

Variable	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
Lifetime	Census Region; Gender; Age; Imputed Marital Status; Imputed Race; Imputed Hispanic/Latino Origin; Imputed Education Level; Population Density; Cigarettes Lifetime Use	Cigarettes Lifetime Use; Imputed Hispanic/Latino Origin; Imputed Race; Census Region; Population Density; Imputed Education Level; Imputed Marital Status; Age

CHAID = Chi-square Automatic Interaction Detection; WSHD = weighted sequential hot deck.

Table A.15 CHAID Results for Simple WSHD, Marijuana: 26 Years or Older

Variable	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
Lifetime	Census Region; Gender; Age; Imputed Marital Status; Imputed Race; Imputed Hispanic/Latino Origin; Imputed Education Level; Population Density; Cigarettes Lifetime Use	Cigarettes Lifetime Use; Imputed Hispanic/Latino Origin; Imputed Race; Census Region; Population Density; Imputed Education Level; Imputed Marital Status; Age

CHAID = Chi-square Automatic Interaction Detection; WSHD = weighted sequential hot deck.

Table A.16 CHAID Results for Complex WSHD, Marijuana: 12 to 17 Years

Variable	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
Lifetime	Age; Gender; Imputed Race; Population Density; Census Region; Imputed Marital Status; Imputed Education Level; Imputed Hispanic/Latino Origin; State Rank; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; Imputed Inhalants Lifetime Use	Age; Imputed Race; Population Density; Census Region; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; Imputed Inhalants Lifetime Use
Recency	Age; Gender; Imputed Race; Population Density; Census Region; Imputed Marital Status; Imputed Education Level; Imputed Hispanic/Latino Origin; State Rank; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; Imputed Inhalants Lifetime Use; Imputed Marijuana Lifetime Use; Imputed Pain Relievers Lifetime Use; Imputed Heroin Lifetime Use; Imputed Cocaine Lifetime Use	Age; Imputed Race; Population Density; Census Region; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; State Rank
12-Month Frequency	Age; Gender; Imputed Race; Population Density; Census Region; Imputed Marital Status; Imputed Education Level; Imputed Hispanic/Latino Origin; State Rank; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; Imputed Inhalants Lifetime Use; Imputed Marijuana Lifetime Use; Imputed Pain Relievers Lifetime Use; Imputed Heroin Lifetime Use; Imputed Cocaine Lifetime Use; Imputed Marijuana Recency	Gender; Imputed Race; Cigarettes Lifetime Use; Imputed Pain Relievers Lifetime Use; Imputed Cocaine Lifetime Use; Imputed Marijuana Recency
30-Day Frequency	Age; Gender; Imputed Race; Population Density; Census Region; Imputed Marital Status; Imputed Education Level; Imputed Hispanic/Latino Origin; State Rank; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; Imputed Inhalants Lifetime Use; Imputed Marijuana Lifetime Use; Imputed Pain Relievers Lifetime Use; Imputed Heroin Lifetime Use; Imputed Cocaine Lifetime Use; Imputed Marijuana Recency; Imputed Marijuana 12-Month Frequency	Age; Imputed Pain Relievers Lifetime Use; Imputed Marijuana 12-Month Frequency
Age at First Use	Age; Gender; Imputed Race; Population Density; Census Region; Imputed Marital Status; Imputed Education Level; Imputed Hispanic/Latino Origin; State Rank; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; Imputed Inhalants Lifetime Use; Imputed Marijuana Lifetime Use; Imputed Pain Relievers Lifetime Use; Imputed Heroin Lifetime Use; Imputed Cocaine Lifetime Use; Imputed Marijuana Recency; Imputed Marijuana 12-Month Frequency; Imputed Marijuana 30-Day Frequency	Age; Gender; Census Region; Imputed Marital Status; Cigarettes Lifetime Use; Imputed Inhalants Lifetime Use; Imputed Pain Relievers Lifetime Use; Imputed Cocaine Lifetime Use; Imputed Marijuana 12-Month Frequency

CHAID = Chi-square Automatic Interaction Detection; WSHD = weighted sequential hot deck.

Table A.17 CHAID Results for Complex WSHD, Marijuana: 18 to 25 Years

Variable	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
Lifetime	Age; Gender; Imputed Race; Population Density; Census Region; Imputed Marital Status; Imputed Education Level; Imputed Hispanic/Latino Origin; State Rank; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; Imputed Inhalants Lifetime Use	Age; Imputed Race; Census Region; Imputed Education Level; Imputed Hispanic/Latino Origin; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; Imputed Inhalants Lifetime Use
Recency	Age; Gender; Imputed Race; Population Density; Census Region; Imputed Marital Status; Imputed Education Level; Imputed Hispanic/Latino Origin; State Rank; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; Imputed Inhalants Lifetime Use; Imputed Marijuana Lifetime Use; Imputed Pain Relievers Lifetime Use; Imputed Heroin Lifetime Use; Imputed Cocaine Lifetime Use	Age; Imputed Race; Population Density; Census Region; Imputed Education Level; State Rank
12-Month Frequency	Age; Gender; Imputed Race; Population Density; Census Region; Imputed Marital Status; Imputed Education Level; Imputed Hispanic/Latino Origin; State Rank; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; Imputed Inhalants Lifetime Use; Imputed Marijuana Lifetime Use; Imputed Pain Relievers Lifetime Use; Imputed Heroin Lifetime Use; Imputed Cocaine Lifetime Use; Imputed Marijuana Recency	Age; Gender; Imputed Race; Imputed Education Level; Imputed Hispanic/Latino Origin; Imputed Inhalants Lifetime Use; Imputed Pain Relievers Lifetime Use; Imputed Heroin Lifetime Use; Imputed Cocaine Lifetime Use; Imputed Marijuana Recency
30-Day Frequency	Age; Gender; Imputed Race; Population Density; Census Region; Imputed Marital Status; Imputed Education Level; Imputed Hispanic/Latino Origin; State Rank; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; Imputed Inhalants Lifetime Use; Imputed Marijuana Lifetime Use; Imputed Pain Relievers Lifetime Use; Imputed Heroin Lifetime Use; Imputed Cocaine Lifetime Use; Imputed Marijuana Recency; Imputed Marijuana 12-Month Frequency	Gender; Imputed Education Level; Imputed Inhalants Lifetime Use; Imputed Marijuana 12-Month Frequency
Age at First Use	Age; Gender; Imputed Race; Population Density; Census Region; Imputed Marital Status; Imputed Education Level; Imputed Hispanic/Latino Origin; State Rank; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; Imputed Inhalants Lifetime Use; Imputed Marijuana Lifetime Use; Imputed Pain Relievers Lifetime Use; Imputed Heroin Lifetime Use; Imputed Cocaine Lifetime Use; Imputed Marijuana Recency; Imputed Marijuana 12-Month Frequency; Imputed Marijuana 30-Day Frequency	Age; Census Region; Imputed Education Level; Cigarettes Lifetime Use; Imputed Inhalants Lifetime Use; Imputed Pain Relievers Lifetime Use; Imputed Heroin Lifetime Use; Imputed Cocaine Lifetime Use; Imputed Marijuana Recency; Imputed Marijuana 12-Month Frequency; Imputed Marijuana 30-Day Frequency; State Rank

CHAID = Chi-square Automatic Interaction Detection; WSHD = weighted sequential hot deck.

Table A.18 CHAID Results for Complex WSHD, Marijuana: 26 Years or Older

Variable	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
Lifetime	Age; Gender; Imputed Race; Population Density; Census Region; Imputed Marital Status; Imputed Education Level; Imputed Hispanic/Latino Origin; State Rank; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; Imputed Inhalants Lifetime Use	Age; Imputed Race; Population Density; Imputed Marital Status; Imputed Education Level; Imputed Hispanic/Latino Origin; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; Imputed Inhalants Lifetime Use
Recency	Age; Gender; Imputed Race; Population Density; Census Region; Imputed Marital Status; Imputed Education Level; Imputed Hispanic/Latino Origin; State Rank; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; Imputed Inhalants Lifetime Use; Imputed Marijuana Lifetime Use; Imputed Pain Relievers Lifetime Use; Imputed Heroin Lifetime Use; Imputed Cocaine Lifetime Use	Age; Imputed Race; Imputed Marital Status; Imputed Pain Relievers Lifetime Use; Imputed Heroin Lifetime Use; Imputed Cocaine Lifetime Use
12-Month Frequency	Age; Gender; Imputed Race; Population Density; Census Region; Imputed Marital Status; Imputed Education Level; Imputed Hispanic/Latino Origin; State Rank; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; Imputed Inhalants Lifetime Use; Imputed Marijuana Lifetime Use; Imputed Pain Relievers Lifetime Use; Imputed Heroin Lifetime Use; Imputed Cocaine Lifetime Use; Imputed Marijuana Recency	Age; Gender; Imputed Education Level; Imputed Inhalants Lifetime Use; Imputed Pain Relievers Lifetime Use; Imputed Marijuana Recency
30-Day Frequency	Age; Gender; Imputed Race; Population Density; Census Region; Imputed Marital Status; Imputed Education Level; Imputed Hispanic/Latino Origin; State Rank; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; Imputed Inhalants Lifetime Use; Imputed Marijuana Lifetime Use; Imputed Pain Relievers Lifetime Use; Imputed Heroin Lifetime Use; Imputed Cocaine Lifetime Use; Imputed Marijuana Recency; Imputed Marijuana 12-Month Frequency	Imputed Marijuana Lifetime Use; Imputed Pain Relievers Lifetime Use
Age at First Use	Age; Gender; Imputed Race; Population Density; Census Region; Imputed Marital Status; Imputed Education Level; Imputed Hispanic/Latino Origin; State Rank; Cigarettes Lifetime Use; Imputed Alcohol Lifetime Use; Imputed Inhalants Lifetime Use; Imputed Marijuana Lifetime Use; Imputed Pain Relievers Lifetime Use; Imputed Heroin Lifetime Use; Imputed Cocaine Lifetime Use; Imputed Marijuana Recency; Imputed Marijuana 12-Month Frequency; Imputed Marijuana 30-Day Frequency	Age; Gender; Imputed Education Level; Imputed Hispanic/Latino Origin; Imputed Inhalants Lifetime Use; Imputed Heroin Lifetime Use; Imputed Cocaine Lifetime Use; Imputed Marijuana 12-Month Frequency

CHAID = Chi-square Automatic Interaction Detection; WSHD = weighted sequential hot deck.

A.3 Model Summaries for the IVEware Imputation Method

Table A.19 Model Summaries for IVEware, Demographics: 15 Years or Older

Variable	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycles 2–4	Variables Included in Predictive Mean Model, Cycle 5
Marital Status[1]	Age; Gender; Household Type; Census Region; Population Density; Percentage Hispanic/Latino in Segment; Percentage Black/African American in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Owner Occupied in Segment; Employment Status; CBSA; Cigarette Lifetime Indicator; Intermediate Imputed Education Level; Intermediate Lifetime Indicators for Cigars, Chewing Tobacco, and Pipes; Preliminary Analysis Weight	Age; Gender; Household Type; Census Region; Population Density; Percentage Hispanic/Latino in Segment; Percentage Black/African American in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Owner Occupied in Segment; Employment Status; CBSA; Cigarette Lifetime Indicator; Intermediate Imputed Education Level; Intermediate Imputed Hispanic/Latino Origin Indicator; Intermediate Imputed Race; Intermediate Lifetime Indicators for Cigars, Chewing Tobacco, Pipes, Snuff, Smokeless Tobacco, Alcohol, Heroin, Cocaine, Crack, Marijuana, PCP, LSD, Methamphetamine, Ecstasy, Inhalants, Tranquilizers, Stimulants, Sedatives, OxyContin, Hallucinogens, Pain Relievers, Other Pain Relievers, Other Stimulants, Other Hallucinogens; Preliminary Analysis Weight	Age; Gender; Household Type; Census Region; Population Density; Percentage Hispanic/Latino in Segment; Percentage Black/African American in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Owner Occupied in Segment; Employment Status; CBSA; Cigarette Lifetime Indicator; Imputation-Revised Education Level; Intermediate Imputed Hispanic/Latino Origin Indicator; Intermediate Imputed Race; Imputation-Revised Lifetime Indicators for Cigars, Chewing Tobacco, Pipes; Intermediate Lifetime Indicators for Snuff, Smokeless Tobacco, Alcohol, Heroin, Cocaine, Crack, Marijuana, PCP, LSD, Methamphetamine, Ecstasy, Inhalants, Tranquilizers, Stimulants, Sedatives, OxyContin, Hallucinogens, Pain Relievers, Other Pain Relievers, Other Stimulants, Other Hallucinogens; Preliminary Analysis Weight

CBSA = core-based statistical area.

[1] Respondents aged 12 to 14 were assigned a skip code and were not included in any imputation steps for marital status.

Table A.20 Model Summaries for IVEware, Demographics: 12 Years or Older

Variable	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycles 2–4	Variables Included in Predictive Mean Model, Cycle 5
Race	Age; Gender; Household Type; Census Region; Population Density; Percentage Hispanic/Latino in Segment; Percentage Black/African American in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Owner Occupied in Segment; Employment Status; CBSA; Cigarette Lifetime Indicator; Intermediate Imputed Education Level; Intermediate Imputed Hispanic/Latino Origin Indicator; Intermediate Imputed Marital Status; Intermediate Lifetime Indicators for Cigars, Chewing Tobacco, Pipes, Snuff, Smokeless Tobacco, Alcohol, Heroin, Cocaine, Crack, Marijuana, PCP, LSD, Methamphetamine, Ecstasy, Inhalants, Tranquilizers, Stimulants, Sedatives, OxyContin, Hallucinogens, Pain Relievers, Other Pain Relievers, Other Stimulants, Other Hallucinogens; Preliminary Analysis Weight	Age; Gender; Household Type; Census Region; Population Density; Percentage Hispanic/Latino in Segment; Percentage Black/African American in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Owner Occupied in Segment; Employment Status; CBSA; Cigarette Lifetime Indicator; Intermediate Imputed Education Level; Intermediate Imputed Hispanic/Latino Origin Indicator; Intermediate Imputed Marital Status; Intermediate Lifetime Indicators for Cigars, Chewing Tobacco, Pipes, Snuff, Smokeless Tobacco, Alcohol, Heroin, Cocaine, Crack, Marijuana, PCP, LSD, Methamphetamine, Ecstasy, Inhalants, Tranquilizers, Stimulants, Sedatives, OxyContin, Hallucinogens, Pain Relievers, Other Pain Relievers, Other Stimulants, Other Hallucinogens; Preliminary Analysis Weight	Age; Gender; Household Type; Census Region; Population Density; Percentage Hispanic/Latino in Segment; Percentage Black/African American in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Owner Occupied in Segment; Employment Status; CBSA; Cigarette Lifetime Indicator; Imputation-Revised Education Level; Imputation-Revised Hispanic/Latino Origin Indicator; Imputation-Revised Marital Status; Imputation-Revised Lifetime Indicators for Cigars, Chewing Tobacco, Pipes, Snuff, Smokeless Tobacco, Alcohol, Heroin, Cocaine, Crack, Marijuana, PCP, LSD, Methamphetamine, Ecstasy, Inhalants, Tranquilizers, Stimulants, Sedatives, OxyContin, Hallucinogens, Pain Relievers, Other Pain Relievers, Other Stimulants, Other Hallucinogens; Preliminary Analysis Weight

Table A.20 Model Summaries for IVEware, Demographics: 12 Years or Older (continued)

Variable	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycles 2–4	Variables Included in Predictive Mean Model, Cycle 5
Hispanic/Latino Origin	Age; Gender; Household Type; Census Region; Population Density; Percentage Hispanic/Latino in Segment; Percentage Black/African American in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Owner Occupied in Segment; Employment Status; CBSA; Cigarette Lifetime Indicator; Intermediate Imputed Education Level; Intermediate Imputed Marital Status; Intermediate Lifetime Indicators for Cigars, Chewing Tobacco, Pipes, Snuff, Smokeless Tobacco, Alcohol, Heroin, Cocaine, Crack, Marijuana, PCP, LSD, Methamphetamine, Ecstasy; Preliminary Analysis Weight	Age; Gender; Household Type; Census Region; Population Density; Percentage Hispanic/Latino in Segment; Percentage Black/African American in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Owner Occupied in Segment; Employment Status; CBSA; Cigarette Lifetime Indicator; Intermediate Imputed Education Level; Intermediate Imputed Race; Intermediate Imputed Marital Status; Intermediate Lifetime Indicators for Cigars, Chewing Tobacco, Pipes, Snuff, Smokeless Tobacco, Alcohol, Heroin, Cocaine, Crack, Marijuana, PCP, LSD, Methamphetamine, Ecstasy, Inhalants, Tranquilizers, Stimulants, Sedatives, OxyContin, Hallucinogens, Pain Relievers, Other Pain Relievers, Other Stimulants, Other Hallucinogens; Preliminary Analysis Weight	Age; Gender; Household Type; Census Region; Population Density; Percentage Hispanic/Latino in Segment; Percentage Black/African American in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Owner Occupied in Segment; Employment Status; CBSA; Cigarette Lifetime Indicator; Imputation-Revised Education Level; Intermediate Imputed Race; Imputation-Revised Marital Status; Imputation-Revised Lifetime Indicators for Cigars, Chewing Tobacco, Pipes, Snuff, Smokeless Tobacco, Alcohol, Heroin, Cocaine, Crack, Marijuana, PCP, LSD, Methamphetamine, Ecstasy; Intermediate Lifetime Indicators for Inhalants, Tranquilizers, Stimulants, Sedatives, OxyContin, Hallucinogens, Pain Relievers, Other Pain Relievers, Other Stimulants, Other Hallucinogens; Preliminary Analysis Weight

Table A.20 Model Summaries for IVEware, Demographics: 12 Years or Older (continued)

Variable	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycles 2–4	Variables Included in Predictive Mean Model, Cycle 5
Education Level[1]	Age; Gender; Household Type; Census Region; Population Density; Percentage Hispanic/Latino in Segment; Percentage Black/African American in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Owner Occupied in Segment; Employment Status; CBSA; Preliminary Analysis Weight	Age; Gender; Household Type; Census Region; Population Density; Percentage Hispanic/Latino in Segment; Percentage Black/African American in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Owner Occupied in Segment; Employment Status; CBSA; Cigarette Lifetime Indicator; Intermediate Imputed Hispanic/Latino Origin Indicator; Intermediate Imputed Race; Intermediate Imputed Marital Status; Intermediate Lifetime Indicators for Cigars, Chewing Tobacco, Pipes, Snuff, Smokeless Tobacco, Alcohol, Heroin, Cocaine, Crack, Marijuana, PCP, LSD, Methamphetamine, Ecstasy, Inhalants, Tranquilizers, Stimulants, Sedatives, OxyContin, Hallucinogens, Pain Relievers, Other Pain Relievers, Other Stimulants, Other Hallucinogens; Preliminary Analysis Weight	Age; Gender; Household Type; Census Region; Population Density; Percentage Hispanic/Latino in Segment; Percentage Black/African American in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Owner Occupied in Segment; Employment Status; CBSA; Cigarette Lifetime Indicator; Intermediate Imputed Hispanic/Latino Origin Indicator; Intermediate Imputed Race; Intermediate Imputed Marital Status; Intermediate Lifetime Indicators for Cigars, Chewing Tobacco, Pipes, Snuff, Smokeless Tobacco, Alcohol, Heroin, Cocaine, Crack, Marijuana, PCP, LSD, Methamphetamine, Ecstasy, Inhalants, Tranquilizers, Stimulants, Sedatives, OxyContin, Hallucinogens, Pain Relievers, Other Pain Relievers, Other Stimulants, Other Hallucinogens; Preliminary Analysis Weight

CBSA = core-based statistical area.

[1] Respondents aged 12 to 17 were assigned a skip code and were not included in any imputation steps for education level.

Table A.21 Model Summaries for IVEware, Marijuana: 12 Years or Older

Variable	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycles 2–4	Variables Included in Predictive Mean Model, Cycle 5
Lifetime	Age; Gender; Household Type; Census Region; Population Density; Percentage Hispanic/Latino in Segment; Percentage Black/African American in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Owner Occupied in Segment; Employment Status; CBSA; Cigarette Lifetime Indicator; Intermediate Imputed Education Level; Intermediate Imputed Marital Status; Intermediate Lifetime Indicators for Cigars, Chewing Tobacco, Pipes, Snuff, Smokeless Tobacco, Alcohol, Heroin, Cocaine, Crack; Preliminary Analysis Weight	Age; Gender; Household Type; Census Region; Population Density; Percentage Hispanic/Latino in Segment; Percentage Black/African American in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Owner Occupied in Segment; Employment Status; CBSA; Cigarette Lifetime Indicator; Intermediate Imputed Hispanic/Latino Origin Indicator; Intermediate Imputed Education Level; Intermediate Imputed Race; Intermediate Imputed Marital Status; Intermediate Lifetime Indicators for Cigars, Chewing Tobacco, Pipes, Snuff, Smokeless Tobacco, Alcohol, Heroin, Cocaine, Crack, PCP, LSD, Methamphetamine, Ecstasy, Inhalants, Tranquilizers, Stimulants, Sedatives, OxyContin, Hallucinogens, Pain Relievers, Other Pain Relievers, Other Stimulants, Other Hallucinogens; Preliminary Analysis Weight	Age; Gender; Household Type; Census Region; Population Density; Percentage Hispanic/Latino in Segment; Percentage Black/African American in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Owner Occupied in Segment; Employment Status; CBSA; Cigarette Lifetime Indicator; Intermediate Imputed Hispanic/Latino Origin Indicator; Imputation-Revised Education Level; Intermediate Imputed Race; Imputation-Revised Marital Status; Imputed Lifetime Indicators for Cigars, Chewing Tobacco, Pipes, Snuff, Smokeless Tobacco, Alcohol, Heroin, Cocaine, Crack; Intermediate Lifetime Indicators for PCP, LSD, Methamphetamine, Ecstasy, Inhalants, Tranquilizers, Stimulants, Sedatives, OxyContin, Hallucinogens, Pain Relievers, Other Pain Relievers, Other Stimulants, Other Hallucinogens; Preliminary Analysis Weight

Table A.21 Model Summaries for IVEware, Marijuana: 12 Years or Older (continued)

Variable	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycles 2–4	Variables Included in Predictive Mean Model, Cycle 5
Recency	Age; Gender; Census Region; CBSA; Imputation-Revised Hispanic/Latino Origin Indicator; Imputation-Revised Race; Imputation-Revised Marital Status; Imputation-Revised Education Level; Employment Status; Imputation-Revised Cigarette Recency; Imputation-Revised Lifetime Indicators for Inhalants, Tranquilizers, Sedatives, Cigars, Alcohol, Cocaine, Crack, Heroin, Pipes, Smokeless Tobacco, Hallucinogens, Pain Relievers, Stimulants; Age Squared; Age Cubed; Age * Gender; Race * Gender; Age * Race	Age; Gender; Census Region; CBSA; Imputation-Revised Hispanic/Latino Origin Indicator; Imputation-Revised Race; Imputation-Revised Marital Status; Imputation-Revised Education Level; Employment Status; Imputation-Revised Cigarette Recency; Imputation-Revised Lifetime Indicators for Inhalants, Tranquilizers, Sedatives, Cigars, Alcohol, Cocaine, Crack, Heroin, Pipes, Smokeless Tobacco, Hallucinogens, Pain Relievers, Stimulants; Age Squared; Age Cubed; Age * Gender; Race * Gender; Age * Race	Age; Gender; Census Region; CBSA; Imputation-Revised Hispanic/Latino Origin Indicator; Imputation-Revised Race; Imputation-Revised Marital Status; Imputation-Revised Education Level; Employment Status; Imputation-Revised Cigarette Recency; Imputation-Revised Lifetime Indicators for Inhalants, Tranquilizers, Sedatives, Cigars, Alcohol, Cocaine, Crack, Heroin, Pipes, Smokeless Tobacco, Hallucinogens, Pain Relievers, Stimulants; Age Squared; Age Cubed; Age * Gender; Race * Gender; Age * Race
12-Month Frequency	Age; Gender; Census Region; CBSA; Imputation-Revised Hispanic/Latino Origin Indicator; Imputation-Revised Race; Imputation-Revised Marital Status; Imputation-Revised Education Level; Employment Status; Imputation-Revised Cigarette Recency; Imputation-Revised Lifetime Indicators for Inhalants, Tranquilizers, Sedatives, Cigars, Alcohol, Cocaine, Crack, Heroin, Pipes, Smokeless Tobacco, Hallucinogens, Pain Relievers, Stimulants;	Age; Gender; Census Region; CBSA; Imputation-Revised Hispanic/Latino Origin Indicator; Imputation-Revised Race; Imputation-Revised Marital Status; Imputation-Revised Education Level; Employment Status; Imputation-Revised Cigarette Recency; Imputation-Revised Lifetime Indicators for Inhalants, Tranquilizers, Sedatives, Cigars, Alcohol, Cocaine, Crack, Heroin, Pipes, Smokeless Tobacco, Hallucinogens, Pain Relievers, Stimulants; Age Squared; Age Cubed;	Age; Gender; Census Region; CBSA; Imputation-Revised Hispanic/Latino Origin Indicator; Imputation-Revised Race; Imputation-Revised Marital Status; Imputation-Revised Education Level; Employment Status; Imputation-Revised Cigarette Recency; Imputation-Revised Lifetime Indicators for Inhalants, Tranquilizers, Sedatives, Cigars, Alcohol, Cocaine, Crack, Heroin, Pipes, Smokeless Tobacco, Hallucinogens,

Table A.21 Model Summaries for IVEware, Marijuana: 12 Years or Older (continued)

Variable	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycles 2–4	Variables Included in Predictive Mean Model, Cycle 5
12-Month Frequency (continued)	Age Squared; Age Cubed; Age * Gender; Race * Gender; Age * Race	Age * Gender; Race * Gender; Age * Race	Pain Relievers, Stimulants; Age Squared; Age Cubed; Age * Gender; Race * Gender; Age * Race
30-Day Frequency	Age; Gender; Census Region; CBSA; Imputation-Revised Hispanic/Latino Origin Indicator; Imputation-Revised Race; Imputation-Revised Marital Status; Imputation-Revised Education Level; Employment Status; Imputation-Revised Cigarette Recency; Imputation-Revised Lifetime Indicators for Inhalants, Tranquilizers, Sedatives, Cigars, Alcohol, Cocaine, Crack, Heroin, Pipes, Smokeless Tobacco, Hallucinogens, Pain Relievers, Stimulants; Age Squared; Age Cubed; Age * Gender; Race * Gender; Age * Race; Imputation-Revised Marijuana 12-Month Frequency	Age; Gender; Census Region; CBSA; Imputation-Revised Hispanic/Latino Origin Indicator; Imputation-Revised Race; Imputation-Revised Marital Status; Imputation-Revised Education Level; Employment Status; Imputation-Revised Cigarette Recency; Imputation-Revised Lifetime Indicators for Inhalants, Tranquilizers, Sedatives, Cigars, Alcohol, Cocaine, Crack, Heroin, Pipes, Smokeless Tobacco, Hallucinogens, Pain Relievers, Stimulants; Age Squared; Age Cubed; Age * Gender; Race * Gender; Age * Race; Imputation-Revised Marijuana 12-Month Frequency	Age; Gender; Census Region; CBSA; Imputation-Revised Hispanic/Latino Origin Indicator; Imputation-Revised Race; Imputation-Revised Marital Status; Imputation-Revised Education Level; Employment Status; Imputation-Revised Cigarette Recency; Imputation-Revised Lifetime Indicators for Inhalants, Tranquilizers, Sedatives, Cigars, Alcohol, Cocaine, Crack, Heroin, Pipes, Smokeless Tobacco, Hallucinogens, Pain Relievers, Stimulants; Age Squared; Age Cubed; Age * Gender; Race * Gender; Age * Race; Imputation-Revised Marijuana 12-Month Frequency
Age at First Use	Age; Gender; Census Region; CBSA; Imputation-Revised Hispanic/Latino Origin Indicator; Imputation-Revised Race; Imputation-Revised Marital Status; Imputation-Revised Education Level; Employment Status; Imputation-Revised Cigarette Recency;	Age; Gender; Census Region; CBSA; Imputation-Revised Hispanic/Latino Origin Indicator; Imputation-Revised Race; Imputation-Revised Marital Status; Imputation-Revised Education Level; Employment Status; Imputation-Revised Cigarette Recency; Imputation-Revised	Age; Gender; Census Region; CBSA; Imputation-Revised Hispanic/Latino Origin Indicator; Imputation-Revised Race; Imputation-Revised Marital Status; Imputation-Revised Education Level; Employment Status; Imputation-Revised

Table A.21 Model Summaries for IVEware, Marijuana: 12 Years or Older (continued)

Variable	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycles 2–4	Variables Included in Predictive Mean Model, Cycle 5
Age at First Use (continued)	Imputation-Revised Lifetime Indicators for Inhalants, Tranquilizers, Sedatives, Cigars, Alcohol, Cocaine, Crack, Heroin, Pipes, Smokeless Tobacco, Hallucinogens, Pain Relievers, Stimulants; Age Squared; Age Cubed; Age * Gender; Race * Gender; Age * Race; Imputation-Revised Marijuana 12-Month Frequency; Imputation-Revised Marijuana 30-Day Frequency; Imputation-Revised Marijuana Recency; Imputation-Revised Cigarette Age at First Use; Imputation-Revised Alcohol Age at First Use	Lifetime Indicators for Inhalants, Tranquilizers, Sedatives, Cigars, Alcohol, Cocaine, Crack, Heroin, Pipes, Smokeless Tobacco, Hallucinogens, Pain Relievers, Stimulants; Age Squared; Age Cubed; Age * Gender; Race * Gender; Age * Race; Imputation-Revised Marijuana 12-Month Frequency; Imputation-Revised Marijuana 30-Day Frequency; Imputation-Revised Marijuana Recency; Imputation-Revised Cigarette Age at First Use; Imputation-Revised Alcohol Age at First Use	Cigarette Recency; Imputation-Revised Lifetime Indicators for Inhalants, Tranquilizers, Sedatives, Cigars, Alcohol, Cocaine, Crack, Heroin, Pipes, Smokeless Tobacco, Hallucinogens, Pain Relievers, Stimulants; Age Squared; Age Cubed; Age * Gender; Race * Gender; Age * Race; Imputation-Revised Marijuana 12-Month Frequency; Imputation-Revised Marijuana 30-Day Frequency; Imputation-Revised Marijuana Recency; Imputation-Revised Cigarette Age at First Use; Imputation-Revised Alcohol Age at First Use

CBSA = core-based statistical area.

Note: An asterisk "*" represents an interaction between two variables.

A.4 Model Summaries for the Modified Predictive Mean Neighborhood Multiple Imputation Method

Table A.22 Model Summaries for modPMN-MI, Demographics: 15 Years or Older

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model
Marital Status[1]	Census Region; Gender; Population Density; Age Category; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Age Category * Gender	Age; Percentage Black/African American in Segment; Percentage Owner Occupied in Segment; Gender; Age * Gender; Census Region; Population Density; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment

modPMN-MI = modified predictive mean neighborhood multiple imputation.

Note: An asterisk "*" represents an interaction between two variables.

[1] Respondents aged 12 to 14 were assigned a skip code and were not included in any imputation steps for marital status.

Table A.23 Model Summaries for modPMN-MI, Demographics: 12 to 17 Years

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model
Race	Census Region; Household Type; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment	Census Region; Household Type; Age; Age Squared; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment
Hispanic/Latino Origin	Census Region; Imputed Race; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment	Census Region; Imputed Race; Household Type; Age; Age Squared; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment
Education Level[1]	No model used: no nonrespondents	No model used: no nonrespondents

modPMN-MI = modified predictive mean neighborhood multiple imputation.

Note: An asterisk "*" represents an interaction between two variables.

[1] Respondents aged 12 to 17 were assigned a skip code and were not included in any imputation steps for education level.

Table A.24 Model Summaries for modPMN-MI, Demographics: 18 to 25 Years

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model
Race	Census Region; Household Type; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Imputed Marital Status	Census Region; Household Type; Age; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Imputed Marital Status
Hispanic/Latino Origin	Census Region; Imputed Race; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment	Census Region; Imputed Race; Household Type; Age; Age Squared; Age Cubed; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Imputed Marital Status
Education Level	Census Region; Imputed Race; Gender; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment	Census Region; Imputed Race; Gender; Age; Age * Gender; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Imputed Marital Status

modPMN-MI = modified predictive mean neighborhood multiple imputation.

Note: An asterisk "*" represents an interaction between two variables.

Table A.25 Model Summaries for modPMN-MI, Demographics: 26 Years or Older

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model
Race	Census Region; Household Type; Age Category; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Imputed Marital Status; Age	Census Region; Household Type; Age Category; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Imputed Marital Status
Hispanic/Latino Origin	Census Region; Imputed Race; Age Category; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment	Household Type; Age; Age Squared; Age Cubed; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Imputed Marital Status
Education Level	Census Region; Imputed Race; Gender; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Percentage Owner Occupied in Segment	Census Region; Imputed Race; Gender; Age; Age * Gender; Percentage Black/African American in Segment; Percentage American Indian/Alaska Native in Segment; Percentage Asian/Other Pacific Islander in Segment; Percentage Hispanic/Latino in Segment; Imputed Marital Status

modPMN-MI = modified predictive mean neighborhood multiple imputation.

Note: An asterisk "*" represents an interaction between two variables.

Table A.26 Model Summaries for modPMN-MI, Marijuana: 12 to 17 Years

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycle 2
Lifetime[1]	Gender; Race; Gender * Race; CBSA; Census Region; Cigarette Lifetime Indicator	Cigarette Lifetime Indicator; Intermediate Lifetime Indicators for Alcohol and Inhalants; Age; Gender; Race; Age Squared; Age Cubed; Gender * Race; Age * Gender; Age * Race; State Rank; CBSA; Census Region	Cigarette Lifetime Indicator; Age; Gender; Race; Age Squared; Age Cubed; Gender * Race; Age * Gender; Age * Race; State Rank; CBSA; Census Region; Intermediate Lifetime Indicators for Pain Relievers, Cocaine, and Heroin; Imputed Lifetime Indicators for Alcohol and Inhalants
Recency[2]	Gender; Race; Gender * Race; CBSA; Census Region; Cigarette Lifetime Indicator; Imputed Lifetime Indicators for Alcohol, Inhalants, Marijuana, Pain Relievers, Cocaine, and Heroin	Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Census Region; CBSA; State Rank; Imputed Recencies for Cigarettes, Alcohol, Inhalants; Imputed Lifetime Indicators for Pain Relievers, Cocaine, and Heroin	Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Census Region; CBSA; State Rank; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Pain Relievers, and Cocaine
12-Month Frequency[3]	Imputed Recencies for Alcohol, Inhalants, Marijuana, Pain Relievers, and Cocaine; Race; Gender; Census Region; CBSA; Imputed Recency for Alcohol * Race; Imputed Recency for Inhalants * Race; Imputed Recency for Marijuana * Race; Imputed Recency for Pain Relievers * Race; Imputed Recency for Cocaine * Race; Imputed Recency for Alcohol * Gender; Imputed Recency for Inhalants * Gender; Imputed Recency for Marijuana * Gender; Imputed Recency for Pain Relievers * Gender; Imputed Recency for Cocaine * Gender; Imputed Recency for Alcohol * Census Region; Imputed Recency for Inhalants * Census Region; Imputed Recency for Marijuana * Census Region;	Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Census Region; CBSA; State Rank; Imputed Recencies for Cigarettes, Alcohol, Inhalants Pain Relievers, Cocaine, and Heroin; Intermediate 12-Month Frequencies for Alcohol and Inhalants	Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Census Region; CBSA; State Rank; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Pain Relievers, Cocaine, and Heroin; Imputed 12-Month Frequencies for Alcohol and Inhalants; Intermediate 12-Month Frequencies for Pain Relievers, Cocaine, and Heroin

Table A.26 Model Summaries for modPMN-MI, Marijuana: 12 to 17 Years (continued)

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycle 2
12-Month Frequency[3] (continued)	Imputed Recency for Pain Relievers * Census Region; Imputed Recency for Cocaine * Census Region; Imputed Recency for Alcohol * CBSA; Imputed Recency for Inhalants * CBSA; Imputed Recency for Marijuana * CBSA; Imputed Recency for Pain Relievers * CBSA; Imputed Recency for Cocaine * CBSA		
30-Day Frequency[4]	Imputed Recencies for Cigarettes, Alcohol, and Marijuana; Race; Gender; Census Region; CBSA; Imputed Recency for Cigarettes * Race; Imputed Recency for Alcohol * Race; Imputed Recency for Marijuana * Race; Imputed Recency for Cigarettes * Gender; Imputed Recency for Alcohol * Gender; Imputed Recency for Marijuana * Gender; Imputed Recency for Cigarettes * Census Region; Imputed Recency for Alcohol * Census Region; Imputed Recency for Marijuana * Census Region; Imputed Recency for Cigarettes * CBSA; Imputed Recency for Alcohol * CBSA; Imputed Recency for Marijuana * CBSA	Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Census Region; CBSA; State Rank; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Pain Relievers, Cocaine, and Heroin; Imputed 12-Month Frequency for Marijuana; Intermediate 30-Day Frequencies for Cigarettes, Alcohol, and Inhalants	Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Census Region; CBSA; State Rank; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Pain Relievers, Cocaine, and Heroin; Imputed 12-Month Frequency for Marijuana; Imputed 30-Day Frequencies for Cigarettes, Alcohol, and Inhalants; Intermediate 30-Day Frequencies for Cocaine and Heroin
Age at First Use[5]	Cigarette Lifetime Indicator; Imputed Lifetime Indicators for Alcohol, Inhalants, Marijuana, Pain Relievers, Cocaine, and Heroin; Gender; Race; Census Region; CBSA; State Rank; Education Level; Employment Status; Marital Status; Cigarette Lifetime * Race; Imputed Lifetime Alcohol * Race; Imputed Lifetime Inhalants * Race; Imputed Lifetime Marijuana * Race; Imputed Lifetime Pain Relievers * Race; Imputed Lifetime Cocaine * Race; Imputed Lifetime Heroin * Race; Cigarette Lifetime * Gender; Imputed Lifetime Alcohol * Gender; Imputed Lifetime Inhalants * Gender; Imputed Lifetime Marijuana * Gender;	Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Age Squared * Gender; Age Squared * Race; Census Region; CBSA; State Rank; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Marijuana, Pain Relievers, Cocaine, and Heroin; Imputed 12-Month Frequency for Marijuana;	Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Age Squared * Gender; Age Squared * Race; Census Region; CBSA; State Rank; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Marijuana, Pain Relievers, Cocaine, and Heroin; Imputed 12-Month Frequency for Marijuana;

Table A.26 Model Summaries for modPMN-MI, Marijuana: 12 to 17 Years (continued)

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycle 2
Age at First Use[5] (continued)	Imputed Lifetime Pain Relievers * Gender; Imputed Lifetime Cocaine * Gender; Imputed Lifetime Heroin * Gender; Cigarette Lifetime * Census Region; Imputed Lifetime Alcohol * Census Region; Imputed Lifetime Inhalants * Census Region; Imputed Lifetime Marijuana * Census Region; Imputed Lifetime Pain Relievers * Census Region; Imputed Lifetime Cocaine * Census Region; Imputed Lifetime Heroin * Census Region; Cigarette Lifetime * CBSA; Imputed Lifetime Alcohol * CBSA; Imputed Lifetime Inhalants * CBSA; Imputed Lifetime Marijuana * CBSA; Imputed Lifetime Pain Relievers * CBSA; Imputed Lifetime Cocaine * CBSA; Imputed Lifetime Heroin * CBSA	Imputed 30-Day Frequency for Marijuana; Intermediate Age at First Use for Cigarettes, Alcohol, and Inhalants	Imputed 30-Day Frequency for Marijuana; Imputed Age at First Use for Cigarettes, Alcohol, and Inhalants; Intermediate Age at First Use for Pain Relievers, Cocaine, and Heroin

CBSA = core-based statistical area; modPMN-MI = modified predictive mean neighborhood multiple imputation.

Note: An asterisk "*" represents an interaction between two variables.

[1] A single response propensity model was fit for all drugs within the age group.

[2] A single response propensity model was fit for all drugs within the age group.

[3] A single response propensity model was fit for alcohol, inhalants, marijuana, pain relievers, and cocaine within the age group.

[4] A single response propensity model was fit for cigarettes, alcohol, and marijuana within the age group.

[5] A single response propensity model was fit for all drugs within the age group.

Table A.27 Model Summaries for modPMN-MI, Marijuana: 18 to 25 Years

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycle 2
Lifetime[1]	Gender; Race; Gender * Race; Marital Status; Education Level; Employment Status; CBSA; Census Region; Cigarette Lifetime Indicator	Cigarette Lifetime Indicator; Age; Gender; Race; Age Squared; Age Cubed; Gender * Race; Age * Gender; Age * Race; State Rank; CBSA; Census Region; Education Level; Employment Status; Marital Status; Intermediate Lifetime Indicators for Alcohol and Inhalants	Cigarette Lifetime Indicator; Age; Gender; Race; Age Squared; Age Cubed; Gender * Race; Gender * Age; Age * Race; State Rank; CBSA; Census Region; Education Level; Employment Status; Marital Status; Intermediate Lifetime Indicators for Pain Relievers, Cocaine, and Heroin; Imputed Lifetime Indicators for Alcohol and Inhalants
Recency[2]	Gender; Race; Gender * Race; Marital Status; Education Level; Employment Status; CBSA; Census Region; Cigarette Lifetime Indicator; Imputed Recencies for Cigarettes and Alcohol; Imputed Lifetime Indicators for Alcohol, Inhalants, Marijuana, Pain Relievers, Cocaine, and Heroin	Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Marital Status; Education Level; Employment Status; Census Region; CBSA; State Rank; Imputed Recencies for Cigarettes, Alcohol, Inhalants; Imputed Lifetime Indicators for Pain Relievers, Cocaine, and Heroin	Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Marital Status; Education Level; Employment Status; Census Region; CBSA; State Rank; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Pain Relievers, Cocaine, and Heroin
12-Month Frequency[3]	Imputed Recencies for Alcohol, Inhalants, Marijuana, Pain Relievers, and Cocaine; Race; Gender; Census Region; CBSA; Education Level; Employment Status; Marital Status; Imputed Recency for Alcohol * Race; Imputed Recency for Inhalants * Race; Imputed Recency for Marijuana * Race; Imputed Recency for Pain Relievers * Race; Imputed Recency for Cocaine * Race; Imputed Recency for Alcohol * Gender; Imputed Recency for Inhalants * Gender; Imputed Recency for Marijuana *	Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Census Region; CBSA; State Rank; Education Level; Employment Status; Marital Status; Imputed Recencies for Cigarettes, Alcohol, Inhalants Pain Relievers, Cocaine, and Heroin;	Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Census Region; CBSA; State Rank; Education Level; Employment Status; Marital Status; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Marijuana, Pain Relievers, Cocaine, and Heroin;

Table A.27 Model Summaries for modPMN-MI, Marijuana: 18 to 25 Years (continued)

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycle 2
12-Month Frequency[3] (continued)	Gender; Imputed Recency for Pain Relievers * Gender; Imputed Recency for Cocaine * Gender; Imputed Recency for Alcohol * Census Region; Imputed Recency for Inhalants * Census Region; Imputed Recency for Marijuana * Census Region; Imputed Recency for Pain Relievers * Census Region; Imputed Recency for Cocaine * Census Region; Imputed Recency for Alcohol * CBSA; Imputed Recency for Inhalants * CBSA; Imputed Recency for Marijuana * CBSA; Imputed Recency for Pain Relievers * CBSA; Imputed Recency for Cocaine * CBSA; Imputed Recency for Alcohol * Education Level; Imputed Recency for Inhalants * Education Level; Imputed Recency for Marijuana * Education Level; Imputed Recency for Pain Relievers * Education Level; Imputed Recency for Cocaine * Education Level; Imputed Recency for Alcohol * Employment Status; Imputed Recency for Inhalants * Employment Status; Imputed Recency for Marijuana * Employment Status; Imputed Recency for Pain Relievers * Employment Status; Imputed Recency for Cocaine * Employment Status; Imputed Recency for Alcohol * Marital Status; Imputed Recency for Inhalants * Marital Status; Imputed Recency for Marijuana * Marital Status; Imputed Recency for Pain Relievers * Marital Status; Imputed Recency for Cocaine * Marital Status	Intermediate 12-Month Frequencies for Alcohol and Inhalants	Imputed 12-Month Frequencies for Alcohol and Inhalants; Intermediate 12-Month Frequencies for Pain Relievers, Cocaine, and Heroin
30-Day Frequency[4]	Imputed Recencies for Cigarettes, Alcohol, and Marijuana; Race; Gender; Census Region; CBSA; Education Level; Employment Status; Marital Status; Imputed Recency for Cigarettes * Race; Imputed Recency for Alcohol * Race;	Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Census Region; CBSA; State Rank; Education Level;	Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Census Region; CBSA; State Rank; Education Level;

Table A.27 Model Summaries for modPMN-MI, Marijuana: 18 to 25 Years (continued)

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycle 2
30-Day Frequency[4] (continued)	Imputed Recency for Marijuana * Race; Imputed Recency for Cigarettes * Gender; Imputed Recency for Alcohol * Gender; Imputed Recency for Marijuana * Gender; Imputed Recency for Cigarettes * Census Region; Imputed Recency for Alcohol * Census Region; Imputed Recency for Marijuana * Census Region; Imputed Recency for Cigarettes * CBSA; Imputed Recency for Alcohol * CBSA; Imputed Recency for Marijuana * CBSA; Imputed Recency for Cigarettes * Education Level; Imputed Recency for Alcohol * Education Level; Imputed Recency for Marijuana * Education Level; Imputed Recency for Cigarettes * Employment Status; Imputed Recency for Alcohol * Employment Status; Imputed Recency for Marijuana * Employment Status; Imputed Recency for Cigarettes * Marital Status; Imputed Recency for Alcohol * Marital Status; Imputed Recency for Marijuana * Marital Status	Employment Status; Marital Status; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Pain Relievers, Cocaine, and Heroin; Imputed 12-Month Frequency for Marijuana; Intermediate 30-Day Frequencies for Cigarettes, Alcohol, and Inhalants	Employment Status; Marital Status; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Pain Relievers, Cocaine, and Heroin; Imputed 12-Month Frequency for Marijuana; Imputed 30-Day Frequencies for Cigarettes, Alcohol, and Inhalants; Intermediate 30-Day Frequencies for Cocaine and Heroin
Age at First Use[5]	Cigarette Lifetime Indicator; Imputed Lifetime Indicators for Alcohol, Inhalants, Marijuana, Pain Relievers, Cocaine, and Heroin; Gender; Race; Census Region; CBSA; State Rank; Education Level; Employment Status; Marital Status; Cigarette Lifetime * Race; Imputed Lifetime Alcohol * Race; Imputed Lifetime Inhalants * Race; Imputed Lifetime Marijuana * Race; Imputed Lifetime Pain Relievers * Race; Imputed Lifetime Cocaine * Race; Cigarette Lifetime * Gender; Imputed Lifetime Alcohol * Gender; Imputed Lifetime Inhalants * Gender; Imputed Lifetime Marijuana * Gender; Imputed Lifetime Pain Relievers * Gender; Imputed Lifetime Cocaine * Gender; Imputed Lifetime Heroin * Gender; Cigarette Lifetime * Census	Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Age Squared * Gender; Age Squared * Race; Census Region; CBSA; State Rank; Education Level; Employment Status; Marital Status; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Marijuana, Pain Relievers, Cocaine, and Heroin; Imputed 12-Month Frequency for Marijuana; Imputed 30-Day Frequency	Age; Age Squared; Age Cubed; Gender; Race; Gender * Race; Age * Gender; Age * Race; Age Squared * Gender; Age Squared * Race; Census Region; CBSA; State Rank; Education Level; Employment Status; Marital Status; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Marijuana, Pain Relievers, Cocaine, and Heroin; Imputed 12-Month Frequency for Marijuana; Imputed 30-Day Frequency

Table A.27 Model Summaries for modPMN-MI, Marijuana: 18 to 25 Years (continued)

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycle 2
Age at First Use[5] (continued)	Region; Imputed Lifetime Alcohol * Census Region; Imputed Lifetime Inhalants * Census Region; Imputed Lifetime Marijuana * Census Region; Imputed Lifetime Pain Relievers * Census Region; Imputed Lifetime Cocaine * Census Region; Cigarette Lifetime * CBSA; Imputed Lifetime Alcohol * CBSA; Imputed Lifetime Inhalants * CBSA; Imputed Lifetime Marijuana * CBSA; Imputed Lifetime Pain Relievers * CBSA; Imputed Lifetime Cocaine * CBSA; Cigarette Lifetime * Education Level; Imputed Lifetime Alcohol * Education Level; Imputed Lifetime Inhalants * Education Level; Imputed Lifetime Marijuana * Education Level; Imputed Lifetime Pain Relievers * Education Level; Imputed Lifetime Cocaine * Education Level; Cigarette Lifetime * Employment Status; Imputed Lifetime Alcohol * Employment Status; Imputed Lifetime Inhalants * Employment Status; Imputed Lifetime Marijuana * Employment Status; Imputed Lifetime Pain Relievers * Employment Status; Imputed Lifetime Cocaine * Employment Status; Cigarette Lifetime * Marital Status; Imputed Lifetime Alcohol * Marital Status; Imputed Lifetime Inhalants * Marital Status; Imputed Lifetime Marijuana * Marital Status; Imputed Lifetime Pain Relievers * Marital Status; Imputed Lifetime Cocaine * Marital Status	for Marijuana; Intermediate Age at First Use for Cigarettes, Alcohol, and Inhalants	for Marijuana; Imputed Age at First Use for Cigarettes, Alcohol, and Inhalants; Intermediate Age at First Use for Pain Relievers, Cocaine, and Heroin

CBSA = core-based statistical area; modPMN-MI = modified predictive mean neighborhood multiple imputation.

Note: An asterisk "*" represents an interaction between two variables.

[1] A single response propensity model was fit for all drugs within the age group.

[2] A single response propensity model was fit for all drugs within the age group.

[3] A single response propensity model was fit for alcohol, inhalants, marijuana, pain relievers, and cocaine within the age group.

[4] A single response propensity model was fit for cigarettes, alcohol, and marijuana within the age group.

[5] A single response propensity model was fit for all drugs within the age group.

Table A.28 Model Summaries for modPMN-MI, Marijuana: 26 Years or Older

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycle 2
Lifetime[1]	Age Category; Gender; Race; Gender * Race; Marital Status; Education Level; Employment Status; CBSA; Census Region; Cigarette Lifetime Indicator	Cigarette Lifetime Indicator; Age; Gender; Race; Age Squared; Age Cubed; Gender * Race; Gender * Age; Age * Race; State Rank; CBSA; Census Region; Education Level; Employment Status; Marital Status; Intermediate Lifetime Indicators for Alcohol and Inhalants	Cigarette Lifetime Indicator; Age; Gender; Race; Age Squared; Age Cubed; Gender * Race; Age * Gender; Age * Race; State Rank; CBSA; Census Region; Education Level; Employment Status; Marital Status; Intermediate Lifetime Indicators for Pain Relievers, Cocaine, and Heroin; Imputed Lifetime Indicators for Alcohol and Inhalants
Recency[2]	Gender; Race; Gender * Race; CBSA; Census Region; Cigarette Lifetime Indicator; Imputed Recencies for Cigarettes and Alcohol; Imputed Lifetime Indicators for Alcohol, Inhalants, Marijuana, Pain Relievers, Tranquilizers, Cocaine, and Heroin	Age; Gender; Race; Gender * Race; Age * Gender; Age * Race; Marital Status; Education Level; Employment Status; Census Region; CBSA; State Rank; Imputed Recencies for Cigarettes, Alcohol, and Inhalants; Imputed Lifetime Indicators for Pain Relievers, Cocaine, and Heroin	Age; Gender; Race; Gender * Race; Age * Gender; Age * Race; Marital Status; Education Level; Employment Status; Census Region; CBSA; State Rank; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Pain Relievers, Cocaine, and Heroin
12-Month Frequency[3]	Imputed Recencies for Alcohol, Inhalants, Marijuana, and Pain Relievers; Race; Gender; Census Region; CBSA; Education Level; Employment Status; Marital Status; Imputed Recency for Alcohol * Race; Imputed Recency for Inhalants * Race; Imputed Recency for Marijuana * Race; Imputed Recency for Pain Relievers * Race; Imputed Recency for Alcohol * Gender; Imputed Recency for Inhalants * Gender; Imputed Recency for Marijuana * Gender;	Age; Gender; Race; Gender * Race; Age * Gender; Age * Race; Census Region; CBSA; State Rank; Education Level; Employment Status; Marital Status; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Marijuana, Pain Relievers, Cocaine, and Heroin; Intermediate 12-Month	Age; Gender; Race; Gender * Race; Age * Gender; Age * Race; Census Region; CBSA; State Rank; Education Level; Employment Status; Marital Status; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Marijuana, Pain Relievers, Cocaine, and Heroin; Imputed 12-Month

Table A.28 Model Summaries for modPMN-MI, Marijuana: 26 Years or Older (continued)

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycle 2
12-Month Frequency[3] (continued)	Imputed Recency for Pain Relievers * Gender; Imputed Recency for Alcohol * Census Region; Imputed Recency for Inhalants * Census Region; Imputed Recency for Marijuana * Census Region; Imputed Recency for Pain Relievers * Census Region; Imputed Recency for Alcohol * CBSA; Imputed Recency for Marijuana * CBSA; Imputed Recency for Pain Relievers * CBSA; Imputed Recency for Cocaine * CBSA; Imputed Recency for Alcohol * Marital Status; Imputed Recency for Marijuana * Marital Status; Imputed Recency for Pain Relievers * Marital Status; Imputed Recency for Cocaine * Marital Status; Imputed Recency for Alcohol * Age; Imputed Recency for Inhalants * Age; Imputed Recency for Marijuana * Age; Imputed Recency for Pain Relievers * Age; Imputed Recency for Cocaine * Age	Frequencies for Alcohol and Inhalants	Frequencies for Alcohol and Inhalants; Intermediate 12-Month Frequencies for Pain Relievers, Cocaine, and Heroin
30-Day Frequency[4]	Imputed Recencies for Cigarettes, Alcohol, and Marijuana; Race; Gender; Census Region; CBSA; Education Level; Employment Status; Marital Status; Imputed Recency for Cigarettes * Race; Imputed Recency for Marijuana * Race; Imputed Recency for Cigarettes * Gender; Imputed Recency for Alcohol * Gender; Imputed Recency for Marijuana * Gender; Imputed Recency for Cigarettes * Census Region; Imputed Recency for Alcohol * Census Region; Imputed Recency for Marijuana * Census Region; Imputed Recency for Cigarettes * CBSA; Imputed Recency for Alcohol * CBSA; Imputed Recency for Marijuana * CBSA; Imputed Recency for Cigarettes * Education Level;	Age; Gender; Race; Gender * Race; Age * Gender; Age * Race; Census Region; CBSA; State Rank; Education Level; Employment Status; Marital Status; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Pain Relievers, Cocaine, and Heroin; Imputed 12-Month Frequency for Marijuana; Intermediate 30-Day Frequencies for Cigarettes, Alcohol, and Inhalants	Age; Gender; Race; Gender * Race; Age * Gender; Age * Race; Census Region; CBSA; State Rank; Education Level; Employment Status; Marital Status; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Pain Relievers, Cocaine, and Heroin; Imputed 12-Month Frequency for Marijuana; Imputed 30-Day Frequencies for Cigarettes, Alcohol, and Inhalants; Intermediate 30-Day Frequencies for Cocaine and Heroin

Table A.28 Model Summaries for modPMN-MI, Marijuana: 26 Years or Older (continued)

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycle 2
30-Day Frequency[4] (continued)	Imputed Recency for Alcohol * Education Level; Imputed Recency for Marijuana * Education Level; Imputed Recency for Cigarettes * Employment Status; Imputed Recency for Alcohol * Employment Status; Imputed Recency for Marijuana * Employment Status; Imputed Recency for Cigarettes * Marital Status; Imputed Recency for Alcohol * Marital Status		
Age at First Use[5]	Cigarette Lifetime Indicator; Imputed Lifetime Indicators for Alcohol, Inhalants, Marijuana, Pain Relievers, Cocaine, and Heroin; Gender; Race; Census Region; CBSA; State Rank; Education Level; Employment Status; Marital Status; Cigarette Lifetime * Race; Imputed Lifetime Alcohol * Race; Imputed Lifetime Inhalants * Race; Imputed Lifetime Marijuana * Race; Imputed Lifetime Pain Relievers * Race; Imputed Lifetime Cocaine * Race; Imputed Lifetime Heroin * Race; Cigarette Lifetime * Gender; Imputed Lifetime Alcohol * Gender; Imputed Lifetime Inhalants * Gender; Imputed Lifetime Marijuana * Gender; Imputed Lifetime Pain Relievers * Gender; Imputed Lifetime Cocaine * Gender; Imputed Lifetime Heroin * Gender; Cigarette Lifetime * Census Region; Imputed Lifetime Alcohol * Census Region; Imputed Lifetime Inhalants * Census Region; Imputed Lifetime Marijuana * Census Region; Imputed Lifetime Pain Relievers * Census Region; Imputed Lifetime Cocaine * Census Region; Imputed Lifetime Heroin * Census Region; Cigarette Lifetime * CBSA; Imputed Lifetime Alcohol * CBSA; Imputed Lifetime Inhalants * CBSA;	Age; Gender; Race; Gender * Race; Age * Gender; Age * Race; Census Region; CBSA; State Rank; Education Level; Employment Status; Marital Status; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Marijuana, Pain Relievers, Cocaine, and Heroin; Imputed 12-Month Frequency for Marijuana; Imputed 30-Day Frequency for Marijuana; Intermediate Age at First Use for Cigarettes, Alcohol, and Inhalants	Age; Gender; Race; Gender * Race; Age * Gender; Age * Race; Census Region; CBSA; State Rank; Education Level; Employment Status; Marital Status; Imputed Recencies for Cigarettes, Alcohol, Inhalants, Marijuana, Pain Relievers, Cocaine, and Heroin; Imputed 12-Month Frequency for Marijuana; Imputed 30-Day Frequency for Marijuana; Imputed Age at First Use for Cigarettes, Alcohol, and Inhalants; Intermediate Age at First Use for Pain Relievers, Cocaine, and Heroin

Table A.28 Model Summaries for modPMN-MI, Marijuana: 26 Years or Older (continued)

Variable	Variables Included in Response Propensity Model	Variables Included in Predictive Mean Model, Cycle 1	Variables Included in Predictive Mean Model, Cycle 2
Age at First Use[5] (continued)	Imputed Lifetime Marijuana * CBSA; Imputed Lifetime Pain Relievers * CBSA; Imputed Lifetime Cocaine * CBSA; Imputed Lifetime Heroin * CBSA; Cigarette Lifetime * Education Level; Imputed Lifetime Alcohol * Education Level; Imputed Lifetime Inhalants * Education Level; Imputed Lifetime Marijuana * Education Level; Imputed Lifetime Pain Relievers * Education Level; Imputed Lifetime Cocaine * Education Level; Imputed Lifetime Heroin * Education Level; Cigarette Lifetime * Employment Status; Imputed Lifetime Alcohol * Employment Status; Imputed Lifetime Inhalants * Employment Status; Imputed Lifetime Marijuana * Employment Status; Imputed Lifetime Pain Relievers * Employment Status; Imputed Lifetime Cocaine * Employment Status; Imputed Lifetime Heroin * Employment Status; Cigarette Lifetime * Marital Status; Imputed Lifetime Alcohol * Marital Status; Imputed Lifetime Inhalants * Marital Status; Imputed Lifetime Marijuana * Marital Status; Imputed Lifetime Pain Relievers * Marital Status; Imputed Lifetime Cocaine * Marital Status; Imputed Lifetime Heroin * Marital Status		

CBSA = core-based statistical area; modPMN-MI = modified predictive mean neighborhood multiple imputation.

Note: An asterisk "*" represents an interaction between two variables.

[1] A single response propensity model was fit for all drugs within the age group.

[2] A single response propensity model was fit for all drugs within the age group.

[3] A single response propensity model was fit for alcohol, inhalants, marijuana, pain relievers, and cocaine within the age group.

[4] A single response propensity model was fit for cigarettes, alcohol, and marijuana within the age group.

[5] A single response propensity model was fit for all drugs within the age group.

Appendix B: Multiple Imputation Results

This page intentionally left blank

This appendix describes the methodology for assessing the increase in variance due to imputation through the use of multiple imputation (MI). A benefit of MI is the introduction of a random mechanism that under ideal conditions can account for the additional uncertainty produced when imputing missing values. Suppose that the primary interest is in determining a point estimate such as a mean, a proportion, or a regression coefficient. The combined point estimate is the average of the point estimates obtained from the *m* datasets where *m* is the number of completed datasets generated by the MI procedure. The analysis of the *m* completed datasets resulting from MI proceeds as follows: (1) analyze each of the *m* completed datasets separately, (2) extract the point estimate and the estimated standard error from each analysis, and (3) combine the point estimates and the estimated standard errors to arrive at a single point estimate, its estimated standard error, and the associated confidence interval or significance test.

The estimated variance of the combined point estimate is computed by adding two components. The first component is the average of the estimated variances obtained from the *m* completed datasets and can be thought of as the within-imputation variance. If *m* were infinite, the second component would be the variation among the point estimates obtained from the *m* completed datasets. The latter component represents the uncertainty due to imputing the missing values and can be thought of as the between-imputation variance. Technical details on how to analyze multiply imputed data are described in Rubin (1987).

The SUDAAN DESCRIPT procedure (RTI International, 2008) produces descriptive statistics for continuous and categorical variables and takes into consideration the sample design and weighting effects when producing estimates. Multiply imputed datasets are handled in the following manner. Suppose that there are *m* ($m \geq 2$) distinct imputed datasets and a parameter θ is being estimated. The estimate of this parameter from the i^{th} imputed dataset is $\hat{\theta}_i$ and the variance is $\hat{V}(\hat{\theta}_i)$. The MI estimator is then

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i ,$$

and the variance of the MI estimator $\hat{\theta}$ is

$$V(\hat{\theta}) = \frac{1}{m}\sum_{i=1}^{m}V(\hat{\theta}_i) + \frac{m+1}{m}\left\{\frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\theta}_i - \hat{\theta}\right)^2\right\}$$

This also may be written as

$$V(\hat{\theta}) = W_m + (1+\frac{1}{m})B_m ,$$

where W_m is the (estimated) within-imputation variance, which is the average of variances of the *m* completed estimates, and B_m is the (estimated) between-imputation variance of the *m* imputed estimates.

The extra $\frac{1}{m} B_m$ term in $V(\hat{\theta})$ results from *m* being finite. If only a single set of imputations were used operationally, but five sets were computed for variance-estimation purposes, $(1+\frac{1}{m})$ in $V(\hat{\theta})$ would be replaced by 2.

For this evaluation, the effects of the five multiple imputations from the IVEware and modified predictive mean neighborhood multiple imputation (modPMN-MI) methods were analyzed using these procedures. For each of these two methods, the total variance $V(\hat{\theta})$, within-imputation variance (W_m), and between-imputation variance (B_m) components were calculated. To achieve a more comprehensive understanding of the components of variance inflation, the estimated relative increase in variance due to imputation given by the following was examined:

$$R = \frac{V(\hat{\theta})}{W_m} - 1 .$$

Table B.1 Estimates, Variances, and 95 Percent Confidence Intervals for IVEware

Variable	Percentage Imputed[1]	Imputed Estimate[1]	Variance Components			Relative Increase in Variance Due to Imputation (%)	95 Percent Confidence Interval	
			Between Variance	Within Variance	Total Variance		Lower Level[1]	Upper Level[1]
Race								
American Indian/Alaska Native	24.6	2.3	0.0067	0.0127	0.0208	63.52	2.0	2.6
Asian/Other Pacific Islander	3.3	4.9	0.0001	0.0421	0.0422	0.20	4.5	5.3
Black/African American	0.4	12.3	0.0002	0.1429	0.1432	0.19	11.6	13.1
White	1.6	80.5	0.0038	0.1851	0.1897	2.48	79.7	81.4
Hispanic/Latino Origin								
Hispanic/Latino	0.1	13.8	0.0000	0.1168	0.1168	0.00	13.2	14.5
Non-Hispanic/Latino	0.2	86.2	0.0000	0.1168	0.1168	0.00	85.5	86.9
Marital Status								
Married	0.1	52.4	0.0000	0.2191	0.2191	0.00	51.4	53.3
Widowed	0.0	5.7	0.0000	0.0526	0.0526	0.00	5.2	6.1
Divorced/Separated	0.1	12.6	0.0000	0.0804	0.0805	0.03	12.1	13.2
Never Been Married	0.0	29.4	0.0000	0.1202	0.1202	0.01	28.7	30.1
Education Level								
Less than High School	0.0	16.3	0.0000	0.1317	0.1317	0.01	15.6	17.0
High School Graduate	0.0	30.6	0.0001	0.1616	0.1617	0.06	29.8	31.4
Some College	0.0	25.8	0.0000	0.1269	0.1270	0.03	25.1	26.5
College Graduate	0.0	27.3	0.0001	0.2005	0.2006	0.09	26.4	28.1
Cigarettes Recency								
Within Past 30 Days	1.1	24.3	0.0001	0.1063	0.1064	0.10	23.7	25.0
More than 30 Days Ago but within Past 12 Months	2.7	4.2	0.0000	0.0160	0.0160	0.14	4.0	4.5
More than 12 Months Ago but within Past 3 Years	2.9	3.9	0.0000	0.0146	0.0147	0.24	3.7	4.1
More than 3 Years Ago	0.6	32.8	0.0001	0.1679	0.1680	0.06	32.0	33.6
Never Smoked Cigarettes	0.0	34.7	0.0000	0.1257	0.1257	0.00	34.0	35.4
Cigarettes								
30-Day Frequency	2.2	22.6 days	0.0000	0.0225	0.0225	0.05	22.3	22.9
Age at First Use	1.4	15.7 years	0.0000	0.0019	0.0020	2.09	15.6	15.8

Table B.1 Estimates, Variances, and 95 Percent Confidence Intervals for IVEware (continued)

Variable	Percentage Imputed[1]	Imputed Estimate[1]	Variance Components			Relative Increase in Variance Due to Imputation (%)	95 Percent Confidence Interval	
			Between Variance	Within Variance	Total Variance		Lower Level[1]	Upper Level[1]
Alcohol Recency								
Within Past 30 Days	1.3	51.1	0.0017	0.1686	0.1706	1.20	50.3	52.0
More than 30 Days Ago but within Past 12 Months	3.0	14.5	0.0002	0.0627	0.0630	0.44	14.0	15.0
More than 12 Months Ago	1.9	16.6	0.0009	0.0999	0.1009	1.07	16.0	17.3
Never Used Alcohol	0.1	17.7	0.0000	0.0841	0.0841	0.01	17.1	18.3
Alcohol								
12-Month Frequency	5.5	86.8 days	0.0009	0.9163	0.9173	0.12	84.9	88.7
30-Day Frequency	2.8	8.4 days	0.0000	0.0088	0.0088	0.10	8.2	8.6
Age at First Use	1.6	17.0 years	0.0000	0.0023	0.0023	0.58	17.0	17.1
Marijuana Recency								
Within Past 30 Days	2.0	5.8	0.0003	0.0207	0.0210	1.80	5.5	6.1
More than 30 Days Ago but within Past 12 Months	1.9	4.2	0.0002	0.0148	0.0150	1.72	4.0	4.5
More than 12 Months Ago	1.2	30.5	0.0004	0.1307	0.1312	0.36	29.8	31.3
Never Used Marijuana	0.1	59.5	0.0000	0.1508	0.1509	0.02	58.7	60.2
Marijuana								
12-Month Frequency	9.4	102.7 days	0.0240	4.9416	4.9704	0.58	98.3	107.0
30-Day Frequency	3.0	12.9 days	0.0009	0.0704	0.0715	1.47	12.3	13.4
Age at First Use	1.0	18.0 years	0.0000	0.0057	0.0057	0.08	17.9	18.2

[1] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table B.2 Estimates, Variances, and 95 Percent Confidence Intervals for modPMN-MI

Variable	Percentage Imputed[1]	Imputed Estimate[1]	Variance Components			Relative Increase in Variance Due to Imputation (%)	95 Percent Confidence Interval	
			Between Variance	Within Variance	Total Variance		Lower Level[1]	Upper Level[1]
Race								
American Indian/Alaska Native	10.7	1.5	0.0018	0.0069	0.0091	31.90	1.3	1.7
Asian/Other Pacific Islander	2.0	4.9	0.0001	0.0423	0.0423	0.15	4.5	5.3
Black/African American	1.1	12.4	0.0006	0.1434	0.1441	0.46	11.6	13.1
White	2.6	81.2	0.0033	0.1794	0.1834	2.23	80.4	82.1
Hispanic/Latino Origin								
Hispanic/Latino	0.1	13.8	0.0000	0.1168	0.1168	0.00	13.2	14.5
Non-Hispanic/Latino	0.2	86.2	0.0000	0.1168	0.1168	0.00	85.5	86.9
Marital Status								
Married	0.0	52.4	0.0000	0.2191	0.2191	0.01	51.4	53.3
Widowed	0.0	5.7	0.0000	0.0526	0.0526	0.00	5.2	6.1
Divorced/Separated	0.1	12.60	0.0000	0.0804	0.0804	0.00	12.1	13.2
Never Been Married	0.0	29.38	0.0000	0.1202	0.1202	0.01	28.7	30.1
Education Level								
Less than High School	0.0	16.3	0.0000	0.1317	0.1317	0.01	15.6	17.0
High School Graduate	0.0	30.6	0.0001	0.1620	0.1621	0.06	29.8	31.4
Some College	0.0	25.8	0.0001	0.1270	0.1271	0.12	25.1	26.5
College Graduate	0.1	27.2	0.0002	0.2004	0.2006	0.12	26.4	28.1
Cigarettes Recency								
Within Past 30 Days	0.1	24.2	0.0000	0.1064	0.1065	0.04	23.6	24.9
More than 30 Days Ago but within Past 12 Months	2.8	4.2	0.0000	0.0159	0.0160	0.16	4.0	4.5
More than 12 Months Ago but within Past 3 Years	5.6	4.0	0.0000	0.0146	0.0147	0.13	3.7	4.2
More than 3 Years Ago	1.1	32.9	0.0000	0.1682	0.1683	0.03	32.1	33.7
Never Smoked Cigarettes	0.0	34.7	0.0000	0.1257	0.1257	0.00	34.0	35.4
Cigarettes								
30-Day Frequency	1.3	22.6 days	0.0001	0.0227	0.0228	0.35	22.3	22.9
Age at First Use	1.4	15.7 years	0.0000	0.0019	0.0020	1.14	15.6	15.8

Table B.2 Estimates, Variances, and 95 Percent Confidence Intervals for modPMN-MI (continued)

Variable	Percentage Imputed[1]	Imputed Estimate[1]	Variance Components			Relative Increase in Variance Due to Imputation (%)	95 Percent Confidence Interval	
			Between Variance	Within Variance	Total Variance		Lower Level[1]	Upper Level[1]
Alcohol Recency								
Within Past 30 Days	1.4	51.2	0.0005	0.1685	0.1692	0.36	50.4	52.0
More than 30 Days Ago but within Past 12 Months	3.0	14.5	0.0007	0.0625	0.0633	1.30	14.0	15.0
More than 12 Months Ago	1.3	16.6	0.0001	0.0997	0.0998	0.09	16.0	17.2
Never Used Alcohol	0.1	17.70	0.0000	0.0840	0.0841	0.04	17.1	18.3
Alcohol								
12-Month Frequency	5.5	86.7 days	0.0013	0.9163	0.9178	0.17	84.8	88.6
30-Day Frequency	3.0	8.4 days	0.0001	0.0088	0.0090	1.71	8.2	8.6
Age at First Use	1.6	17.0 years	0.0000	0.0023	0.0023	0.84	16.9	17.1
Inhalants Recency								
Within Past 30 Days	10.4	0.3	0.0002	0.0008	0.0011	37.19	0.2	0.3
More than 30 Days Ago but within Past 12 Months	8.8	0.6	0.0001	0.0012	0.0014	11.01	0.5	0.7
More than 12 Months Ago	3.3	8.2	0.0002	0.0374	0.0376	0.55	7.8	8.6
Never Used Inhalants	0.2	90.9	0.0001	0.0411	0.0412	0.27	90.5	91.3
Inhalants								
12-Month Frequency	23.8	29.4 days	0.5064	17.0960	17.7037	3.55	21.1	37.7
30-Day Frequency	16.7	4.1 days	0.0076	0.1493	0.1585	6.14	3.3	4.8
Age at First Use	7.9	17.3 years	0.0003	0.0186	0.0189	2.04	17.0	17.6
Marijuana Recency								
Within Past 30 Days	2.6	5.8	0.0001	0.0209	0.0210	0.42	5.5	6.1
More than 30 Days Ago but within Past 12 Months	3.3	4.3	0.0000	0.0154	0.0155	0.36	4.1	4.5
More than 12 Months Ago	0.5	30.4	0.0000	0.1317	0.1318	0.04	29.7	31.2
Never Used Marijuana	0.01	59.5	0.0000	0.1509	0.1509	0.02	58.7	60.2
Marijuana								
12-Month Frequency	9.9	102.6 days	0.0292	4.9202	4.9553	0.71	97.9	106.6
30-Day Frequency	3.6	12.9 days	0.0001	0.0699	0.0700	0.09	12.4	13.4
Age at First Use	1.0	18.0 years	0.0000	0.0057	0.0057	0.11	17.9	18.2

Table B.2 Estimates, Variances, and 95 Percent Confidence Intervals for modPMN-MI (continued)

Variable	Percentage Imputed[1]	Imputed Estimate[1]	Variance Components			Relative Increase in Variance Due to Imputation (%)	95 Percent Confidence Interval	
			Between Variance	Within Variance	Total Variance		Lower Level[1]	Upper Level[1]
Pain Relievers Recency								
Within Past 30 Days	5.2	2.1	0.0001	0.0075	0.0075	0.94	1.9	2.3
More than 30 Days Ago but within Past 12 Months	5.0	3.0	0.0002	0.0089	0.0091	2.58	2.8	3.1
More than 12 Months Ago	3.1	8.3	0.0003	0.0358	0.0361	0.98	7.9	8.7
Never Used Analgesics	0.60	86.7	0.0002	0.0516	0.0519	0.45	86.2	87.1
Pain Relievers								
12-Month Frequency	13.9	44.7 days	0.3911	3.6785	4.1478	12.76	40.7	48.7
Age at First Use	8.00	22.1 years	0.0009	0.0302	0.0312	3.45	21.7	22.4
Cocaine Recency								
Within Past 30 Days	9.1	0.8	0.0001	0.0037	0.0038	3.11	0.7	1.0
More than 30 Days Ago but within Past 12 Months	5.9	1.5	0.0002	0.0052	0.0055	5.65	1.3	1.6
More than 12 Months Ago	1.0	12.2	0.0001	0.0631	0.0633	0.22	11.7	12.6
Never Used Cocaine	0.1	85.5	0.0000	0.0707	0.0708	0.06	85.0	86.1
Cocaine								
12-Month Frequency	15.2	43.3 days	0.4565	9.1908	9.7386	5.96	37.2	49.4
30-Day Frequency	11.8	6.1 days	0.0375	0.2570	0.3019	17.50	5.1	7.2
Age at First Use	2.7	21.9 years	0.0000	0.0213	0.0213	0.21	21.6	22.2
Heroin Recency								
Within Past 30 Days	5.6	0.1	0.0000	0.0003	0.0003	11.81	0.0	0.1
More than 30 Days Ago but within Past 12 Months	6.5	0.1	0.0000	0.0002	0.0002	21.41	0.1	0.1
More than 12 Months Ago	1.2	1.4	0.0000	0.0083	0.0083	0.05	1.2	1.6
Never Used Heroin	0.1	98.5	0.0000	0.0093	0.0093	0.00	98.3	98.7
Heroin								
12-Month Frequency	24.0	95.6 days	39.8903	284.0855	331.9538	16.85	59.6	131.6
30-Day Frequency	5.6	15.0 days	1.1082	6.0828	7.4126	21.86	9.6	20.4
Age at First Use	5.3	22.9 years	0.0000	0.2466	0.2467	0.01	21.9	23.8

modPMN-MI = modified predictive mean neighborhood multiple imputation.

[1] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

This page intentionally left blank

Appendix C: Methodology for Weighting and Data Augmentation for the Modified Predictive Mean Neighborhood Multiple Imputation Method

This page intentionally left blank

C.1 Alternative Weighting Procedures to Compensate for Item Nonresponse in Imputation

Survey weights are traditionally used to compensate both for the population units not selected into the sample and for sampled units that fail to respond at all to the survey. In principle, the weight returned by the WTADJUST procedure in SUDAAN® (RTI International, 2013) for each respondent is an estimate of the inverse of the unit's joint probability of sample selection and unit response. Sampled units that respond to at least a minimum number of survey items are deemed "unit respondents."

C.1.1 Response Modeling

This same procedure can be used to estimate the probability that a sample unit responds to a particular item on the survey. To do this, one needs to make the following assumptions:

- The probability of item response for a unit is a function of the unit's characteristics and is independent of everything else (like the unit's probability of responding at all to the survey or whether other units respond to the item in question).
- These characteristics and the mathematical form of the response function are known.

Although the response function must have a known form, the parameters of that model are estimated. This is what SUDAAN's WTADJUST procedure does. Starting with a weight, d_k, for unit k equal to the inverse of the unit's estimated joint probability of sample selection and unit response, WTADJUST (implicitly) estimates a probability of item response, p_k, by calculating a new weight, $a_k = d_k / p_k$.

These item-specific weights are not typically created or used in analyses. This new weight is not generally used directly because a unit would potentially have a different weight for every item with item nonresponse. Moreover, the subsample of item respondents usually varies from item to item. Consequently, instead of reweighting the subsample of item respondents for each item with missing values, imputation is performed for the missing item values of unit respondents.

C.1.2 Prediction Modeling

The main tool of item imputation is prediction modeling (also called imputation modeling), that is, predicting what a missing item value is. The assumed model used in this prediction is different from the response model discussed above. Nevertheless, incorporating the response model into the imputation process can provide some protection against bias when the assumed prediction model is wrong.

On this note, consider a particular item that some unit respondents provide responses to and other unit respondents do not. The simplest prediction model assumes that all unit respondents can be separated into mutually exclusive groups based on their known characteristics, that is, characteristics known for all the unit respondents. Although these characteristics may be the same as the characteristics in the response model described previously,

they do not have to be the same. Each group contains both item respondents and item nonrespondents. What the members of a group have in common is the expected value of the item of interest under the prediction model. When the item of interest is a 0/1 variable, its expected value is the probability of it being 1.

If only the prediction model was relied upon, then the item value for any item respondent in Group *g*, for example, could be used to impute for the missing value of a fellow member of the group without causing a bias. Alternatively, the average item value of all item respondents in the group could be used as the imputed values for all item nonrespondent group members.

Suppose the prediction model does not hold, but the response model does hold. Recall that in the response model every unit in the respondent sample has an estimable probability of item response using WTADJUST. In principle, every unit in the entire population also has an estimable probability of item response. Moreover, each population unit can, in principle, be put into a prediction-model group based on its characteristics, even though the prediction model may not hold.

Let *y* denote the item of interest. The expected value under the response model of the population mean value among item nonrespondents in Group *g* (i.e., population units in the group that would have been item nonrespondents had they been sampled and responded at all to the survey) is approximately

$$y_{inrg} = \frac{\sum_{U_g} y_k(1 - p_k)}{\sum_{U_g} (1 - p_k)}.$$

The summations are over all population units in the group, and $1 - p_k$ is the estimated probability that unit *k* is an item nonrespondent. The right-hand side of the equation is only approximately unbiased because, among other things, the p_k probabilities are estimates.

If all the missing values in Group *g* with y_{inrg} were imputed, then it would not be necessary to assume every unit had a common mean, because this is (approximately) the average value of all item nonrespondents meeting the group definition. Under the response model, little or no bias results from estimating the overall population *y*-mean if every item nonrespondent in the population had its *y*-value replaced by such a group-specific construct.

If desired, y_{inrg} could be estimated using only item respondents in Group *g* with

$$y_{inrg} = \frac{\sum_{U_g} a_k y_k(1 - p_k)}{\sum_{U_g} a_k(1 - p_k)},$$

where $a_k = d_k / p_k$ for item respondents in the group and $a_k = 0$ for all other units. It is not hard to show that using y_{inrg} as the imputed value for item nonrespondents in Group *g* returns an estimator for the overall population *y*-mean that is nearly unbiased under the response model and the probability mechanisms generating the unit respondent sample.

Although the above development avoided assuming a prediction model, the resulting estimator is also nearly unbiased under the prediction model because the missing values in Group *g* were imputed with

$$y_{inrg} = \frac{\sum_{R_g} w_k y_k}{\sum_{R_g} w_k},$$

where

$$w_k = d_k \frac{1 - p_k}{p_k},$$

and the summations are over the item respondents in the group.

Observe that in this last formulation, y_{inrg} is simply a weighted average of the item-respondent *y*-values in the group, that is, a prediction-model unbiased estimator of the predicted mean in the group. Nevertheless, the choice of weights, w_k, is a bit unusual because it is dictated from the response model.

Rather than using the predicted mean of a unit with a missing item value, predictive mean neighborhood (PMN) imputation chooses a single donor from among item respondents with predicted means "in the neighborhood" of the predicted mean. With the simple prediction model described above, the group defines a neighborhood containing item respondents and nonrespondents with the exact same predicted mean. Expressed this way, there is little difference between PMN and weighted sequential hot deck (WSHD). Moreover, by giving each item respondent *k* in a group a probability of being a donor proportional to w_k, the resulting estimator would remain nearly unbiased under the response model since the donor is expected to have a *y*-value of y_{inrg}. Here, expectation is defined with respect to the donor-selection mechanism.[1]

When employing a more complicated prediction model using linear or logistic regression (as was done with modified predictive mean neighborhood multiple imputation, which is described in Chapter 5), the estimated predicted mean for a unit *k* has the form $f(\mathbf{x}_k'\boldsymbol{\beta})$, where $\mathbf{x}_k$ is a vector of characteristics and $\boldsymbol{\beta}$ is a vector of unknown regression coefficients. Using reasoning similar to that described earlier, one can estimate $\boldsymbol{\beta}$ by solving $\sum_R w_k \left(y_k - f(\mathbf{x}_k'\boldsymbol{\beta})\right)\mathbf{x}_k = 0$, where the summation is over the item respondents. This will return estimated predicted means that, when used as imputed values for missing *y*, will produce an estimator for the overall population *y*-mean that is nearly unbiased under the response model even when the prediction model $E(y_k) = f(\mathbf{x}_k'\boldsymbol{\beta})$ fails. See Kott and Folsom (2010) for a more rigorous treatment.

[1] In the two versions of WSHD discussed in Chapter 3 (and generally in practice), every unit respondent in an imputation group was implicitly assumed equally likely to be an item respondent (i.e., p_k was assumed constant within each group). This was reflected by d_k in a group being treated like w_k.

C.2 Steps for Implementing Data Augmentation for the Modified Predictive Mean Neighborhood Multiple Imputation Method

This appendix describes the steps that were used to correct the variance-covariance matrix in SAS when the non-positive definite matrix was generated. The practice of drawing a random vector from a multivariate normal distribution is easily done using the VNORMAL routine in PROC IML in SAS. If the regression model was over fitted, the VNORMAL routine would not execute. Instead, PROC IML would report that the variance-covariance matrix is not positive definite. However, the truth is that the variance-covariance matrix is positive definite, but barely so, since some of its eigenvalues are positive but very small (Strang, 1988, p. 331).[2] There are two reasonable solutions to this problem.

The solution used for this problem involved factoring the variance-covariance matrix into its eigenvector and eigenvalues matrices (Strang, 1988, p. 254) and to reassemble the variance-covariance matrix by "correcting" the small positive eigenvalues until the VNORMAL routine performed correctly. The following steps were performed to ensure the VNORMAL routine performed correctly:

- Step 1: Use the EIGEN call routine in PROC IML in SAS to create a column vector of eigenvalues M and a matrix of eigenvectors E that can be used to factor the variance-covariance matrix Σ, as shown in Strang (1988).
- Step 2: Choose a small positive number ε that is larger than the smallest element of M. Create the vector N by replacing all elements in M that are less than ε with ε.
- Step 3: Create a diagonal matrix D from the column vector N using the DIAG function in PROC IML in SAS.
- Step 4: Create a matrix similar to Σ, called Σ^*, using matrix multiplication: $\Sigma^* = EDE^T$.
- Step 5: Verify that the VNORMAL routine performs correctly using Σ^* instead of Σ. If this does not work, try Steps 2 through 5 again using a larger value for ε.

Out of the more than 1,600 regression models that were fit, the correction was implemented about 300 times. These corrections are not expected to have a significant impact on the results of the comparison of methods.

[2] Theorem 6B, Condition II: A matrix is positive definite if, and only if, its eigenvalues are positive.

Appendix D: Methodology for Evaluating the Different Imputation Methods

This page intentionally left blank

This appendix details the methodology used to evaluate four alternative imputation methods as compared with the predictive mean neighborhood (PMN) method that is used operationally in the National Survey on Drug Use and Health (NSDUH). The four alternative methods that were evaluated are (1) simple weighted sequential hot deck (WSHD), (2) complex WSHD, (3) IVEware, and (4) modified PMN multiple imputation (modPMN-MI). The following analyses were conducted for all ages (12 or older) and split by age group (12 to 17, 18 to 25, and 26 or older):

- an examination of weighted and unweighted means before and after imputation for each of the five methods,
- a repeated measures analysis to test for statistical differences among the methods,
- an analysis of bias ratios and associated confidence interval coverage probabilities assuming PMN was an unbiased method, and
- an examination of two-way drug relationships for a subset of drug measures.

D.1 Estimates by Imputation Method

The first comparisons that were examined were the changes in the variable distributions across methods. The unweighted frequency counts and weighted percentages for each of the imputation methods were calculated before and after imputation and only using the imputed or logically assigned records. The counts and percentages were calculated using the SUDAAN® DESCRIPT procedure (RTI International, 2013). For the weighted percentages, the final person-level weight (ANALWT) from the 2007 NSDUH was used. The unweighted frequency counts were not reported by age group due to small sample sizes and the need to apply suppression rules.

As have been used in other NSDUH reports, the suppression rules for unreliable estimates (Aldworth et al., 2009) were applied in this report to caution the reader. If one of the suppression criteria was met, then an asterisk was used to denote low precision for that estimate.

D.2 Checking for Statistical Differences across Imputation Methods: Repeated Measures Methodology

Of particular interest to this evaluation was whether or not the five imputation methods differed significantly from each other. In this scenario, each imputation was a treatment applied to a set of missing data (i.e., a set of experimental units). Because the set of missing data was the same for each imputation method, the experimental units were correlated such that the assumption of independent outcome measurements, required for a regular regression model, was violated. To account for this correlation, a repeated measures model was used. The data used for this analysis included all of the survey data—the imputed, logically assigned, and nonimputed data—in addition to the analysis weights.

Typically, in a repeated measures analysis, one challenge for the researcher is describing the correlation matrix. In most analyses, this correlation matrix is a nuisance parameter; that is, it may be needed to obtain meaningful inference from the analysis, but the actual values of the matrix are not important to the researcher. The use of generalized estimating equations (GEE) (Binder, 1983; Zeger & Liang, 1986) alleviates this problem. Estimation of the exact correlation

structure is unnecessary when using GEE to determine variance estimates from a complex sample. Each cluster is allowed to have a unique correlation structure that does not need to be estimated in order to estimate the variance of a parameter of interest.

The analysis can be described with the following working model:

$$\mathbf{y}_{hijk} = g(\alpha + \delta_k) + \varepsilon_{hijk},$$
$$Var(\boldsymbol{\varepsilon}_{hijk}) = \sigma^2_{hijk},$$

where *h* is an index for the stratum variable, *i* is an index for the primary sampling unit (PSU) variable, *j* is an index for each person (imputed data, if missing, or survey response, if not missing), and *k* is an index for the imputation method (i.e., the repeated measure) within a PSU.

The y_{hijk} value is repeated in cases where the response is imputed, and it is duplicated in cases where the response is the respondent's actual answer to the survey question. There is a common intercept α and a parameter δ_k that is set equal to zero for one of the methods. The error term, ε_{hijk}, has a mean of zero and is allowed to be correlated across respondents within primary sampling units and across the imputation methods within respondents. The functional form of *g*(*) depends on the type of regression. For linear regression, *g*(*) is the identity link function; for logistic regression, it is the logit link function; and for multinomial models, it is the generalized or cumulative logit link function (Agresti, 1990).

The repeated measures part of the statistical analysis was conducted with two different approaches. Since the main interest of the evaluation was how PMN compared with the other imputation methods, an initial analysis of pairwise comparisons for PMN versus each imputation method was conducted. This analysis included four comparisons per variable (PMN vs. simple WSHD, PMN vs. complex WSHD, PMN vs. IVEware, and PMN vs. modPMN-MI). The results for this first analysis are found in each chapter explaining the respective imputation methods. Next, an exploratory analysis looking at all possible pairwise comparisons was conducted. To control the error rates, each variable was first tested for global differences among the sample-weighted means or proportions (depending on the variable being examined). If differences were found at the global level (i.e., a test that all $\delta_k = 0$ failed), then pairwise comparisons between imputation methods were performed to determine which imputation methods differed. To further understand where the statistical differences were for the categorical variables (demographics and drug recency), the exploratory analysis was conducted within each level (i.e., white, black/African American, Asian/Other Pacific Islander, etc.) of the variables. The exploratory analysis was also conducted by age group. These results can be found in Chapter 6.

D.3 Bias Ratios and Confidence Intervals for Coverage Probabilities for PMN Imputation Methods versus Other Imputation Methods

Due to the large sample size and high correlation within the repeated measures, there were many significant differences among the imputation methods. Some of these differences were not meaningful while others were very meaningful. As a method of sorting through the significant differences of PMN versus the other four imputation methods, an ad hoc assessment

calculating the bias ratio and confidence intervals for coverage probabilities (Cochran, 1977, pp. 12-15) was conducted. Among comparisons with significant differences, the following was performed:

1. Assumed the PMN-based estimates were unbiased.[1]
2. Calculated the ratio of the estimated bias:
   Bias Ratio = (alternative imputation method estimate – PMN imputation estimate)/standard error (alternative imputation method estimate).
3. Based on the standard normal 95 percent confidence interval, calculated the upper and lower limits using the bias ratio estimator:
   [Bias Ratio – 1.96], [Bias Ratio + 1.96].
4. Calculated the area under the standard (mean 0) normal curve.

As in the repeated analysis, the data used for this analysis included the imputed, logically assigned, and nonimputed data as well as the analysis weights. Note that since the standard errors used to compute the bias ratio did not include the contributions to variance resulting from the imputation process, the associated calculations were too large. Therefore, the quoted coverage probabilities for the alternative methods were somewhat smaller than expected.

[1] Although, the PMN estimates are assumed to be unbiased for this analysis, this is probably not true. The PMN estimates were chosen as the "unbiased estimates" because it is the operational NSDUH imputation method.

This page intentionally left blank

Appendix E: Before and After Imputation Distributions

This page intentionally left blank

Table E.1 Before and After Imputation Distributions, by Imputation Method for Race

Variable	Unweighted Frequency Counts		Weighted Percentages		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
PMN					
American Indian/Alaska Native	2,800	398	1.1	3.4	1.9
Asian/Other Pacific Islander	3,000	50	5.0	1.8	4.9
Black/African American	8,900	103	12.6	1.8	12.3
White	51,300	1,309	81.	93.0	81.6
Total	66,000	1,860	100.0	100.0	100.0
Simple WSHD					
American Indian/Alaska Native	2,800	84	1.1	4.6	1.2
Asian/Other Pacific Islander	3,000	83	5.0	3.9	4.9
Black/African American	8,900	249	12.6	9.8	12.5
White	51,300	1,444	81.3	81.7	81.4
Total	66,000	1,860	100.0	100.0	100.0
IVEware					
American Indian/Alaska Native	2,800	900	1.1	48.5	2.3
Asian/Other Pacific Islander	3,000	103	5.0	4.0	4.9
Black/African American	8,900	33	12.6	1.0	12.3
White	51,300	824	81.3	46.5	80.5
Total	66,000	1,860	100.0	100.0	100.0
modPMN-MI					
American Indian/Alaska Native	2,800	330	1.1	15.6	1.5
Asian/Other Pacific Islander	3,000	61	5.0	2.7	4.9
Black/African American	8,900	98	12.6	3.2	12.4
White	51,300	1,371	81.3	78.5	81.3
Total	66,000	1,860	100.0	100.0	100.0

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table E.2 Before and After Imputation Distributions, by Imputation Method for Hispanic/Latino Origin

Variable	Unweighted Frequency Counts		Weighted Percentages		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
PMN					
Hispanic/Latino	10,300	9	13.8	11.1*	13.8
Non-Hispanic/Latino	57,500	100	86.2	88.9*	86.2
Total	67,800	109	100.0	100.0	100.0
Simple WSHD					
Hispanic/Latino	10,300	9	13.8	6.9*	13.8
Non-Hispanic/Latino	57,500	100	86.2	93.1*	86.2
Total	67,800	109	100.0	100.0	100.0
IVEware					
Hispanic/Latino	10,300	7	13.8	6.1*	13.8
Non-Hispanic/Latino	57,500	102	86.2	93.9*	86.2
Total	67,800	109	100.0	100.0	100.0
modPMN-MI					
Hispanic/Latino	10,300	10	13.8	7.2*	13.8
Non-Hispanic/Latino	57,500	99	86.2	92.8*	86.2
Total	67,800	109	100.0	100.0	100.0

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table E.3 Before and After Imputation Distributions, by Imputation Method for Marital Status

Variable	Unweighted Frequency Counts		Weighted Percentages		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
PMN					
Married	17,300	5	52.4	35.3*	52.4
Widowed	1,000	0	5.7	0.0*	5.7
Divorced/Separated	4,000	3	12.6	28.8*	12.6
Never Been Married	34,400	10	29.4	35.9*	29.4
Total	56,800	18	100.0	100.0	100.0
Simple WSHD					
Married	17,300	8	52.4	67.6*	52.4
Widowed	1,000	0	5.7	0.0*	5.7
Divorced/Separated	4,000	2	12.6	15.1*	12.6
Never Been Married	34,400	8	29.4	17.3*	29.4
Total	56,800	18	100.0	100.0	100.0
IVEware					
Married	17,300	8	52.4	50.3*	52.4
Widowed	1,000	0	5.7	0.0*	5.7
Divorced/Separated	4,000	3	12.6	34.9*	12.6
Never Been Married	34,400	7	29.4	14.8*	29.4
Total	56,800	18	100.0	100.0	100.0
modPMN-MI					
Married	17,300	5	52.4	50.1*	52.4
Widowed	1,000	0	5.7	0.0*	5.7
Divorced/Separated	4,000	2	12.6	13.2*	12.6
Never Been Married	34,400	11	29.4	36.7*	29.4
Total	56,800	18	100.0	100.0	100.0

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table E.4 Before and After Imputation Distributions, by Imputation Method for Education Level

Variable	Unweighted Frequency Counts		Weighted Percentages		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
PMN					
Less than High School	7,700	0	16.3	0.0*	16.3
High School Graduate	14,800	2	30.6	33.1*	30.6
Some College	13,100	4	25.8	23.0*	25.8
College Graduate	9,800	4	27.2	44.0*	27.3
Total	45,400	10	100.0	100.0	100.0
Simple WSHD					
Less than High School	7,700	2	16.3	25.6*	16.3
High School Graduate	14,800	3	30.6	28.1*	30.6
Some College	13,100	4	25.8	15.8*	25.8
College Graduate	9,800	1	27.2	30.6*	27.2
Total	45,400	10	100.0	100.0	100.0
IVEware					
Less than High School	7,700	2	16.3	11.3*	16.3
High School Graduate	14,800	2	30.6	54.1*	30.6
Some College	13,100	3	25.8	8.0*	25.8
College Graduate	9,800	3	27.2	26.5*	27.2
Total	45,400	10	100.0	100.0	100.0
modPMN-MI					
Less than High School	7,700	2	16.3	14.7*	16.3
High School Graduate	14,800	2	30.6	32.6*	30.6
Some College	13,100	1	25.8	1.7*	25.8
College Graduate	9,800	5	27.2	51.0*	27.3
Total	45,400	10	100.0	100.0	100.0

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table E.5 Before and After Imputation Distributions, by Imputation Method for Cigarettes Recency

Variable	Unweighted Frequency Counts		Weighted Percentages		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
PMN					
Within Past 30 Days	17,000	15	24.3	2.7	24.2
More than 30 Days Ago but within Past 12 Months	4,000	122	4.2	20.3	4.2
More than 12 Months Ago but within Past 3 Years	3,300	191	3.9	37.3	4.0
More than 3 Years Ago	12,100	141	32.8	39.7	32.9
Never Smoked Cigarettes	31,000	0	34.8	0.0*	34.7
Total	67,400	469	100.0	100.0	100.0
Simple WSHD					
Within Past 30 Days	17,000	24	24.3	10.4*	24.3
More than 30 Days Ago but within Past 12 Months	4,000	135	4.2	23.5	4.3
More than 12 Months Ago but within Past 3 Years	3,300	173	3.9	31.4	3.9
More than 3 Years Ago	12,100	137	32.8	34.7	32.8
Never Smoked Cigarettes	31,000	0	34.8	0.0*	34.7
Total	67,400	469	100.0	100.0	100.0
Complex WSHD					
Within Past 30 Days	17,000	23	24.3	4.3	24.2
More than 30 Days Ago but within Past 12 Months	4,000	125	4.2	23.5	4.3
More than 12 Months Ago but within Past 3 Years	3,300	182	3.9	33.6	3.9
More than 3 Years Ago	12,100	139	32.8	38.6	32.9
Never Smoked Cigarettes	31,000	0	34.8	0.0*	34.7
Total	67,400	469	100.0	100.0	100.0
IVEware					
Within Past 30 Days	17,000	187	24.3	32.4	24.3
More than 30 Days Ago but within Past 12 Months	4,000	110	4.2	21.7	4.2
More than 12 Months Ago but within Past 3 Years	3,300	98	3.9	19.6	3.9
More than 3 Years Ago	12,100	74	32.8	26.4*	32.8
Never Smoked Cigarettes	31,000	0	34.8	0.0*	34.7
Total	67,400	469	100.0	100.0	100.0

Table E.5 Before and After Imputation Distributions, by Imputation Method for Cigarettes Recency (continued)

Variable	Unweighted Frequency Counts		Weighted Percentages		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
modPMN-MI					
Within Past 30 Days	17,000	19	24.3	3.2	24.2
More than 30 Days Ago but within Past 12 Months	4,000	116	4.2	22.4	4.2
More than 12 Months Ago but within Past 3 Years	3,300	197	3.9	34.0	4.0
More than 3 Years Ago	12,100	137	32.8	40.3	32.9
Never Smoked Cigarettes	31,000	0	34.8	0.0*	34.7
Total	67,400	469	100.0	100.0	100.0

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table E.6 Before and After Imputation Distributions, by Imputation Method for Alcohol Recency

Variable	Unweighted Frequency Counts		Weighted Percentages		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
PMN					
Within Past 30 Days	30,100	405	51.1	54.9	51.1
More than 30 Days Ago but within Past 12 Months	10,600	337	14.4	34.8	14.6
More than 12 Months Ago	7,400	116	16.7	8.7	16.6
Never Used Alcohol	18,900	21	17.9	1.6	17.7
Total	67,000	879	100.0	100.0	100.0
Simple WSHD					
Within Past 30 Days	30,100	464	51.1	64.0	51.2
More than 30 Days Ago but within Past 12 Months	10,600	295	14.4	28.1	14.5
More than 12 Months Ago	7,400	99	16.7	6.4	16.6
Never Used Alcohol	18,900	21	17.9	1.5	17.7
Total	67,000	879	100.0	100.0	100.0
Complex WSHD					
Within Past 30 Days	30,100	488	51.1	67.2	51.3
More than 30 Days Ago but within Past 12 Months	10,600	290	14.4	26.8	14.5
More than 12 Months Ago	7,400	80	16.7	4.5	16.5
Never Used Alcohol	18,900	21	17.9	1.5	17.7
Total	67,000	879	100.0	100.0	100.0
IVEware					
Within Past 30 Days	30,100	392	51.1	47.9	51.1
More than 30 Days Ago but within Past 12 Months	10,600	323	14.4	28.9	14.5
More than 12 Months Ago	7,400	141	16.7	21.2	16.7
Never Used Alcohol	18,900	23	17.9	2.0	17.7
Total	67,000	879	100.0	100.0	100.0

Table E.6 Before and After Imputation Distributions, by Imputation Method for Alcohol Recency (continued)

Variable	Unweighted Frequency Counts		Weighted Percentages		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
modPMN-MI					
Within Past 30 Days	30,100	434	51.1	59.1	51.2
More than 30 Days Ago but within Past 12 Months	10,600	324	14.4	31.8	14.5
More than 12 Months Ago	7,400	98	16.7	7.0	16.6
Never Used Alcohol	18,900	23	17.9	2.1	17.7
Total	67,000	879	100.0	100.0	100.0

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table E.7 Before and After Imputation Distributions, by Imputation Method for Inhalants Recency

Variable	Unweighted Frequency Counts		Weighted Percentages		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
PMN					
Within Past 30 Days	400	32	0.2	7.5	0.3
More than 30 Days Ago but within Past 12 Months	900	81	0.6	14.2	0.6
More than 12 Months Ago	5,700	203	8.1	48.3	8.2
Never Used Inhalants	60,500	123	91.1	30.0	90.9
Total	67,400	439	100.0	100.0	100.0
Simple WSHD					
Within Past 30 Days	400	37	0.2	5.7	0.2
More than 30 Days Ago but within Past 12 Months	900	90	0.6	17.8	0.6
More than 12 Months Ago	5,700	192	8.1	43.1	8.2
Never Used Inhalants	60,500	120	91.1	33.5	90.9
Total	67,400	439	100.0	100.0	100.0
Complex WSHD					
Within Past 30 Days	400	54	0.2	8.4	0.3
More than 30 Days Ago but within Past 12 Months	900	98	0.6	20.6	0.6
More than 12 Months Ago	5,700	168	8.1	39.5*	8.2
Never Used Inhalants	60,500	119	91.1	31.5	90.9
Total	67,400	439	100.0	100.0	100.0
modPMN-MI					
Within Past 30 Days	400	41	0.2	5.6	0.2
More than 30 Days Ago but within Past 12 Months	900	86	0.6	17.4	0.6
More than 12 Months Ago	5,700	194	8.1	47.6	8.2
Never Used Inhalants	60,500	118	91.1	29.4	90.9
Total	67,400	439	100.0	100.0	100.0

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table E.8 Before and After Imputation Distributions, by Imputation Method for Marijuana Recency

Variable	Unweighted Frequency Counts		Weighted Percentages		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
PMN					
Within Past 30 Days	6,200	160	5.7	31.6	5.8
More than 30 Days Ago but within Past 12 Months	4,500	146	4.2	32.8	4.3
More than 12 Months Ago	15,200	96	30.4	28.4	30.4
Never Used Marijuana	41,500	26	59.7	7.3	59.4
Total	67,400	428	100.0	100.0	100.0
Simple WSHD					
Within Past 30 Days	6,200	168	5.7	35.2	5.8
More than 30 Days Ago but within Past 12 Months	4,500	139	4.2	29.2	4.3
More than 12 Months Ago	15,200	91	30.4	25.3	30.4
Never Used Marijuana	41,500	30	59.7	10.2	59.5
Total	67,400	428	100.0	100.0	100.0
Complex WSHD					
Within Past 30 Days	6,200	180	5.7	42.5	5.9
More than 30 Days Ago but within Past 12 Months	4,500	133	4.2	25.2	4.3
More than 12 Months Ago	15,200	87	30.4	21.8	30.4
Never Used Marijuana	41,500	28	59.7	10.5	59.5
Total	67,400	428	100.0	100.0	100.0
IVEware					
Within Past 30 Days	6,200	128	5.7	18.4	5.8
More than 30 Days Ago but within Past 12 Months	4,500	87	4.2	12.0	4.2
More than 12 Months Ago	15,200	185	30.4	61.5	30.6
Never Used Marijuana	41,500	28	59.7	8.1	59.5
Total	67,400	428	100.0	100.0	100.0

Table E.8 Before and After Imputation Distributions, by Imputation Method for Marijuana Recency (continued)

Variable	Unweighted Frequency Counts		Weighted Percentages		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
modPMN-MI					
Within Past 30 Days	6,200	167	5.7	28.7	5.8
More than 30 Days Ago but within Past 12 Months	4,500	154	4.2	34.4	4.3
More than 12 Months Ago	15,200	79	30.4	26.8	30.4
Never Used Marijuana	41,500	28	59.7	10.1	59.5
Total	67,400	428	100.0	100.0	100.0

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table E.9 Before and After Imputation Distributions, by Imputation Method for Pain Relievers Recency

Variable	Unweighted Frequency Counts		Weighted Percentages		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
PMN					
Within Past 30 Days	1,900	108	2.0	13.6	2.1
More than 30 Days Ago but within Past 12 Months	3,000	152	2.9	16.4	2.9
More than 12 Months Ago	5,600	181	8.2	28.1	8.3
Never Used Analgesics	56,600	345	87.0	41.9	86.7
Total	67,100	786	100.0	100.0	100.0
Simple WSHD					
Within Past 30 Days	1,900	111	2.0	12.3	2.1
More than 30 Days Ago but within Past 12 Months	3,000	164	2.9	17.2	3.0
More than 12 Months Ago	5,600	169	8.2	29.7	8.3
Never Used Analgesics	56,600	342	87.0	40.7	86.7
Total	67,100	786	100.0	100.0	100.0
Complex WSHD					
Within Past 30 Days	1,900	119	2.0	15.6	2.1
More than 30 Days Ago but within Past 12 Months	3,000	151	2.9	13.8	2.9
More than 12 Months Ago	5,600	170	8.2	28.2	8.3
Never Used Analgesics	56,600	346	87.0	42.5	86.7
Total	67,100	786	100.0	100.0	100.0
modPMN-MI					
Within Past 30 Days	1,900	106	2.0	10.7	2.1
More than 30 Days Ago but within Past 12 Months	3,000	159	2.9	18.3	3.0
More than 12 Months Ago	5,600	179	8.2	33.2	8.3
Never Used Analgesics	56,600	342	87.0	37.8	86.6
Total	67,100	786	100.0	100.0	100.0

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table E.10 Before and After Imputation Distributions, by Imputation Method for Cocaine Recency

Variable	Unweighted Frequency Counts		Weighted Percentages		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
PMN					
Within Past 30 Days	600	56	0.8	25.1	0.8
More than 30 Days Ago but within Past 12 Months	1,400	91	1.4	38.7	1.5
More than 12 Months Ago	5,700	61	12.1	20.5	12.2
Never Used Cocaine	59,800	29	85.8	15.7*	85.5
Total	67,600	237	100.0	100.0	100.0
Simple WSHD					
Within Past 30 Days	600	60	0.8	31.2*	0.9
More than 30 Days Ago but within Past 12 Months	1,400	93	1.4	34.4	1.5
More than 12 Months Ago	5,700	50	12.1	15.3	12.2
Never Used Cocaine	59,800	34	85.8	19.1*	85.5
Total	67,600	237	100.0	100.0	100.0
Complex WSHD					
Within Past 30 Days	600	64	0.8	30.7*	0.9
More than 30 Days Ago but within Past 12 Months	1,400	97	1.4	36.3*	1.5
More than 12 Months Ago	5,700	46	12.1	18.8*	12.2
Never Used Cocaine	59,800	30	85.8	14.2*	85.5
Total	67,600	237	100.0	100.0	100.0
modPMN-MI					
Within Past 30 Days	600	63	0.8	29.5*	0.9
More than 30 Days Ago but within Past 12 Months	1,400	89	1.4	37.2	1.5
More than 12 Months Ago	5,700	55	12.1	17.7	12.2
Never Used Cocaine	59,800	30	85.8	15.7*	85.5
Total	67,600	237	100.0	100.0	100.0

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table E.11 Before and After Imputation Distributions, by Imputation Method for Heroin Recency

Variable	Unweighted Frequency Counts		Weighted Percentages		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
PMN					
Within Past 30 Days	100	1	0.1	0.3*	0.1
More than 30 Days Ago but within Past 12 Months	100	8	0.1	34.5*	0.1
More than 12 Months Ago	600	8	1.4	18.2*	1.4
Never Used Heroin	67,000	36	98.5	47.0*	98.5
Total	67,800	53	100.0	100.0	100.0
Simple WSHD					
Within Past 30 Days	100	4	0.1	3.1*	0.1
More than 30 Days Ago but within Past 12 Months	100	5	0.1	32.4*	0.1
More than 12 Months Ago	600	8	1.4	17.6*	1.4
Never Used Heroin	67,000	36	98.5	47.0*	98.5
Total	67,800	53	100.0	100.0	100.0
Complex WSHD					
Within Past 30 Days	100	1	0.1	0.3*	0.1
More than 30 Days Ago but within Past 12 Months	100	6	0.1	31.6*	0.1
More than 12 Months Ago	600	10	1.4	21.1*	1.4
Never Used Heroin	67,000	36	98.5	47.0*	98.5
Total	67,800	53	100.0	100.0	100.0
modPMN-MI					
Within Past 30 Days	100	3	0.1	28.0*	0.1
More than 30 Days Ago but within Past 12 Months	100	6	0.1	5.2*	0.1
More than 12 Months Ago	600	8	1.4	19.9*	1.4
Never Used Heroin	67,000	36	98.5	47.0*	98.5
Total	67,800	53	100.0	100.0	100.0

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table E.12 Before and After Imputation Distributions, by Imputation Method for Cigarettes Frequency

Variable	Unweighted Sample Size		Weighted Mean Estimates		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
PMN					
30-Day Frequency	16,800	211	22.7 days	18.7 days	22.6 days
Age at First Use	36,400	532	15.7 years	14.4 years	15.7 years
Simple WSHD					
30-Day Frequency	16,800	220	22.7 days	19.5 days	22.6 days
Age at First Use	36,400	532	15.7 years	15.1 years	15.7 years
Complex WSHD					
30-Day Frequency	16,800	219	22.7 days	19.8 days	22.6 days
Age at First Use	36,400	532	15.7 years	15.0 years	15.7 years
IVEware					
30-Day Frequency	16,800	383	22.7 days	18.6 days	22.6 days
Age at First Use	36,400	532	15.7 years	15.6 years	15.7 years
modPMN-MI					
30-Day Frequency	16,800	215	22.7 days	17.5 days	22.6 days
Age at First Use	36,400	532	15.7 years	14.7 years	15.7 years

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table E.13 Before and After Imputation Distributions, by Imputation Method for Alcohol Frequency

Variable	Unweighted Sample Size		Weighted Mean Estimates		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
PMN					
12-Month Frequency	39,200	2,276	84.5 days	160.2 days	86.9 days
30-Day Frequency	29,700	875	8.4 days	7.9 days	8.4 days
Age at First Use	48,200	802	17.0 years	17.5 years	17.0 years
Simple WSHD					
12-Month Frequency	39,200	2,293	84.5 days	141.4 days	86.3 days
30-Day Frequency	29,700	934	8.4 days	7.0 days	8.4 days
Age at First Use	48,200	802	17.0 years	16.9 years	17.0 years
Complex WSHD					
12-Month Frequency	39,200	2,312	84.5 days	151.7 days	86.7 days
30-Day Frequency	29,700	958	8.4 days	7.8 days	8.4 days
Age at First Use	48,200	802	17.0 years	17.4 years	17.0 years
IVEware					
12-Month Frequency	39,200	2,296	84.6 days	153.3 days	86.8 days
30-Day Frequency	29,700	862	8.4 days	6.8 days	8.4 days
Age at First Use	48,200	802	17.0 years	17.7 years	17.0 years
modPMN-MI					
12-Month Frequency	39,200	2,292	84.5 days	153.4 days	86.7 days
30-Day Frequency	29,700	904	8.4 days	7.2 days	8.4 days
Age at First Use	48,200	802	17.0 years	17.4 years	17.0 years

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table E.14 Before and After Imputation Distributions, by Imputation Method for Inhalants Frequency

Variable	Unweighted Sample Size		Weighted Mean Estimates		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
PMN					
12-Month Frequency	1,000	312	23.1 days	52.2 days	28.6 days
30-Day Frequency	300	57	3.8 days	5.4 days	4.0 days
Age at First Use	6,800	584	17.3 years	16.2 years	17.3 years
Simple WSHD					
12-Month Frequency	1,000	326	23.1 days	51.4 days	28.6 days
30-Day Frequency	300	62	3.8 days	3.5 days	3.7 days
Age at First Use	6,800	584	17.3 years	15.3 years	17.3 years
Complex WSHD					
12-Month Frequency	1,000	351	23.1 days	53.7 days	29.5 days
30-Day Frequency	300	79	3.8 days	5.0 days	4.0 days
Age at First Use	6,800	584	17.3 years	15.8 years	17.3 years
modPMN-MI					
12-Month Frequency	1,000	326	23.1 days	52.3 days	28.7 days
30-Day Frequency	300	66	3.8 days	6.1 days	4.1 days
Age at First Use	6,800	584	17.3 years	16.4 years	17.3 years

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table E.15 Before and After Imputation Distributions, by Imputation Method for Marijuana Frequency

Variable	Unweighted Sample Size		Weighted Mean Estimates		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
PMN					
12-Month Frequency	9,900	1,076	98.4 days	147.8 days	101.9 days
30-Day Frequency	6,100	218	13.0 days	10.0 days	12.9 days
Age at First Use	26,100	255	18.0 years	17.6 years	18.0 years
Simple WSHD					
12-Month Frequency	9,900	1,077	98.4 days	154.6 days	102.3 days
30-Day Frequency	6,100	226	13.0 days	11.2 days	12.9 days
Age at First Use	26,100	255	18.0 years	17.3 years	18.0 years
Complex WSHD					
12-Month Frequency	9,900	1,083	98.4 days	157.2 days	102.6 days
30-Day Frequency	6,100	238	13.0 days	12.9 days	13.0 days
Age at First Use	26,100	255	18.0 years	16.7 years	18.0 years
IVEware					
12-Month Frequency	9,900	1,022	99.0 days	159.7 days	102.9 days
30-Day Frequency	6,100	186	13.0 days	8.8 days	12.9 days
Age at First Use	26,100	255	18.0 years	17.1 years	18.0 years
modPMN-MI					
12-Month Frequency	9,900	1,091	98.4 days	151.7 days	102.1 days
30-Day Frequency	6,100	225	13.0 days	10.6 days	12.9 days
Age at First Use	26,100	255	18.0 years	17.3 years	18.0 years

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table E.16 Before and After Imputation Distributions, by Imputation Method for Pain Relievers Frequency

Variable	Unweighted Sample Size		Weighted Mean Estimates		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
PMN					
12-Month Frequency	4,500	713	43.8 days	67.4 days	46.2 days
Age at First Use	10,300	896	22.1 years	22.8 years	22.1 years
Simple WSHD					
12-Month Frequency	4,500	728	43.8 days	58.7 days	45.3 days
Age at First Use	10,300	896	22.1 years	22.0 years	22.2 years
Complex WSHD					
12-Month Frequency	4,500	723	43.8 days	57.3 days	45.1 days
Age at First Use	10,300	896	22.1 years	22.7 years	22.1 years
modPMN-MI					
12-Month Frequency	4,500	718	43.7 days	47.7 days	44.2 days
Age at First Use	10,300	896	22.1 years	22.7 years	22.1 years

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table E.17 Before and After Imputation Distributions, by Imputation Method for Cocaine Frequency

Variable	Unweighted Sample Size		Weighted Mean Estimates		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
PMN					
12-Month Frequency	1,900	331	36.6 days	79.6 days	43.3 days
30-Day Frequency	600	75	5.2 days	10.5 days	6.0 days
Age at First Use	7,800	213	21.9 years	20.8 years	21.9 years
Simple WSHD					
12-Month Frequency	1,900	337	36.6 days	69.1 days	41.8 days
30-Day Frequency	600	79	5.2 days	8.0 days	5.7 days
Age at First Use	7,800	213	21.9 years	20.8 years	21.9 years
Complex WSHD					
12-Month Frequency	1,900	345	36.6 days	68.3 days	41.7 days
30-Day Frequency	600	83	5.2 days	8.7 days	5.8 days
Age at First Use	7,800	213	21.9 years	20.4 years	21.9 years
modPMN-MI					
12-Month Frequency	1,900	336	36.6 days	76.6 days	43.0 days
30-Day Frequency	600	82	5.2 days	9.7 days	5.9 days
Age at First Use	7,800	213	21.9 years	21.1 years	21.9 years

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table E.18 Before and After Imputation Distributions, by Imputation Method for Heroin Frequency

Variable	Unweighted Sample Size		Weighted Mean Estimates		
	Before Imputation[1]	Imputed or Logically Assigned	Before Imputation[2]	Imputed or Logically Assigned[2]	After Imputation[2]
PMN					
12-Month Frequency	100	35	99.2 days	69.4 days	92.8 days
30-Day Frequency	100	1	15.4 days	20.0* days	15.5 days
Age at First Use	800	44	22.8 years	25.6* years	22.9 years
Simple WSHD					
12-Month Frequency	100	35	99.2 days	63.8* days	91.6 days
30-Day Frequency	100	4	15.4 days	20.8* days	15.5 days
Age at First Use	800	44	22.8 years	17.2* years	22.8 years
Complex WSHD					
12-Month Frequency	100	33	99.2 days	81.3* days	95.5 days
30-Day Frequency	100	1	15.4 days	3.0* days	15.4 days
Age at First Use	800	44	22.8 years	29.2* years	22.9 years
modPMN-MI					
12-Month Frequency	100	35	99.2 days	69.1 days	92.9 days
30-Day Frequency	100	3	15.4 days	4.7* days	13.9 days
Age at First Use	800	44	22.8 years	26.0* years	22.9 years

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

[1] Counts have been rounded to the nearest hundred to ensure respondent confidentiality.

[2] Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

This page intentionally left blank

Appendix F: Estimates of Demographic and Drug Variables by Imputation Method and Significance Results of Pairwise Comparisons

This page intentionally left blank

Table F.1 Estimates of Demographic Variables, by Imputation Method: 12 Years or Older

Demographic Variable	PMN	WSHD	IVEware	modPMN-MI
Race				
American Indian/Alaska Native	1.2	1.2	2.3	1.5
Asian/Other Pacific Islander	4.9	4.9	4.9	4.9
Black/African American	12.3	12.5	12.3	12.4
White	81.6	81.4	80.5	81.3
Hispanic/Latino Origin				
Hispanic/Latino	13.8	13.8	13.8	13.8
Non-Hispanic/Latino	86.2	86.2	86.2	86.2
Marital Status				
Married	52.4	52.4	52.4	52.4
Widowed	5.7	5.7	5.7	5.7
Divorced/Separated	12.6	12.6	12.6	12.6
Never Been Married	29.4	29.4	29.4	29.4
Education Level				
Less than High School	16.3	16.3	16.3	16.3
High School Graduate	30.6	30.6	30.6	30.6
Some College	25.8	25.8	25.8	25.8
College Graduate	27.3	27.2	27.2	27.3

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table F.2 P-values for Comparisons of Demographic Variables, by Imputation Method: 12 Years or Older

Demographic Variable	Global Test	Pairwise Tests					
		PMN vs. WSHD	PMN vs. IVEware	PMN vs. modPMN-MI	WSHD vs. IVEware	WSHD vs. modPMN-MI	IVEware vs. modPMN-MI
Race							
American Indian/Alaska Native	0.0000	0.2565	0.0000	0.0000	0.0000	0.0000	0.0000
Asian/Other Pacific Islander	0.0446	0.0213	0.0289	0.2621	0.9125	0.2515	0.2541
Black/African American	0.0000	0.0000	0.0580	0.0397	0.0000	0.0000	0.0015
White	0.0000	0.0000	0.0000	0.0000	0.0000	0.2075	0.0000
Hispanic/Latino Origin							
Hispanic/Latino	0.2716	N/A	N/A	N/A	N/A	N/A	N/A
Non-Hispanic/Latino	0.2716	N/A	N/A	N/A	N/A	N/A	N/A
Marital Status							
Married	0.4308	N/A	N/A	N/A	N/A	N/A	N/A
Widowed	1.0000	N/A	N/A	N/A	N/A	N/A	N/A
Divorced/Separated	0.7365	N/A	N/A	N/A	N/A	N/A	N/A
Never Been Married	0.1935	N/A	N/A	N/A	N/A	N/A	N/A
Education Level							
Less than High School	0.2710	N/A	N/A	N/A	N/A	N/A	N/A
High School Graduate	0.6059	N/A	N/A	N/A	N/A	N/A	N/A
Some College	0.2405	N/A	N/A	N/A	N/A	N/A	N/A
College Graduate	0.6572	N/A	N/A	N/A	N/A	N/A	N/A

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

N/A = Not applicable. The pairwise tests were not performed since the global test p-value was not significant at the 0.05 level.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table F.3 Estimates of Demographic Variables, by Imputation Method: 12 to 17 Years

Demographic Variable	PMN	WSHD	IVEware	modPMN-MI
Race				
American Indian/Alaska Native	1.6	1.6	2.8	2.1
Asian/Other Pacific Islander	4.9	5.0	5.2	4.9
Black/African American	16.4	16.9	16.5	16.5
White	77.1	76.6	75.6	76.5
Hispanic/Latino Origin				
Hispanic/Latino	18.5	18.5	18.5	18.5
Non-Hispanic/Latino	81.5	81.5	81.5	81.5

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table F.4 P-values for Comparisons of Demographic Variables, by Imputation Method: 12 to 17 Years

Demographic Variable	Global Test	Pairwise Tests					
		PMN vs. WSHD	PMN vs. IVEware	PMN vs. modPMN-MI	WSHD vs. IVEware	WSHD vs. modPMN-MI	IVEware vs. modPMN-MI
Race							
American Indian/Alaska Native	0.0000	0.7621	0.0000	0.0000	0.0000	0.0000	0.0000
Asian/Other Pacific Islander	0.0001	0.0594	0.0000	0.1255	0.0028	0.5784	0.0005
Black/African American	0.0000	0.0000	0.7670	0.1229	0.0000	0.0001	0.1692
White	0.0000	0.0000	0.0000	0.0000	0.0000	0.5869	0.0000
Hispanic/Latino Origin							
Hispanic/Latino	0.4454	N/A	N/A	N/A	N/A	N/A	N/A
Non-Hispanic/Latino	0.4454	N/A	N/A	N/A	N/A	N/A	N/A

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

N/A = Not applicable. The pairwise tests were not performed since the global test p-value was not significant at the 0.05 level.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table F.5 Estimates of Demographic Variables, by Imputation Method: 18 to 25 Years

Demographic Variable	PMN	WSHD	IVEware	modPMN-MI
Race				
American Indian/Alaska Native	1.4	1.4	3.1	1.7
Asian/Other Pacific Islander	5.4	5.5	5.5	5.4
Black/African American	14.7	15.0	14.7	14.7
White	78.6	78.1	76.7	78.1
Hispanic/Latino Origin				
Hispanic/Latino	17.7	17.7	17.7	17.7
Non-Hispanic/Latino	82.3	82.3	82.3	82.3
Marital Status				
Married	12.9	12.9	12.9	12.9
Widowed	0.1	0.1	0.1	0.1
Divorced/Separated	2.0	2.0	2.0	2.0
Never Been Married	85.1	85.1	85.1	85.1
Education Level				
Less than High School	19.0	19.0	19.0	19.0
High School Graduate	34.0	34.0	34.0	34.0
Some College	33.5	33.5	33.5	33.5
College Graduate	13.5	13.5	13.5	13.5

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table F.6 P-values for Comparisons of Demographic Variables, by Imputation Method: 18 to 25 Years

Demographic Variable	Global Test	Pairwise Tests					
		PMN vs. WSHD	PMN vs. IVEware	PMN vs. modPMN-MI	WSHD vs. IVEware	WSHD vs. modPMN-MI	IVEware vs. modPMN-MI
Race							
American Indian/Alaska Native	0.0000	0.5462	0.0000	0.0000	0.0000	0.0000	0.0000
Asian/Other Pacific Islander	0.0011	0.0012	0.0156	0.0305	0.9659	0.1987	0.3543
Black/African American	0.0000	0.0000	0.1966	0.1345	0.0000	0.0002	0.0260
White	0.0000	0.0000	0.0000	0.0000	0.0000	0.7374	0.0000
Hispanic/Latino Origin							
Hispanic/Latino	0.3174	N/A	N/A	N/A	N/A	N/A	N/A
Non-Hispanic/Latino	0.8011	N/A	N/A	N/A	N/A	N/A	N/A
Marital Status							
Married	0.3923	N/A	N/A	N/A	N/A	N/A	N/A
Widowed	1.0000	N/A	N/A	N/A	N/A	N/A	N/A
Divorced/Separated	1.0000	N/A	N/A	N/A	N/A	N/A	N/A
Never Been Married	0.3923	N/A	N/A	N/A	N/A	N/A	N/A
Education Level							
Less than High School	0.3684	N/A	N/A	N/A	N/A	N/A	N/A
High School Graduate	0.3177	N/A	N/A	N/A	N/A	N/A	N/A
Some College	0.3685	N/A	N/A	N/A	N/A	N/A	N/A
College Graduate	0.3688	N/A	N/A	N/A	N/A	N/A	N/A

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

N/A = Not applicable. The pairwise tests were not performed since the global test p-value was not significant at the 0.05 level.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table F.7 Estimates of Demographic Variables, by Imputation Method: 26 Years or Older

Demographic Variable	PMN	WSHD	IVEware	modPMN-MI
Race				
American Indian/Alaska Native	1.1	1.1	2.1	1.4
Asian/Other Pacific Islander	4.8	4.8	4.8	4.8
Black/African American	11.4	11.5	11.3	11.4
White	82.8	82.5	81.8	82.5
Hispanic/Latino Origin				
Hispanic/Latino	12.5	12.5	12.5	12.5
Non-Hispanic/Latino	87.5	87.5	87.5	87.5
Marital Status				
Married	62.7	62.7	62.7	62.7
Widowed	7.0	7.0	7.0	7.0
Divorced/Separated	15.3	15.3	15.3	15.3
Never Been Married	15.0	15.0	15.0	15.0
Education Level				
Less than High School	15.9	15.9	15.9	15.9
High School Graduate	30.0	30.0	30.0	30.0
Some College	24.5	24.5	24.5	24.5
College Graduate	29.6	29.6	29.6	29.6

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table F.8 P-values for Comparisons of Demographic Variables, by Imputation Method: 26 Years or Older

Demographic Variable	Global Test	Pairwise Tests					
		PMN vs. WSHD	PMN vs. IVEware	PMN vs. modPMN-MI	WSHD vs. IVEware	WSHD vs. modPMN-MI	IVEware vs. modPMN-MI
Race							
American Indian/Alaska Native	0.0000	0.3281	0.0000	0.0000	0.0000	0.0003	0.0000
Asian/Other Pacific Islander	0.7204	N/A	N/A	N/A	N/A	N/A	N/A
Black/African American	0.0000	0.0001	0.0470	0.1734	0.0000	0.0004	0.0143
White	0.0000	0.0001	0.0000	0.0000	0.0000	0.2567	0.0000
Hispanic/Latino Origin							
Hispanic/Latino	0.3177	N/A	N/A	N/A	N/A	N/A	N/A
Non-Hispanic/Latino	0.3177	N/A	N/A	N/A	N/A	N/A	N/A
Marital Status							
Married	0.4566	N/A	N/A	N/A	N/A	N/A	N/A
Widowed	1.0000	N/A	N/A	N/A	N/A	N/A	N/A
Divorced/Separated	0.7364	N/A	N/A	N/A	N/A	N/A	N/A
Never Been Married	0.2239	N/A	N/A	N/A	N/A	N/A	N/A
Education Level							
Less than High School	0.2965	N/A	N/A	N/A	N/A	N/A	N/A
High School Graduate	0.5298	N/A	N/A	N/A	N/A	N/A	N/A
Some College	0.2730	N/A	N/A	N/A	N/A	N/A	N/A
College Graduate	0.7378	N/A	N/A	N/A	N/A	N/A	N/A

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

N/A = Not applicable. The pairwise tests were not performed since the global test p-value was not significant at the 0.05 level.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table F.9 Estimates of Drug Variables, by Imputation Method: 12 Years or Older

Drug Variable	PMN	Simple WSHD	Complex WSHD	IVEware	modPMN-MI
Cigarettes Recency					
Within Past 30 Days	24.2	24.3	24.2	24.3	24.2
More than 30 Days Ago but within Past 12 Months	4.2	4.3	4.3	4.2	4.2
More than 12 Months Ago but within Past 3 Years	4.0	3.9	3.9	3.9	4.0
More than 3 Years Ago	32.9	32.8	32.9	32.8	32.9
Never Smoked Cigarettes	34.7	34.7	34.7	34.7	34.7
Cigarettes					
30-Day Frequency (days)	22.6	22.6	22.6	22.6	22.6
Age at First Use (years)	15.7	15.7	15.7	15.7	15.7
Alcohol Recency					
Within Past 30 Days	51.1	51.2	51.3	51.1	51.2
More than 30 Days Ago but within Past 12 Months	14.6	14.5	14.5	14.5	14.5
More than 12 Months Ago	16.6	16.6	16.5	16.7	16.6
Never Used Alcohol	17.7	17.7	17.7	17.7	17.7
Alcohol					
12-Month Frequency (days)	86.9	86.3	86.7	86.8	86.7
30-Day Frequency (days)	8.4	8.4	8.4	8.4	8.4
Age at First Use (years)	17.0	17.0	17.0	17.0	17.0
Inhalants Recency					
Within Past 30 Days	0.3	0.2	0.3	--	0.2
More than 30 Days Ago but within Past 12 Months	0.6	0.6	0.6	--	0.6
More than 12 Months Ago	8.2	8.2	8.2	--	8.2
Never Used Inhalants	90.9	90.9	90.9	--	90.9
Inhalants					
12-Month Frequency (days)	28.6	28.6	29.5	--	28.7
30-Day Frequency (days)	4.0	3.7	4.0	--	4.1
Age at First Use (years)	17.3	17.3	17.3	--	17.3
Marijuana Recency					
Within Past 30 Days	5.8	5.8	5.9	5.8	5.8
More than 30 Days Ago but within Past 12 Months	4.3	4.3	4.3	4.2	4.3
More than 12 Months Ago	30.4	30.4	30.4	30.6	30.4
Never Used Marijuana	59.4	59.5	59.5	59.5	59.5

Table F.9 Estimates of Drug Variables, by Imputation Method: 12 Years or Older (continued)

Drug Variable	PMN	Simple WSHD	Complex WSHD	IVEware	modPMN-MI
Marijuana					
12-Month Frequency (days)	101.9	102.3	102.6	102.9	102.1
30-Day Frequency (days)	12.9	12.9	13.0	12.9	12.9
Age at First Use (years)	18.0	18.0	18.0	18.0	18.0
Pain Relievers Recency					
Within Past 30 Days	2.1	2.1	2.1	--	2.1
More than 30 Days Ago but within Past 12 Months	2.9	3.0	2.9	--	3.0
More than 12 Months Ago	8.3	8.3	8.3	--	8.3
Never Used Pain Relievers	86.7	86.7	86.7	--	86.6
Pain Relievers					
12-Month Frequency (days)	46.2	45.	45.1	--	44.2
Age at First Use (years)	22.1	22.1	22.1	--	22.1
Cocaine Recency					
Within Past 30 Days	0.8	0.9	0.9	--	0.9
More than 30 Days Ago but within Past 12 Months	1.5	1.5	1.5	--	1.5
More than 12 Months Ago	12.2	12.2	12.2	--	12.2
Never Used Cocaine	85.5	85.5	85.5	--	85.5
Cocaine					
12-Month Frequency (days)	43.3	41.8	41.7	--	43.0
30-Day Frequency (days)	6.0	5.7	5.8	--	5.9
Age at First Use (years)	21.9	21.9	21.9	--	21.9
Heroin Recency					
Within Past 30 Days	0.1	0.1	0.1	--	0.1
More than 30 Days Ago but within Past 12 Months	0.1	0.1	0.1	--	0.1
More than 12 Months Ago	1.4	1.4	1.4	--	1.4
Never Used Heroin	98.5	98.5	98.5	--	98.5

Table F.9 Estimates of Drug Variables, by Imputation Method: 12 Years or Older (continued)

Drug Variable	PMN	Simple WSHD	Complex WSHD	IVEware	modPMN-MI
Heroin					
12-Month Frequency (days)	92.8	91.6	95.5	--	92.9
30-Day Frequency (days)	15.5	15.5	15.4	--	13.9
Age at First Use (years)	22.9	22.8	22.9	--	22.9
Ever Used Any Illicit Drug[1]	44.8	44.8	44.8	--	44.8
Used Any Illicit Drug[1] within Past 30 Days	7.5	7.6	7.6	--	7.5

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

-- Not available. Imputations were not performed for inhalants, pain relievers, cocaine, and heroin for IVEware.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

[1] Illicit drugs include marijuana/hashish, cocaine (including crack), heroin, inhalants, and prescription-type pain relievers used nonmedically.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table F.10 P-values for Comparisons of Drug Variables, by Imputation Method: 12 Years or Older

Drug Variable	Global Test	Pairwise Tests									
		PMN vs. Simple WSHD	PMN vs. Complex WSHD	PMN vs. IVEware	PMN vs. modPMN-MI	Simple WSHD vs. Complex WSHD	Simple WSHD vs. IVEware	Simple WSHD vs. modPMN-MI	Complex WSHD vs. IVEware	Complex WSHD vs. modPMN-MI	IVEware vs. modPMN-MI
Cigarettes Recency											
Within Past 30 Days	0.0000	0.2049	0.0773	0.0000	0.3805	0.3157	0.0013	0.2406	0.0000	0.2813	0.0000
More than 30 Days Ago but within Past 12 Months	0.8137	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 12 Months Ago but within Past 3 Years	0.0000	0.0930	0.3108	0.0000	0.3293	0.4863	0.0001	0.4765	0.0001	0.9104	0.0000
More than 3 Years Ago	0.0005	0.4525	0.7446	0.0001	0.8153	0.5521	0.2225	0.4107	0.0006	0.6388	0.0001
Never Smoked Cigarettes	1.0000	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Cigarettes											
30-Day Frequency (days)	0.0000	0.4455	0.1004	0.0000	0.2384	0.5777	0.0001	0.1503	0.0000	0.0426	0.0017
Age at First Use (years)	0.2475	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Alcohol Recency											
Within Past 30 Days	0.0001	0.1000	0.0022	0.1277	0.3733	0.5120	0.0012	0.2182	0.0000	0.0875	0.0031
More than 30 Days Ago but within Past 12 Months	0.3264	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 12 Months Ago	0.0000	0.1206	0.0061	0.0005	0.0640	0.0574	0.0000	0.6745	0.0000	0.0800	0.0001
Never Used Alcohol	0.2914	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Alcohol											
12-Month Frequency (days)	0.0000	0.0000	0.0892	0.2489	0.0343	0.0019	0.0000	0.0000	0.1448	0.8114	0.2143
30-Day Frequency (days)	0.0387	0.0916	0.7051	0.0067	0.0322	0.2460	0.8512	0.5809	0.2154	0.4124	0.5029
Age at First Use (years)	0.3986	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Inhalants Recency											
Within Past 30 Days	0.3313	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 30 Days Ago but within Past 12 Months	0.1860	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 12 Months Ago	0.4145	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Never Used Inhalants	0.6305	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Inhalants											
12-Month Frequency (days)	0.6778	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
30-Day Frequency (days)	0.2718	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Age at First Use (years)	0.3086	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Marijuana Recency											
Within Past 30 Days	0.0007	0.5273	0.1653	0.0026	0.5345	0.2629	0.0019	0.2549	0.0014	0.0732	0.0043
More than 30 Days Ago but within Past 12 Months	0.0005	0.3081	0.2251	0.0006	0.6977	0.5160	0.0061	0.2103	0.0009	0.1590	0.0005
More than 12 Months Ago	0.0000	0.5764	0.2766	0.0000	0.6167	0.3120	0.0000	0.7539	0.0000	0.3518	0.0000
Never Used Marijuana	0.6110	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table F.10 P-values for Comparisons of Drug Variables, by Imputation Method: 12 Years or Older (continued)

Drug Variable	Global Test	Pairwise Tests									
		PMN vs. Simple WSHD	PMN vs. Complex WSHD	PMN vs. IVEware	PMN vs. modPMN-MI	Simple WSHD vs. Complex WSHD	Simple WSHD vs. IVEware	Simple WSHD vs. modPMN-MI	Complex WSHD vs. IVEware	Complex WSHD vs. modPMN-MI	IVEware vs. modPMN-MI
Marijuana											
12-Month Frequency (days)	0.0111	0.1324	0.1650	0.0004	0.3571	0.6112	0.1065	0.3766	0.5663	0.3739	0.0124
30-Day Frequency (days)	0.7460	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Age at First Use (years)	0.6698	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Pain Relievers Recency											
Within Past 30 Days	0.3279	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 30 Days Ago but within Past 12 Months	0.3180	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 12 Months Ago	0.6840	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Never Used Pain Relievers	0.3468	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Pain Relievers											
12-Month Frequency (days)	0.1490	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Age at First Use (years)	0.6076	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Cocaine Recency											
Within Past 30 Days	0.6610	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 30 Days Ago but within Past 12 Months	0.8754	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 12 Months Ago	0.2586	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Never Used Cocaine	0.5984	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Cocaine											
12-Month Frequency (days)	0.3182	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
30-Day Frequency (days)	0.3090	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Age at First Use (years)	0.3425	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Heroin Recency											
Within Past 30 Days	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 30 Days Ago but within Past 12 Months	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 12 Months Ago	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Never Used Heroin	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table F.10 P-values for Comparisons of Drug Variables, by Imputation Method: 12 Years or Older (continued)

Drug Variable	Global Test	Pairwise Tests									
		PMN vs. Simple WSHD	PMN vs. Complex WSHD	PMN vs. IVEware	PMN vs. modPMN-MI	Simple WSHD vs. Complex WSHD	Simple WSHD vs. IVEware	Simple WSHD vs. modPMN-MI	Complex WSHD vs. IVEware	Complex WSHD vs. modPMN-MI	IVEware vs. modPMN-MI
Heroin											
12-Month Frequency (days)	0.1958	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
30-Day Frequency (days)	0.5774	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Age at First Use (years)	0.2629	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Ever Used Any Illicit Drug[1]	0.5797	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Used Any Illicit Drug[1] within Past 30 Days	0.0555	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

-- Not available. Imputations were not performed for inhalants, pain relievers, cocaine, and heroin for IVEware.

N/A = Not applicable. The pairwise tests were not performed since the global test p-value was not significant at the 0.05 level. The comparison among methods for the heroin recency variable could not be performed due to the lack of differences between the estimates as a result of the small number of item nonrespondents that needed to be imputed.

[1] Illicit drugs include marijuana/hashish, cocaine (including crack), heroin, inhalants, and prescription-type pain relievers used nonmedically.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table F.11 Estimates of Drug Variables, by Imputation Method: 12 to 17 Years

Drug Variable	PMN	Simple WSHD	Complex WSHD	IVEware	modPMN-MI
Cigarettes Recency					
Within Past 30 Days	9.8	9.9	9.9	10.4	9.9
More than 30 Days Ago but within Past 12 Months	5.8	5.9	5.9	5.8	5.8
More than 12 Months Ago but within Past 3 Years	4.4	4.4	4.8	4.1	4.4
More than 3 Years Ago	3.6	3.6	3.58	3.4	3.6
Never Smoked Cigarettes	76.3	76.3	76.34	76.3	76.3
Cigarettes					
30-Day Frequency (days)	14.4	14.4	14.4	14.3	14.4
Age at First Use (years)	12.6	12.7	12.6	12.6	12.6
Alcohol Recency					
Within Past 30 Days	15.9	16.0	16.1	15.9	16.0
More than 30 Days Ago but within Past 12 Months	15.8	15.9	15.9	16.0	15.9
More than 12 Months Ago	7.7	7.5	7.5	7.5	7.5
Never Used Alcohol	60.6	60.6	60.6	60.6	60.6
Alcohol					
12-Month Frequency (days)	36.9	36.0	36.6	36.6	36.5
30-Day Frequency (days)	4.6	4.4	4.5	4.5	4.5
Age at First Use (years)	13.2	13.2	13.2	13.2	13.2
Inhalants Recency					
Within Past 30 Days	1.2	1.2	1.3	--	1.2
More than 30 Days Ago but within Past 12 Months	2.8	2.8	2.9	--	2.8
More than 12 Months Ago	5.7	5.6	5.5	--	5.6
Never Used Inhalants	90.4	90.4	90.4	--	90.4
Inhalants					
12-Month Frequency (days)	30.7	30.8	31.3	--	30.9
30-Day Frequency (days)	4.8	4.5	4.6	--	4.7
Age at First Use (years)	12.4	12.5	12.5	--	12.4
Marijuana Recency					
Within Past 30 Days	6.7	6.8	6.7	6.7	6.7
More than 30 Days Ago but within Past 12 Months	5.8	5.7	5.8	5.7	5.8
More than 12 Months Ago	3.7	3.7	3.7	3.8	3.7
Never Used Marijuana	83.8	83.8	83.8	83.8	83.8
Marijuana					
12-Month Frequency (days)	76.8	77.7	77.6	78.0	76.8
30-Day Frequency (days)	10.3	10.2	10.5	10.2	10.3
Age at First Use (years)	13.8	13.8	13.8	13.8	13.8

Table F.11 Estimates of Drug Variables, by Imputation Method: 12 to 17 Years (continued)

Drug Variable	PMN	Simple WSHD	Complex WSHD	IVEware	modPMN-MI
Pain Relievers Recency					
Within Past 30 Days	2.7	2.7	2.7	--	2.6
More than 30 Days Ago but within Past 12 Months	4.0	4.1	4.1	--	4.1
More than 12 Months Ago	3.1	3.0	3.0	--	3.1
Never Used Pain Relievers	90.3	90.3	90.3	--	90.3
Pain Relievers					
12-Month Frequency (days)	42.0	40.3	41.3	--	40.2
Age at First Use (years)	13.3	13.4	13.4	--	13.4
Cocaine Recency					
Within Past 30 Days	0.4	0.4	0.4	--	0.4
More than 30 Days Ago but within Past 12 Months	1.1	1.1	1.1	--	1.1
More than 12 Months Ago	0.6	0.6	0.6	--	0.6
Never Used Cocaine	97.9	97.9	97.9	--	97.9
Cocaine					
12-Month Frequency (days)	30.5	30.4	31.7	--	30.6
30-Day Frequency (days)	5.4	5.3	5.6	--	5.2
Age at First Use (years)	14.8	14.8	14.8	--	14.8
Heroin Recency					
Within Past 30 Days	0.0	0.0	0.0	--	0.0
More than 30 Days Ago but within Past 12 Months	0.1	0.1	0.1	--	0.1
More than 12 Months Ago	0.1	0.1	0.1	--	0.1
Never Used Heroin	99.8	99.8	99.8	--	99.8
Heroin					
12-Month Frequency (days)	50.1	53.3	57.3	--	52.5
30-Day Frequency (days)	9.0*	10.0*	7.4*	--	9.0*
Age at First Use (years)	13.2	13.2	13.2	--	13.1
Ever Used Any Illicit Drug[1]	25.3	25.3	25.3	--	25.3
Used Any Illicit Drug[1] within Past 30 Days	9.1	9.2	9.2	--	9.0

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

-- Not available. Imputations were not performed for inhalants, pain relievers, cocaine, and heroin for IVEware.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

[1] Illicit drugs include marijuana/hashish, cocaine (including crack), heroin, inhalants, and prescription-type pain relievers used nonmedically.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table F.12 P-values for Comparisons of Drug Variables, by Imputation Method: 12 to 17 Years

Drug Variable	Global Test	Pairwise Tests									
		PMN vs. Simple WSHD	PMN vs. Complex WSHD	PMN vs. IVEware	PMN vs. modPMN-MI	Simple WSHD vs. Complex WSHD	Simple WSHD vs. IVEware	Simple WSHD vs. modPMN-MI	Complex WSHD vs. IVEware	Complex WSHD vs. modPMN-MI	IVEware vs. modPMN-MI
Cigarettes Recency											
Within Past 30 Days	0.0000	0.0840	0.2737	0.0000	0.1651	0.1644	0.0000	0.5244	0.0000	0.9850	0.0000
More than 30 Days Ago but within Past 12 Months	0.7911	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 12 Months Ago but within Past 3 Years	0.0000	0.1573	0.2937	0.0000	0.6422	0.6592	0.0000	0.4709	0.0000	0.6631	0.0000
More than 3 Years Ago	0.0027	0.5516	0.5107	0.0070	0.2700	0.9384	0.0011	0.5966	0.0008	0.6332	0.0003
Never Smoked Cigarettes	1.0000	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Cigarettes											
30-Day Frequency (days)	0.0053	0.9187	0.6707	0.0019	0.3085	0.7811	0.0097	0.6329	0.0015	0.1681	0.0134
Age at First Use (years)	0.0001	0.0000	0.0388	0.0144	0.6053	0.0008	0.0730	0.0009	0.5180	0.1950	0.0599
Alcohol Recency											
Within Past 30 Days	0.4147	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 30 Days Ago but within Past 12 Months	0.4308	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 12 Months Ago	0.0453	0.0665	0.0067	0.0059	0.0116	0.2421	0.3731	0.4859	0.7984	0.6018	0.8275
Never Used Alcohol	0.4759	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Alcohol											
12-Month Frequency (days)	0.0062	0.0045	0.4550	0.3469	0.2695	0.0655	0.0008	0.0320	0.9417	0.8385	0.6291
30-Day Frequency (days)	0.0525	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Age at First Use (years)	0.1782	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Inhalants Recency											
Within Past 30 Days	0.0760	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 30 Days Ago but within Past 12 Months	0.4555	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 12 Months Ago	0.0153	0.7008	0.0025	--	0.7301	0.0158	--	0.9843	--	0.0133	--
Never Used Inhalants	0.6933	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Inhalants											
12-Month Frequency (days)	0.9756	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
30-Day Frequency (days)	0.6989	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Age at First Use (years)	0.0623	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Marijuana Recency											
Within Past 30 Days	0.6782	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 30 Days Ago but within Past 12 Months	0.1714	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 12 Months Ago	0.0138	0.5710	0.9434	0.1947	0.0872	0.4689	0.0134	0.8347	0.0527	0.4944	0.0308
Never Used Marijuana	0.3615	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table F.12 P-values for Comparisons of Drug Variables, by Imputation Method: 12 to 17 Years (continued)

Drug Variable	Global Test	Pairwise Tests									
		PMN vs. Simple WSHD	PMN vs. Complex WSHD	PMN vs. IVEware	PMN vs. modPMN-MI	Simple WSHD vs. Complex WSHD	Simple WSHD vs. IVEware	Simple WSHD vs. modPMN-MI	Complex WSHD vs. IVEware	Complex WSHD vs. modPMN-MI	IVEware vs. modPMN-MI
Marijuana											
12-Month Frequency (days)	0.1900	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
30-Day Frequency (days)	0.1085	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Age at First Use (years)	0.3034	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Pain Relievers Recency											
Within Past 30 Days	0.2773	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 30 Days Ago but within Past 12 Months	0.6516	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 12 Months Ago	0.2647	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Never Used Pain Relievers	0.7698	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Pain Relievers											
12-Month Frequency (days)	0.3603	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Age at First Use (years)	0.0151	0.0077	0.5981	--	0.6545	0.0263	--	0.0274	--	0.8963	--
Cocaine Recency											
Within Past 30 Days	0.4924	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 30 Days Ago but within Past 12 Months	0.7912	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 12 Months Ago	0.4856	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Never Used Cocaine	0.3179	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Cocaine											
12-Month Frequency (days)	0.8535	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
30-Day Frequency (days)	0.4114	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Age at First Use (years)	0.9456	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Heroin Recency											
Within Past 30 Days	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 30 Days Ago but within Past 12 Months	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 12 Months Ago	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Never Used Heroin	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table F.12 P-values for Comparisons of Drug Variables, by Imputation Method: 12 to 17 Years (continued)

Drug Variable	Global Test	Pairwise Tests									
		PMN vs. Simple WSHD	PMN vs. Complex WSHD	PMN vs. IVEware	PMN vs. modPMN-MI	Simple WSHD vs. Complex WSHD	Simple WSHD vs. IVEware	Simple WSHD vs. modPMN-MI	Complex WSHD vs. IVEware	Complex WSHD vs. modPMN-MI	IVEware vs. modPMN-MI
Heroin											
12-Month Frequency (days)	0.3534	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
30-Day Frequency (days)	0.3857	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Age at First Use (years)	0.3123	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Ever Used Any Illicit Drug[1]	0.1162	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Used Any Illicit Drug[1] within Past 30 Days	0.1188	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

-- Not available. Imputations were not performed for inhalants, pain relievers, cocaine, and heroin for IVEware.

N/A = Not applicable. The pairwise tests were not performed since the global test p-value was not significant at the 0.05 level. The comparison among methods for the heroin recency variable could not be performed due to the lack of differences between the estimates as a result of the small number of item nonrespondents that needed to be imputed.

[1] Illicit drugs include marijuana/hashish, cocaine (including crack), heroin, inhalants, and prescription-type pain relievers used nonmedically.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table F.13 Estimates of Drug Variables, by Imputation Method: 18 to 25 Years

Drug Variable	PMN	Simple WSHD	Complex WSHD	IVEware	modPMN-MI
Cigarettes Recency					
Within Past 30 Days	36.2	36.2	36.2	36.3	36.2
More than 30 Days Ago but within Past 12 Months	9.0	9.0	9.0	9.0	9.0
More than 12 Months Ago but within Past 3 Years	7.1	7.1	7.1	7.0	7.1
More than 3 Years Ago	12.5	12.5	12.5	12.4	12.5
Never Smoked Cigarettes	35.3	35.3	35.3	35.3	35.3
Cigarettes					
30-Day Frequency (days)	20.3	20.3	20.3	20.3	20.3
Age at First Use (years)	15.1	15.1	15.1	15.1	15.1
Alcohol Recency					
Within Past 30 Days	61.2	61.3	61.3	61.2	61.2
More than 30 Days Ago but within Past 12 Months	16.7	16.6	16.6	16.7	16.7
More than 12 Months Ago	7.3	7.3	7.3	7.3	7.3
Never Used Alcohol	14.8	14.8	14.8	14.8	14.8
Alcohol					
12-Month Frequency (days)	79.0	78.7	79.0	79.0	79.0
30-Day Frequency (days)	7.4	7.3	7.4	7.4	7.4
Age at First Use (years)	16.0	16.0	16.0	16.0	16.0
Inhalants Recency					
Within Past 30 Days	0.4	0.4	0.4	--	0.4
More than 30 Days Ago but within Past 12 Months	1.2	1.2	1.2	--	1.2
More than 12 Months Ago	9.7	9.7	9.6	--	9.7
Never Used Inhalants	88.7	88.7	88.7	--	88.7
Inhalants					
12-Month Frequency (days)	20.3	20.0	22.1	--	20.6
30-Day Frequency (days)	4.5	3.4	3.9	--	4.5
Age at First Use (years)	16.3	16.3	16.3	--	16.3
Marijuana Recency					
Within Past 30 Days	16.4	16.4	16.5	16.3	16.4
More than 30 Days Ago but within Past 12 Months	11.0	11.0	11.0	11.0	11.0
More than 12 Months Ago	23.3	23.4	23.3	23.5	23.3
Never Used Marijuana	49.2	49.2	49.2	49.2	49.2
Marijuana					
12-Month Frequency (days)	112.2	112.2	112.3	112.6	112.1
30-Day Frequency (days)	13.7	13.7	13.8	13.7	13.7
Age at First Use (years)	15.9	15.9	15.9	15.9	15.9

Table F.13 Estimates of Drug Variables, by Imputation Method: 18 to 25 Years (continued)

Drug Variable	PMN	Simple WSHD	Complex WSHD	IVEware	modPMN-MI
Pain Relievers Recency					
Within Past 30 Days	4.6	4.6	4.6	--	4.6
More than 30 Days Ago but within Past 12 Months	7.5	7.5	7.5	--	7.5
More than 12 Months Ago	12.7	12.7	12.8	--	12.7
Never Used Pain Relievers	75.2	75.2	75.2	--	75.2
Pain Relievers					
12-Month Frequency (days)	41.4	41.4	41.7	--	41.8
Age at First Use (years)	17.4	17.4	17.4	--	17.4
Cocaine Recency					
Within Past 30 Days	1.8	1.8	1.8	--	1.8
More than 30 Days Ago but within Past 12 Months	4.6	4.6	4.7	--	4.7
More than 12 Months Ago	8.7	8.6	8.6	--	8.6
Never Used Cocaine	85.0	85.0	85.0	--	85.0
Cocaine					
12-Month Frequency (days)	28.8	29.2	28.0	--	28.1
30-Day Frequency (days)	4.4	4.3	4.3	--	4.4
Age at First Use (years)	18.1	18.1	18.1	--	18.1
Heroin Recency					
Within Past 30 Days	0.2	0.2	0.2	--	0.1
More than 30 Days Ago but within Past 12 Months	0.3	0.3	0.3	--	0.3
More than 12 Months Ago	1.1	1.1	1.1	--	1.1
Never Used Heroin	98.5	98.5	98.5	--	98.5
Heroin					
12-Month Frequency (days)	95.3	93.2	101.6	--	96.8
30-Day Frequency (days)	15.3	15.6	15.3	--	16.3
Age at First Use (years)	18.3	18.3	18.3	--	18.3
Ever Used Any Illicit Drug[1]	56.2	56.2	56.2	--	56.2
Used Any Illicit Drug[1] within Past 30 Days	18.9	19.0	19.0	--	18.9

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

-- Not available. Imputations were not performed for inhalants, pain relievers, cocaine, and heroin for IVEware.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

[1] Illicit drugs include marijuana/hashish, cocaine (including crack), heroin, inhalants, and prescription-type pain relievers used nonmedically.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table F.14 P-values for Comparisons of Drug Variables, by Imputation Method: 18 to 25 Years

Drug Variable	Global Test	Pairwise Tests									
		PMN vs. Simple WSHD	PMN vs. Complex WSHD	PMN vs. IVEware	PMN vs. modPMN-MI	Simple WSHD vs. Complex WSHD	Simple WSHD vs. IVEware	Simple WSHD vs. modPMN-MI	Complex WSHD vs. IVEware	Complex WSHD vs. modPMN-MI	IVEware vs. modPMN-MI
Cigarettes Recency											
Within Past 30 Days	0.0000	0.2413	0.2266	0.0000	0.7383	0.7184	0.0000	0.1449	0.0000	0.1833	0.0000
More than 30 Days Ago but within Past 12 Months	0.9450	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 12 Months Ago but within Past 3 Years	0.2599	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 3 Years Ago	0.0132	0.9588	0.8872	0.0128	0.7748	0.8117	0.0017	0.8814	0.0063	0.7306	0.0170
Never Smoked Cigarettes	1.0000	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Cigarettes											
30-Day Frequency (days)	0.1481	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Age at First Use (years)	0.4004	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Alcohol Recency											
Within Past 30 Days	0.1004	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 30 Days Ago but within Past 12 Months	0.4059	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 12 Months Ago	0.1269	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Never Used Alcohol	0.3309	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Alcohol											
12-Month Frequency (days)	0.0003	0.0066	0.7921	0.7683	0.8076	0.0072	0.0000	0.0020	0.4358	0.6449	0.9789
30-Day Frequency (days)	0.0024	0.0012	0.4482	0.0199	0.2450	0.0048	0.0379	0.0004	0.1394	0.0942	0.0007
Age at First Use (years)	0.3497	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Inhalants Recency											
Within Past 30 Days	0.7900	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 30 Days Ago but within Past 12 Months	0.4395	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 12 Months Ago	0.4347	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Never Used Inhalants	0.3507	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Inhalants											
12-Month Frequency (days)	0.3638	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
30-Day Frequency (days)	0.3117	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Age at First Use (years)	0.0733	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Marijuana Recency											
Within Past 30 Days	0.0887	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 30 Days Ago but within Past 12 Months	0.7012	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 12 Months Ago	0.0001	0.2567	0.7755	0.0000	0.6481	0.1896	0.0004	0.4344	0.0000	0.3929	0.0000
Never Used Marijuana	0.2412	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table F.14 P-values for Comparisons of Drug Variables, by Imputation Method: 18 to 25 Years (continued)

Drug Variable	Global Test	Pairwise Tests									
		PMN vs. Simple WSHD	PMN vs. Complex WSHD	PMN vs. IVEware	PMN vs. modPMN-MI	Simple WSHD vs. Complex WSHD	Simple WSHD vs. IVEware	Simple WSHD vs. modPMN-MI	Complex WSHD vs. IVEware	Complex WSHD vs. modPMN-MI	IVEware vs. modPMN-MI
Marijuana											
12-Month Frequency (days)	0.1803	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
30-Day Frequency (days)	0.1121	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Age at First Use (years)	0.6161	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Pain Relievers Recency											
Within Past 30 Days	0.6005	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 30 Days Ago but within Past 12 Months	0.7317	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 12 Months Ago	0.8303	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Never Used Pain Relievers	0.2820	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Pain Relievers											
12-Month Frequency (days)	0.7083	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Age at First Use (years)	0.2573	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Cocaine Recency											
Within Past 30 Days	0.6081	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 30 Days Ago but within Past 12 Months	0.4908	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 12 Months Ago	0.1054	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Never Used Cocaine	0.3959	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Cocaine											
12-Month Frequency (days)	0.2893	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
30-Day Frequency (days)	0.4102	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Age at First Use (years)	0.5212	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Heroin Recency											
Within Past 30 Days	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 30 Days Ago but within Past 12 Months	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 12 Months Ago	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Never Used Heroin	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table F.14 P-values for Comparisons of Drug Variables, by Imputation Method: 18 to 25 Years (continued)

Drug Variable	Global Test	Pairwise Tests									
		PMN vs. Simple WSHD	PMN vs. Complex WSHD	PMN vs. IVEware	PMN vs. modPMN-MI	Simple WSHD vs. Complex WSHD	Simple WSHD vs. IVEware	Simple WSHD vs. modPMN-MI	Complex WSHD vs. IVEware	Complex WSHD vs. modPMN-MI	IVEware vs. modPMN-MI
Heroin											
12-Month Frequency (days)	0.1178	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
30-Day Frequency (days)	0.5426	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Age at First Use (years)	0.3213	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Ever Used Any Illicit Drug[1]	0.4312	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Used Any Illicit Drug[1] within Past 30 Days	0.1925	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

-- Not available. Imputations were not performed for inhalants, pain relievers, cocaine, and heroin for IVEware.

N/A = Not applicable. The pairwise tests were not performed since the global test p-value was not significant at the 0.05 level. The comparison among methods for the heroin recency variable could not be performed due to the lack of differences between the estimates as a result of the small number of item nonrespondents that needed to be imputed.

[1] Illicit drugs include marijuana/hashish, cocaine (including crack), heroin, inhalants, and prescription-type pain relievers used nonmedically.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table F.15 Estimates of Drug Variables, by Imputation Method: 26 Years or Older

Drug Variable	PMN	Simple WSHD	Complex WSHD	IVEware	modPMN-MI
Cigarettes Recency					
Within Past 30 Days	24.1	24.1	24.1	24.1	24.1
More than 30 Days Ago but within Past 12 Months	3.2	3.2	3.2	3.2	3.2
More than 12 Months Ago but within Past 3 Years	3.4	3.3	3.4	3.3	3.3
More than 3 Years Ago	40.3	40.2	40.3	40.3	40.3
Never Smoked Cigarettes	29.1	29.1	29.1	29.1	29.1
Cigarettes					
30-Day Frequency (days)	23.7	23.7	23.7	23.7	23.8
Age at First Use (years)	15.9	15.9	15.9	15.9	15.9
Alcohol Recency					
Within Past 30 Days	54.1	54.2	54.2	54.0	54.1
More than 30 Days Ago but within Past 12 Months	14.0	14.0	14.0	13.9	14.0
More than 12 Months Ago	19.4	19.4	19.4	19.5	19.4
Never Used Alcohol	12.5	12.5	12.5	12.5	12.5
Alcohol					
12-Month Frequency (days)	91.6	91.0	91.3	91.5	91.4
30-Day Frequency (days)	8.7	8.7	8.7	8.7	8.7
Age at First Use (years)	17.4	17.4	17.4	17.4	17.4
Inhalants Recency					
Within Past 30 Days	0.1	0.1	0.1	--	0.1
More than 30 Days Ago but within Past 12 Months	0.2	0.2	0.2	--	0.2
More than 12 Months Ago	8.3	8.3	8.3	--	8.3
Never Used Inhalants	91.4	91.4	91.4	--	91.4
Inhalants					
12-Month Frequency (days)	32.6	32.6	33.3	--	32.2
30-Day Frequency (days)	2.5	2.6	2.7	--	2.7
Age at First Use (years)	18.2	18.2	18.2	--	18.3
Marijuana Recency					
Within Past 30 Days	3.9	3.9	3.9	3.8	3.9
More than 30 Days Ago but within Past 12 Months	2.9	2.9	2.9	2.9	2.9
More than 12 Months Ago	35.2	35.2	35.2	35.3	35.2
Never Used Marijuana	58.0	58.0	58.0	58.0	58.0
Marijuana					
12-Month Frequency (days)	100.8	101.5	101.9	102.2	101.3
30-Day Frequency (days)	12.9	13.0	13.0	13.0	13.0
Age at First Use (years)	18.7	18.7	18.7	18.7	18.7

Table F.15 Estimates of Drug Variables, by Imputation Method: 26 Years or Older (continued)

Drug Variable	PMN	Simple WSHD	Complex WSHD	IVEware	modPMN-MI
Pain Relievers Recency					
Within Past 30 Days	1.6	1.6	1.6	--	1.6
More than 30 Days Ago but within Past 12 Months	2.0	2.0	2.0	--	2.0
More than 12 Months Ago	8.3	8.3	8.3	--	8.3
Never Used Pain Relievers	88.2	88.2	88.2	--	88.1
Pain Relievers					
12-Month Frequency (days)	50.0	48.8	48.1	--	46.5
Age at First Use (years)	24.8	24.7	24.7	--	24.7
Cocaine Recency					
Within Past 30 Days	0.7	0.8	0.8	--	0.8
More than 30 Days Ago but within Past 12 Months	1.0	1.0	1.0	--	1.0
More than 12 Months Ago	14.3	14.3	14.3	--	14.3
Never Used Cocaine	84.0	84.0	84.0	--	84.0
Cocaine					
12-Month Frequency (days)	54.1	51.1	51.6	--	54.0
30-Day Frequency (days)	6.6	6.2	6.4	--	6.6
Age at First Use (years)	22.6	22.6	22.6	--	22.6
Heroin Recency					
Within Past 30 Days	0.1	0.1	0.1	--	0.1
More than 30 Days Ago but within Past 12 Months	0.1	0.1	0.1	--	0.0
More than 12 Months Ago	1.6	1.6	1.6	--	1.6
Never Used Heroin	98.3	98.3	98.3	--	98.3
Heroin					
12-Month Frequency (days)	96.1	95.0	95.3	--	94.7
30-Day Frequency (days)	15.7	15.7	15.7	--	13.0
Age at First Use (years)	23.7	23.7	23.7	--	23.7
Ever Used Any Illicit Drug[1]	45.4	45.4	45.4	--	45.5
Used Any Illicit Drug[1] within Past 30 Days	5.3	5.4	5.4	--	5.3

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

-- Not available. Imputations were not performed for inhalants, pain relievers, cocaine, and heroin for IVEware.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

[1] Illicit drugs include marijuana/hashish, cocaine (including crack), heroin, inhalants, and prescription-type pain relievers used nonmedically.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table F.16 P-values for Comparisons of Drug Variables, by Imputation Method: 26 Years or Older

Drug Variable	Global Test	Pairwise Tests									
		PMN vs. Simple WSHD	PMN vs. Complex WSHD	PMN vs. IVEware	PMN vs. modPMN-MI	Simple WSHD vs. Complex WSHD	Simple WSHD vs. IVEware	Simple WSHD vs. modPMN-MI	Complex WSHD vs. IVEware	Complex WSHD vs. modPMN-MI	IVEware vs. modPMN-MI
Cigarettes Recency											
Within Past 30 Days	0.3920	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 30 Days Ago but within Past 12 Months	0.4077	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 12 Months Ago but within Past 3 Years	0.4048	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 3 Years Ago	0.2799	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Never Smoked Cigarettes	1.0000	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Cigarettes											
30-Day Frequency (days)	0.2642	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Age at First Use (years)	0.4082	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Alcohol Recency											
Within Past 30 Days	0.0023	0.2266	0.0181	0.1615	0.3766	0.5966	0.0069	0.4930	0.0006	0.2565	0.0035
More than 30 Days Ago but within Past 12 Months	0.3054	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 12 Months Ago	0.0001	0.7370	0.2431	0.0000	0.3179	0.0866	0.0000	0.6657	0.0000	0.2088	0.0000
Never Used Alcohol	0.3673	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Alcohol											
12-Month Frequency (days)	0.0000	0.0000	0.1332	0.3825	0.0555	0.0098	0.0000	0.0002	0.1490	0.8720	0.1788
30-Day Frequency (days)	0.0589	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Age at First Use (years)	0.3974	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Inhalants Recency											
Within Past 30 Days	0.4902	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 30 Days Ago but within Past 12 Months	0.3690	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 12 Months Ago	0.4405	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Never Used Inhalants	0.3928	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Inhalants											
12-Month Frequency (days)	0.2795	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
30-Day Frequency (days)	0.6675	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Age at First Use (years)	0.2264	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Marijuana Recency											
Within Past 30 Days	0.0037	0.6336	0.2177	0.0098	0.4605	0.2607	0.0113	0.2871	0.0067	0.0923	0.0140
More than 30 Days Ago but within Past 12 Months	0.0067	0.5248	0.2341	0.0042	0.8124	0.3741	0.0125	0.4382	0.0036	0.1947	0.0042
More than 12 Months Ago	0.0027	0.5118	0.2840	0.0006	0.7477	0.3844	0.0003	0.5790	0.0002	0.3028	0.0001
Never Used Marijuana	0.5423	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table F.16 P-values for Comparisons of Drug Variables, by Imputation Method: 26 Years or Older (continued)

Drug Variable	Global Test	Pairwise Tests									
		PMN vs. Simple WSHD	PMN vs. Complex WSHD	PMN vs. IVEware	PMN vs. modPMN-MI	Simple WSHD vs. Complex WSHD	Simple WSHD vs. IVEware	Simple WSHD vs. modPMN-MI	Complex WSHD vs. IVEware	Complex WSHD vs. modPMN-MI	IVEware vs. modPMN-MI
Marijuana											
12-Month Frequency (days)	0.0864	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
30-Day Frequency (days)	0.9605	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Age at First Use (years)	0.7462	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Pain Relievers Recency											
Within Past 30 Days	0.4356	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 30 Days Ago but within Past 12 Months	0.3302	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 12 Months Ago	0.7423	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Never Used Pain Relievers	0.3241	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Pain Relievers											
12-Month Frequency (days)	0.2662	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Age at First Use (years)	0.4320	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Cocaine Recency											
Within Past 30 Days	0.6613	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 30 Days Ago but within Past 12 Months	0.7513	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
More than 12 Months Ago	0.6617	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Never Used Cocaine	0.7647	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Cocaine											
12-Month Frequency (days)	0.2916	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
30-Day Frequency (days)	0.2579	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Age at First Use (years)	0.3474	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Heroin Recency											
Within Past 30 Days	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 30 Days Ago but within Past 12 Months	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
More than 12 Months Ago	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Never Used Heroin	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table F.16 P-values for Comparisons of Drug Variables, by Imputation Method: 26 Years or Older (continued)

Drug Variable	Global Test	Pairwise Tests									
		PMN vs. Simple WSHD	PMN vs. Complex WSHD	PMN vs. IVEware	PMN vs. modPMN-MI	Simple WSHD vs. Complex WSHD	Simple WSHD vs. IVEware	Simple WSHD vs. modPMN-MI	Complex WSHD vs. IVEware	Complex WSHD vs. modPMN-MI	IVEware vs. modPMN-MI
Heroin											
12-Month Frequency (days)	0.2730	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
30-Day Frequency (days)	0.7446	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Age at First Use (years)	0.3628	N/A	N/A	--	N/A	N/A	--	N/A	--	N/A	--
Ever Used Any Illicit Drug[1]	0.5074	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Used Any Illicit Drug[1] within Past 30 Days	0.2977	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

-- Not available. Imputations were not performed for inhalants, pain relievers, cocaine, and heroin for IVEware.

N/A = Not applicable. The pairwise tests were not performed since the global test p-value was not significant at the 0.05 level. The comparison among methods for the heroin recency variable could not be performed due to the lack of differences between the estimates as a result of the small number of item nonrespondents that needed to be imputed.

[1] Illicit drugs include marijuana/hashish, cocaine (including crack), heroin, inhalants, and prescription-type pain relievers used nonmedically.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

This page intentionally left blank

Appendix G: Pairwise Comparisons of Imputation Methods

This page intentionally left blank

Figure G.1 Pairwise Comparisons of Imputation Methods for Race, Percentages

Results for White, 12 to 17 Years:

IVEware$_{75.6}$ < modPMN-MI$_{76.5}$ WSHD$_{76.6}$ < PMN$_{77.1}$

Results for Black/African American, 12 to 17 Years:

PMN$_{16.4}$ IVEware$_{16.5}$ modPMN-MI$_{16.5}$ < WSHD$_{16.9}$

Results for Asian/Other Pacific Islander, 12 to 17 Years:

PMN$_{4.9}$ modPMN-MI$_{4.9}$ WSHD$_{5.0}$ < IVEware$_{5.2}$

Results for American Indian/Alaska Native, 12 to 17 Years:

PMN$_{1.6}$ WSHD$_{1.6}$ < modPMN-MI$_{2.1}$ < IVEware$_{2.8}$

Results for White, 18 to 25 Years:

IVEware$_{76.7}$ < modPMN-MI$_{78.1}$ WSHD$_{78.1}$ < PMN$_{78.6}$

Results for Black/African American, 18 to 25 Years:

IVEware$_{14.7}$ PMN$_{14.7}$ modPMN-MI$_{14.7}$ < WSHD$_{15.0}$

Results for Asian/Other Pacific Islander, 18 to 25 Years:

PMN$_{5.4}$ < modPMN-MI$_{5.4}$ IVEware$_{5.5}$ WSHD$_{5.5}$

Results for American Indian/Alaska Native, 18 to 25 Years:

PMN$_{1.4}$ WSHD$_{1.4}$ < modPMN-MI$_{1.7}$ < IVEware$_{3.1}$

Results for White, 26 Years or Older:

IVEware$_{81.8}$ < modPMN-MI$_{82.5}$ WSHD$_{82.5}$ < PMN$_{82.8}$

Results for Black/African American, 26 Years or Older:

IVEware$_{11.3}$ < PMN$_{11.4}$ modPMN-MI$_{11.4}$ < WSHD$_{11.5}$

Results for American Indian/Alaska Native, 26 Years or Older:

PMN$_{1.1}$ WSHD$_{1.1}$ < modPMN-MI$_{1.4}$ < IVEware$_{2.1}$

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

Note: An underline indicates imputation methods that were not found to be statistically different from each other.
Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Figure G.2 Pairwise Comparisons of Imputation Methods for Cigarettes Recency, Percentages

Results for Within Past 30 Days, 12 to 17 Years:

PMN$_{9.8}$ Complex WSHD$_{9.9}$ modPMN-MI$_{9.9}$ Simple WSHD$_{9.9}$ < IVEware$_{10.4}$

Results for More than 12 Months Ago but within Past 3 Years, 12 to 17 Years:

IVEware$_{4.1}$ < Simple WSHD$_{4.4}$ Complex WSHD$_{4.4}$ modPMN-MI$_{4.4}$ PMN$_{4.4}$

Results for More than 3 Years Ago, 12 to 17 Years:

IVEware$_{3.4}$ < PMN$_{3.6}$ Simple WSHD$_{3.6}$ Complex WSHD$_{3.6}$ modPMN-MI$_{3.6}$

Results for Within Past 30 Days, 18 to 25 Years:

modPMN-MI$_{36.2}$ PMN$_{36.2}$ Simple WSHD$_{36.2}$ Complex WSHD$_{36.2}$ < IVEware$_{36.3}$

Results for More than 3 Years Ago, 18 to 25 Years:

IVEware$_{12.4}$ < Simple WSHD$_{12.5}$ modPMN-MI$_{12.5}$ PMN$_{12.5}$ Complex WSHD$_{12.5}$

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

Note: An underline indicates imputation methods that were not found to be statistically different from each other.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Figure G.3 Pairwise Comparisons of Imputation Methods for Alcohol Recency, Percentages

Results for Within Past 30 Days, 12 Years or Older:

IVEware$_{51.1}$ PMN$_{51.1}$ modPMN-MI$_{51.2}$ Simple WSHD$_{51.2}$ Complex WSHD$_{51.3}$

Results for More than 12 Months Ago, 12 Years or Older:

Complex WSHD$_{16.5}$ Simple WSHD$_{16.6}$ modPMN-MI$_{16.6}$ PMN$_{16.6}$ < IVEware$_{16.7}$

Results for More than 12 Months Ago, 12 to 17 Years:

Complex WSHD$_{7.5}$ IVEware$_{7.5}$ modPMN-MI$_{7.5}$ Simple WSHD$_{7.5}$ PMN$_{7.7}$

Results for Within Past 30 Days, 26 Years or Older:

IVEware$_{54.0}$ PMN$_{54.1}$ modPMN-MI$_{54.1}$ Simple WSHD$_{54.2}$ Complex WSHD$_{54.2}$

Results for More than 12 Months Ago, 26 Years or Older:

Complex WSHD$_{19.4}$ Simple WSHD$_{19.4}$ PMN$_{19.4}$ modPMN-MI$_{19.4}$ < IVEware$_{19.5}$

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

Note: An underline indicates imputation methods that were not found to be statistically different from each other.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Figure G.4 Pairwise Comparisons of Imputation Methods for Marijuana Recency, Percentages

Results for Within Past 30 Days, 12 Years or Older:

IVEware$_{5.8}$ < modPMN-MI$_{5.8}$ PMN$_{5.8}$ Simple WSHD$_{5.8}$ Complex WSHD$_{5.9}$

Results for More than 30 Days Ago but within Past 12 Months, 12 Years or Older:

IVEware$_{4.2}$ < Complex WSHD$_{4.3}$ Simple WSHD$_{4.3}$ PMN$_{4.3}$ modPMN-MI$_{4.3}$

Results for More than 12 Months Ago, 12 Years or Older:

Complex WSHD$_{30.4}$ Simple WSHD$_{30.4}$ modPMN-MI$_{30.4}$ PMN$_{30.4}$ < IVEware$_{30.6}$

Results for More than 12 Months Ago, 12 to 17 Years:

modPMN-MI$_{3.7}$ Simple WSHD$_{3.7}$ PMN$_{3.7}$ Complex WSHD$_{3.7}$ IVEware$_{3.8}$

Results for More than 12 Months Ago, 18 to 25 Years:

PMN$_{23.3}$ Complex WSHD$_{23.3}$ modPMN-MI$_{23.3}$ Simple WSHD$_{23.4}$ < IVEware$_{23.5}$

Results for Within Past 30 Days, 26 Years or Older:

IVEware$_{3.8}$ < modPMN-MI$_{3.9}$ PMN$_{3.9}$ Simple WSHD$_{3.9}$ Complex WSHD$_{3.9}$

Results for More than 30 Days Ago but within Past 12 Months, 26 Years or Older:

IVEware$_{2.9}$ < Complex WSHD$_{2.9}$ Simple WSHD$_{2.9}$ PMN$_{2.9}$ modPMN-MI$_{2.9}$

Results for More than 12 Months Ago, 26 Years or Older:

Complex WSHD$_{35.2}$ Simple WSHD$_{35.2}$ PMN$_{35.2}$ modPMN-MI$_{35.2}$ < IVEware$_{35.3}$

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

Note: An underline indicates imputation methods that were not found to be statistically different from each other.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Figure G.5 Pairwise Comparisons of Imputation Methods for Inhalants Recency, Percentages

Results for More than 12 Months Ago, 12 to 17 Years:

Complex $WSHD_{5.5}$ < Simple $WSHD_{5.6}$ modPMN-$MI_{5.6}$ $PMN_{5.7}$

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

Note: An underline indicates imputation methods that were not found to be statistically different from each other.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Figure G.6 Pairwise Comparisons of Imputation Methods for Drug Use 12-Month Frequency, Days

Results for Alcohol 12-Month Frequency, 12 Years or Older[1]:

Simple WSHD$_{86.3}$ < Complex WSHD$_{86.7}$ modPMN-MI$_{86.7}$ IVEware$_{86.8}$ PMN$_{86.9}$

Results for Marijuana 12-Month Frequency, 12 Years or Older:

PMN$_{101.9}$ modPMN-MI$_{102.1}$ Simple WSHD$_{102.3}$ Complex WSHD$_{102.6}$ IVEware$_{102.9}$

Results for Alcohol 12-Month Frequency, 12 to 17 Years[2]:

Simple WSHD$_{36.0}$ < modPMN-MI$_{36.5}$ Complex WSHD$_{36.6}$ IVEware$_{36.6}$ PMN$_{36.9}$

Results for Alcohol 12-Month Frequency, 18 to 25 Years:

Simple WSHD$_{78.7}$ < Complex WSHD$_{79.0}$ PMN$_{79.0}$ IVEware$_{79.0}$ modPMN-MI$_{79.0}$

Results for Alcohol 12-Month Frequency, 26 Years or Older:

Simple WSHD$_{91.0}$ < Complex WSHD$_{91.3}$ modPMN-MI$_{91.4}$ IVEware$_{91.5}$ PMN$_{91.6}$

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

Note: An underline indicates imputation methods that were not found to be statistically different from each other.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

[1] PMN and modPMN-MI are significantly different but are displayed with an underline due to other nonsignificant comparisons.

[2] Complex WSHD and simple WSHD are not significantly different from each other but could not be displayed with an underline due to other significant comparisons.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Figure G.7 Pairwise Comparisons of Imputation Methods for Drug Use 30-Day Frequency, Days

Results for Cigarettes 30-Day Frequency, 12 Years or Older[1]:

IVEware$_{22.6}$ < PMN$_{22.6}$ modPMN-MI$_{22.6}$ Simple WSHD$_{22.6}$ Complex WSHD$_{22.6}$

Results for Alcohol 30-Day Frequency, 12 Years or Older[2]:

IVEware$_{8.4}$ Simple WSHD$_{8.4}$ modPMN-MI$_{8.4}$ Complex WSHD$_{8.4}$ PMN$_{8.4}$

Results for Cigarettes 30-Day Frequency, 12 to 17 Years:

IVEware$_{14.3}$ < modPMN-MI$_{14.4}$ Simple WSHD$_{14.4}$ PMN$_{14.4}$ Complex WSHD$_{14.4}$

Results for Alcohol 30-Day Frequency, 18 to 25 Years:

Simple WSHD$_{7.3}$ < IVEware$_{7.4}$ Complex WSHD$_{7.4}$ PMN$_{7.4}$ modPMN-MI$_{7.4}$

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

Note: An underline indicates imputation methods that were not found to be statistically different from each other.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

[1] Complex WSHD and modPMN-MI are significantly different but are displayed with an underline due to other nonsignificant comparisons.

[2] PMN and simple WSHD are not significantly different from each other but could not be displayed with an underline due to other significant comparisons.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Figure G.8 Pairwise Comparisons of Imputation Methods for Age at First Use, Years

Results for Cigarettes Age at First Use, 12 to 17 Years:

$\underline{\text{PMN}_{12.6}\quad \text{modPMN-MI}_{12.6}}$ $\quad \text{Complex WSHD}_{12.6} \quad$ $\underline{\text{IVEware}_{12.6}\quad \text{Simple WSHD}_{12.7}}$

Results for Pain Relievers Age at First Use, 12 to 17 Years:

$\underline{\text{PMN}_{13.3}\quad \text{Complex WSHD}_{13.4}\quad \text{modPMN-MI}_{13.4}}$ $< \text{Simple WSHD}_{13.4}$

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

Note: An underline indicates imputation methods that were not found to be statistically different from each other.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Appendix H: Two-Way Drug Comparisons by Imputation Method

This page intentionally left blank

Table H.1 Percentages of Lifetime Drug Use of Specified Drugs Given No Other Lifetime Drug Use, by Imputation Method and Age Group

Age Group	Drug Measure of Interest	PMN (%)	Simple WSHD (%)	Complex WSHD (%)	modPMN-MI (%)
12+	Lifetime Cigarette Use	17.6	17.6	17.6	17.6
	Lifetime Alcohol Use	53.6	53.6	53.6	53.5
	Lifetime Inhalants Use	1.7	1.7	1.7	1.7
	Lifetime Marijuana Use	0.8	0.8	0.8	0.8
	Lifetime Pain Relievers Use	2.6	2.6	2.6	2.6
	Lifetime Cocaine Use	0.2	0.2	0.2	0.2
	Lifetime Heroin Use	0.0*	0.0*	0.0*	0.0*
12-17	Lifetime Cigarette Use	4.9	4.9	4.9	4.9
	Lifetime Alcohol Use	20.9	20.8	20.8	20.8
	Lifetime Inhalants Use	3.4	3.5	3.4	3.4
	Lifetime Marijuana Use	0.9	0.9	0.9	0.9
	Lifetime Pain Relievers Use	2.5	2.6	2.6	2.6
	Lifetime Cocaine Use	0.0	0.0	0.0	0.0
	Lifetime Heroin Use	0.0*	0.0*	0.0*	0.0*
18-25	Lifetime Cigarette Use	15.3	15.3	15.3	15.3
	Lifetime Alcohol Use	59.3	59.3	59.3	59.3
	Lifetime Inhalants Use	1.2	1.2	1.2	1.2
	Lifetime Marijuana Use	1.7	1.7	1.7	1.7
	Lifetime Pain Relievers Use	4.1	4.2	4.1	4.1
	Lifetime Cocaine Use	0.0*	0.0*	0.0*	0.0*
	Lifetime Heroin Use	0.0*	0.0*	0.0*	0.0*
26+	Lifetime Cigarette Use	26.0	26.0	26.0	26.0
	Lifetime Alcohol Use	64.4	64.4	64.4	64.4
	Lifetime Inhalants Use	0.3	0.3	0.3	0.3
	Lifetime Marijuana Use	0.5	0.5	0.5	0.5
	Lifetime Pain Relievers Use	2.2	2.2	2.2	2.2
	Lifetime Cocaine Use	0.3	0.3	0.3	0.3
	Lifetime Heroin Use	0.0*	0.0*	0.0*	0.0*

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

*Low precision.

Note: The pairwise tests were not performed since the global test p-value was not significant at the 0.05 level.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Table H.2 Percentages of Lifetime Cocaine or Heroin Use Given Cigarette, Marijuana, or Alcohol Lifetime Use, by Imputation Method and Age Group

Age Group	Population of Interest	Drug Measure of Interest	PMN (%)	Simple WSHD (%)	Complex WSHD (%)	modPMN-MI (%)
12+	Lifetime Cigarette Use	Lifetime Cocaine Use	21.1	21.1	21.1	21.1
		Lifetime Heroin Use	2.3	2.3	2.3	2.3
	Lifetime Marijuana Use	Lifetime Cocaine Use	34.8	34.7	34.8	34.8
		Lifetime Heroin Use	3.7	3.7	3.7	3.7
	Lifetime Alcohol Use	Lifetime Cocaine Use	17.4	17.4	17.4	17.4
		Lifetime Heroin Use	1.8	1.8	1.8	1.8
12-17	Lifetime Cigarette Use	Lifetime Cocaine Use	8.4	8.4	8.4	8.4
		Lifetime Heroin Use	0.7	0.7	0.7	0.7
	Lifetime Marijuana Use	Lifetime Cocaine Use	12.4	12.4	12.4	12.4
		Lifetime Heroin Use	1.1	1.1	1.1	1.1
	Lifetime Alcohol Use	Lifetime Cocaine Use	5.3	5.3	5.3	5.3
		Lifetime Heroin Use	0.5	0.5	0.5	0.5
18-25	Lifetime Cigarette Use	Lifetime Cocaine Use	22.6	22.6	22.6	22.6
		Lifetime Heroin Use	2.3	2.3	2.3	2.3
	Lifetime Marijuana Use	Lifetime Cocaine Use	28.9	28.8	28.9	28.9
		Lifetime Heroin Use	2.9	2.9	2.9	2.9
	Lifetime Alcohol Use	Lifetime Cocaine Use	17.5	17.5	17.5	17.5
		Lifetime Heroin Use	1.8	1.8	1.8	1.8
26+	Lifetime Cigarette Use	Lifetime Cocaine Use	21.4	21.4	21.4	21.4
		Lifetime Heroin Use	2.3	2.3	2.3	2.3
	Lifetime Marijuana Use	Lifetime Cocaine Use	37.2	37.1	37.2	37.2
		Lifetime Heroin Use	3.9	3.9	3.9	3.9
	Lifetime Alcohol Use	Lifetime Cocaine Use	18.1	18.1	18.1	18.1
		Lifetime Heroin Use	1.9	1.9	1.9	1.9

modPMN-MI = modified predictive mean neighborhood multiple imputation; PMN = predictive mean neighborhood; WSHD = weighted sequential hot deck.

Note: The pairwise tests were not performed since the global test p-value was not significant at the 0.05 level.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2007.

Appendix I: Feasibility Assessment for Using the Other Pair Member's Value in Imputation, by Variable

This page intentionally left blank

Table I.1 Assessment of Feasibility of Using the Other Pair Member's Value in Imputation, by Demographic Variable, 2009 NSDUH

Demographic Variable	Number of Respondents in Domain	Number of Responding Pairs in Domain	Percentage of Pair Agreement	Number Missing in Domain	Number of Nonrespondent Paired with Respondents	Percentage of Nonrespondent Eligible for Edit	Number of Nonrespondent Paired with Respondents that Agree after PMN	Percentage of Eligible Nonrespondent with Pair Agreement after PMN
Born in US	68,700	20,184	91.99	27	8	29.63	7	87.50
Education Level	68,700	20,189	18.11	10	2	20.00	0	0.00
Employment Status: 18+	46,074	9,245	42.90	34	7	20.59	1	14.29
Employment Status	57,817	14,037	37.71	39	10	25.64	2	20.00
Immigrant Age of Entry in US	7,365	1,461	0.62	13	1	7.69	0	0.00
Hispanic/Latino Origin Group	10,777	3,000	91.17	140	41	29.29	32	78.05
Single/Multiple Hispanic/Latino Origin Group	10,777	2,996	89.85	148	44	29.73	32	72.73
Hispanic/Latino Origin	68,700	20,115	95.48	116	71	61.21	65	91.55
Marital Status	57,817	14,043	69.54	8	4	50.00	3	75.00
Race Variable: 15 Levels	68,700	18,883	90.30	2,987	652	21.83	317	48.62
Race: Asian	68,700	19,031	98.27	2,633	579	21.99	526	90.85
Race: Black/African American	68,700	19,112	97.40	2,435	545	22.38	463	84.95
Race: American Indian/Alaska Native	68,700	19,100	95.24	2,456	536	21.82	340	63.43
Race: Native Hawaiian	68,700	19,156	99.56	2,332	514	22.04	511	99.42
Race: Other Pacific Islander	68,700	19,156	99.18	2,332	514	22.04	490	95.33
Race: White	68,700	19,032	94.12	2,606	590	22.64	367	62.20

PMN = predictive mean neighborhood.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

Table I.2 Assessment of Feasibility of Using the Other Pair Member's Value in Imputation, by Drug Variable, 2009 NSDUH

Drug Variable	Number of Respondents in Domain	Number of Responding Pairs in Domain	Percentage of Pair Agreement	Number Missing in Domain	Number of Nonrespondent Paired with Respondents	Percentage of Nonrespondent Eligible for Edit	Number of Nonrespondent Paired with Respondents that Agree after PMN	Percentage of Eligible Nonrespondent with Pair Agreement after PMN
Alcohol Age at First Use	49,078	10,423	12.88	740	303	40.95	26	8.58
Alcohol 5+ Drinks Past Month	31,003	4,769	26.17	1,181	336	28.45	85	25.30
Alcohol Day of First Use	49,078	0	N/A	49,078	0	0.00	N/A	N/A
Alcohol 30-Day Frequency	31,003	4,875	10.75	881	232	26.33	21	9.05
Alcohol 12-Month Frequency	41,901	7,661	7.31	2,214	769	34.73	11	1.43
Alcohol Lifetime Use	68,700	20,177	67.43	22	15	68.18	10	66.67
Alcohol Month of First Use	49,078	207	11.11	44,005	1,891	4.30	154	8.14
Alcohol Recency	49,078	10,334	58.37	918	388	42.27	138	35.57
Alcohol Year of First Use	49,078	246	50.81	43,615	1,998	4.58	75	3.75
Pain Relievers Age at First Use	11,256	943	10.71	617	62	10.05	8	12.90
Pain Relievers Day of First Use	11,256	0	N/A	11,256	0	0.00	N/A	N/A
Pain Relievers 12-Month Frequency	5,279	229	6.11	773	53	6.86	3	5.66
Pain Relievers Lifetime Use	68,700	20,048	77.25	252	142	56.35	117	82.39
Pain Relievers Month of First Use	11,256	30	6.67	9,516	241	2.53	17	7.05
Pain Relievers Recency	11,256	948	45.04	457	62	13.57	26	41.94
Pain Relievers Year of First Use	11,256	32	53.13	9,343	256	2.74	19	7.42
Any Other Pain Relievers Lifetime Use	68,700	20,034	77.43	276	156	56.52	118	75.64

Table I.2 Assessment of Feasibility of Using the Other Pair Member's Value in Imputation, by Drug Variable, 2009 NSDUH (continued)

Drug Variable	Number of Respondents in Domain	Number of Responding Pairs in Domain	Percentage of Pair Agreement	Number Missing in Domain	Number of Nonrespondent Paired with Respondents	Percentage of Nonrespondent Eligible for Edit	Number of Nonrespondent Paired with Respondents that Agree after PMN	Percentage of Eligible Nonrespondent with Pair Agreement after PMN
Cigarettes Daily Day of First Use	18,112	0	N/A	18,112	0	0.00	N/A	N/A
Cigarettes Daily Month of First Use	18,112	17	11.76	16,988	237	1.40	15	6.33
Cigarettes Daily Year of First Use	18,112	19	63.16	16,920	250	1.48	9	3.60
Cigarettes Daily Age at First Use	18,112	2,149	12.01	150	28	18.67	3	10.71
Tend To Avoid Places That Don't Allow Smoking	16,699	2,169	49.70	18	2	11.11	0	0.00
Cravings for Cigarettes like Force Can't Control	16,699	2,174	28.89	6	1	16.67	0	0.00
Crave Cigarettes when Haven't Smoked for a Few Hours	16,699	2,174	28.66	3	1	33.33	0	0.00
Amount of Smoking Has Increased Since Started Smoking	16,699	2,170	25.53	2	0	0.00	N/A	N/A
Number Cigarettes Smoked per Day Influenced by Other Things	16,699	2,166	23.64	6	2	33.33	0	0.00
Need To Smoke To Feel Less Irritable	16,699	2,170	26.18	14	4	28.57	0	0.00
Must Smoke Much More Now Before Start To Feel Anything	16,699	2,158	30.91	50	16	32.00	0	0.00
Number Cigarettes Smoked per Day Often Changes	16,699	2,169	25.31	7	1	14.29	0	0.00

Table I.2 Assessment of Feasibility of Using the Other Pair Member's Value in Imputation, by Drug Variable, 2009 NSDUH (continued)

Drug Variable	Number of Respondents in Domain	Number of Responding Pairs in Domain	Percentage of Pair Agreement	Number Missing in Domain	Number of Nonrespondent Paired with Respondents	Percentage of Nonrespondent Eligible for Edit	Number of Nonrespondent Paired with Respondents that Agree after PMN	Percentage of Eligible Nonrespondent with Pair Agreement after PMN
Feel in Control of Smoking	16,699	2,168	27.54	12	6	50.00	0	0.00
Smoking Not Affected by Other Things	16,699	2,163	26.31	19	5	26.32	0	0.00
No Travel by Airplane Because No Smoking Allowed	16,699	2,167	69.87	27	5	18.52	0	0.00
Cigar Age at First Use	21,532	2,513	11.90	505	109	21.58	10	9.17
Cigar Day of First Use	21,532	0	N/A	21,532	0	0.00	N/A	N/A
Cigar 30-Day Frequency	4,623	224	23.21	74	3	4.05	0	0.00
Smoke Cigarettes Regularly throughout the Day	16,699	2,172	31.40	0	0	4.05	N/A	N/A
Smoke Same Number of Cigarettes from Day to Day	16,699	2,169	28.68	7	1	14.29	0	0.00
Smoke Same Number of Cigarettes on Weekends As on Weekdays	16,699	2,169	30.38	8	2	25.00	0	0.00
Cigar Lifetime Use	68,700	20,182	65.36	10	10	100.0	7	70.00
Cigar Month of First Use	21,532	99	10.10	18,831	508	2.70	41	8.07
Worry about Running Out of Cigarettes	16,699	2,175	34.76	1	0	0.00	N/A	N/A
Cigar Recency	21,532	2,526	32.66	407	94	23.10	34	36.17
Cigar Year of First Use	21,532	107	61.68	18,619	547	2.94	28	5.12
Need To Smoke a Lot More To Be Satisfied	16,699	2,174	33.44	6	0	0.00	N/A	N/A
Smoke Lots of Cigarettes in an Hour, Then No Cigarettes Until Hours Later	16,699	2,165	33.39	19	4	21.05	0	0.00

Table I.2 Assessment of Feasibility of Using the Other Pair Member's Value in Imputation, by Drug Variable, 2009 NSDUH (continued)

Drug Variable	Number of Respondents in Domain	Number of Responding Pairs in Domain	Percentage of Pair Agreement	Number Missing in Domain	Number of Nonrespondent Paired with Respondents	Percentage of Nonrespondent Eligible for Edit	Number of Nonrespondent Paired with Respondents that Agree after PMN	Percentage of Eligible Nonrespondent with Pair Agreement after PMN
Chewing Tobacco Age at First Use	7,818	419	13.84	161	17	10.56	0	0.00
Chewing Tobacco Day of First Use	7,818	0	N/A	7,818	0	0.00	N/A	N/A
Chewing Tobacco 30-Day Frequency	1,220	42	11.90	19	0	0.00	N/A	N/A
Chewing Tobacco Lifetime Use	68,700	20,186	81.79	9	6	66.67	4	66.67
Chewing Tobacco Month of First Use	7,818	18	27.78	7,017	88	1.25	8	9.09
Chewing Tobacco Recency	7,818	422	38.15	150	13	8.67	3	23.08
Chewing Tobacco Year of First Use	7,818	22	45.45	6,947	91	1.31	7	7.69
Cigarettes Age at First Use	36,407	6,246	11.22	504	148	29.37	13	8.78
Cigarettes Day of First Use	36,407	0	N/A	36,407	0	0.00	N/A	N/A
Cigarettes Daily Use	36,407	6,374	62.55	52	20	38.46	13	65.00
Cigarettes 30-Day Frequency	16,699	2,145	44.24	177	35	19.77	7	20.00
Cigarettes Month of First Use	36,407	62	3.23	33,739	817	2.42	62	7.59
Cigarettes Recency	36,407	6,225	50.18	475	168	35.37	29	17.26
Cigarettes Year of First Use	36,407	68	48.53	33,544	886	2.64	28	3.16
Cocaine Age at First Use	7,994	593	11.64	168	29	17.26	1	3.45
Cocaine Day of First Use	7,994	0	N/A	7,994	0	0.00	N/A	N/A

Table I.2 Assessment of Feasibility of Using the Other Pair Member's Value in Imputation, by Drug Variable, 2009 NSDUH (continued)

Drug Variable	Number of Respondents in Domain	Number of Responding Pairs in Domain	Percentage of Pair Agreement	Number Missing in Domain	Number of Nonrespondent Paired with Respondents	Percentage of Nonrespondent Eligible for Edit	Number of Nonrespondent Paired with Respondents that Agree after PMN	Percentage of Eligible Nonrespondent with Pair Agreement after PMN
Cocaine 30-Day Frequency	524	18	27.78	57	0	0.00	N/A	N/A
Cocaine 12-Month Frequency	1,810	78	5.13	235	15	6.38	3	20.00
Cocaine Lifetime Use	68,700	20,181	84.27	29	11	37.93	8	72.73
Cocaine Month of First Use	7,994	10	20.00	7,367	93	1.26	8	8.60
Cocaine Recency	7,994	597	63.48	163	24	14.72	12	50.00
Cocaine Year of First Use	7,994	12	58.33	7,323	100	1.37	3	3.00
Crack Age at First Use	1,920	68	11.76	12	2	16.67	0	0.00
Crack Day of First Use	1,920	0	N/A	1,920	0	0.00	N/A	N/A
Crack 30-Day Frequency	99	3	33.33	6	0	0.00	N/A	N/A
Crack 12-Month Frequency	289	5	20.00	37	2	5.41	0	0.00
Crack Lifetime Use	68,700	20,180	95.62	31	12	38.71	11	91.67
Crack Month of First Use	1,920	1	0.00	1,805	7	0.39	2	28.57
Crack Recency	1,920	68	73.53	30	2	6.67	0	0.00
Crack Year of First Use	1,920	1	100.0	1,801	7	0.39	0	0.00
Ecstasy Age at First Use	4,740	324	10.19	62	13	20.97	0	0.00
Ecstasy Day of First Use	4,740	0	N/A	4,740	0	0.00	N/A	N/A
Ecstasy Lifetime Use	68,700	20,161	90.52	68	31	45.59	31	100.0
Ecstasy Month of First Use	4,740	31	25.81	3,860	62	1.61	8	12.90
Ecstasy Recency	4,740	316	69.94	108	21	19.44	9	42.86
Ecstasy Year of First Use	4,740	32	62.50	3,815	67	1.76	4	5.97
Hallucinogens Age at First Use	9,412	848	14.50	234	40	17.09	3	7.50
Hallucinogens Day of First Use	9,412	0	N/A	9,412	0	0.00	N/A	N/A

Table I.2 Assessment of Feasibility of Using the Other Pair Member's Value in Imputation, by Drug Variable, 2009 NSDUH (continued)

Drug Variable	Number of Respondents in Domain	Number of Responding Pairs in Domain	Percentage of Pair Agreement	Number Missing in Domain	Number of Nonrespondent Paired with Respondents	Percentage of Nonrespondent Eligible for Edit	Number of Nonrespondent Paired with Respondents that Agree after PMN	Percentage of Eligible Nonrespondent with Pair Agreement after PMN
Hallucinogens 30-Day Frequency	657	20	30.00	58	4	6.90	0	0.00
Hallucinogens 12-Month Frequency	2,515	132	21.21	310	32	10.32	3	9.38
Hallucinogens Lifetime Use	68,700	20,071	83.10	217	121	55.76	106	87.60
Hallucinogens Month of First Use	9,412	33	24.24	8,298	164	1.98	12	7.32
Hallucinogens Recency	9,412	821	67.11	323	68	21.05	29	42.65
Hallucinogens Year of First Use	9,412	37	62.16	8,191	167	2.04	7	4.19
Heroin Age at First Use	920	23	13.04	6	0	0.00	N/A	N/A
Heroin Day of First Use	920	0	N/A	920	0	0.00	N/A	N/A
Heroin 30-Day Frequency	82	4	0.00	7	0	0.00	N/A	N/A
Heroin 12-Month Frequency	225	4	0.00	45	2	4.44	0	0.00
Heroin Lifetime Use	68,700	20,172	97.79	37	20	54.05	20	100.0
Heroin Month of First Use	920	1	0.00	793	7	0.88	0	0.00
Heroin Recency	920	22	68.18	22	1	4.55	0	0.00
Heroin Year of First Use	920	1	0.00	786	7	0.89	1	14.29
Any Other Hallucinogens Lifetime Use	68,700	20,075	87.62	207	117	56.52	106	90.60
Inhalants Age at First Use	6,889	362	12.71	398	23	5.78	2	8.70
Inhalants Day of First Use	6,889	0	N/A	6,889	0	0.00	N/A	N/A
Inhalants 30-Day Frequency	354	5	40.00	64	0	0.00	N/A	N/A
Inhalants 12-Month Frequency	1,378	30	16.67	319	9	2.82	2	22.22
Inhalants Lifetime Use	68,700	20,133	84.38	109	59	54.13	48	81.36

Table I.2 Assessment of Feasibility of Using the Other Pair Member's Value in Imputation, by Drug Variable, 2009 NSDUH (continued)

Drug Variable	Number of Respondents in Domain	Number of Responding Pairs in Domain	Percentage of Pair Agreement	Number Missing in Domain	Number of Nonrespondent Paired with Respondents	Percentage of Nonrespondent Eligible for Edit	Number of Nonrespondent Paired with Respondents that Agree after PMN	Percentage of Eligible Nonrespondent with Pair Agreement after PMN
Inhalants Month of First Use	6,889	18	16.67	6,056	78	1.29	5	6.41
Inhalants Recency	6,889	361	71.75	309	25	8.09	11	44.00
Inhalants Year of First Use	6,889	21	76.19	5,950	80	1.34	5	6.25
LSD Age at First Use	4,788	255	13.33	92	9	9.78	1	11.11
LSD Day of First Use	4,788	0	N/A	4,788	0	0.00	N/A	N/A
LSD Lifetime Use	68,700	20,157	90.12	59	35	59.32	30	85.71
LSD Month of First Use	4,788	5	20.00	4,439	36	0.81	2	5.56
LSD Recency	4,788	252	82.54	85	12	14.12	4	33.33
LSD Year of First Use	4,788	5	60.00	4,418	39	0.88	4	10.26
Marijuana Age at First Use	26,974	4,254	12.20	218	56	25.69	8	14.29
Marijuana Day of First Use	26,974	0	N/A	26,974	0	0.00	N/A	N/A
Marijuana 30-Day Frequency	7,185	667	15.59	259	33	12.74	1	3.03
Marijuana 12-Month Frequency	12,194	1,217	7.23	1,298	258	19.88	2	0.78
Marijuana Lifetime Use	68,700	20,176	66.07	36	16	44.44	11	68.75
Marijuana Month of First Use	26,974	74	8.11	24,362	692	2.84	51	7.37
Marijuana Recency	26,974	4,180	49.83	449	130	28.95	49	37.69
Marijuana Year of First Use	26,974	84	51.19	24,162	742	3.07	30	4.04
Methamphetamine Age at First Use	2,189	78	14.10	49	1	2.04	0	0.00
Methamphetamine Day of First Use	2,189	0	N/A	2,189	0	0.00	N/A	N/A

Table I.2 Assessment of Feasibility of Using the Other Pair Member's Value in Imputation, by Drug Variable, 2009 NSDUH (continued)

Drug Variable	Number of Respondents in Domain	Number of Responding Pairs in Domain	Percentage of Pair Agreement	Number Missing in Domain	Number of Nonrespondent Paired with Respondents	Percentage of Nonrespondent Eligible for Edit	Number of Nonrespondent Paired with Respondents that Agree after PMN	Percentage of Eligible Nonrespondent with Pair Agreement after PMN
Methamphetamine 12-Month Frequency	335	6	0.00	54	0	0.00	N/A	N/A
Methamphetamine Lifetime Use	68,700	20,165	94.82	59	26	44.07	25	96.15
Methamphetamine Month of First Use	2,189	0	N/A	2,074	7	0.34	0	0.00
Methamphetamine Recency	2,189	79	73.42	39	1	2.56	0	0.00
Methamphetamine Year of First Use	2,189	0	N/A	2,066	8	0.39	1	12.50
OxyContin Age at First Use	2,406	95	11.58	82	5	6.10	0	0.00
OxyContin Day of First Use	2,406	0	N/A	2,406	0	0.00	N/A	N/A
OxyContin 12-Month Frequency	903	16	0.00	173	8	4.62	0	0.00
OxyContin Lifetime Use	68,700	20,070	94.14	208	120	57.69	106	88.33
OxyContin Month of First Use	2,406	2	0.00	1,934	33	1.71	2	6.06
OxyContin Recency	2,406	94	61.70	107	6	5.61	0	0.00
OxyContin Year of First Use	2,406	3	100.0	1,889	33	1.75	3	9.09
PCP Age at First Use	1,180	25	16.00	40	0	0.00	N/A	N/A
PCP Day of First Use	1,180	0	N/A	1,180	0	0.00	N/A	N/A
PCP Lifetime Use	68,700	20,163	97.04	57	29	50.88	29	100.0
PCP Month of First Use	1,180	0	N/A	1,118	2	0.18	0	0.00
PCP Recency	1,180	25	84.00	24	0	0.00	N/A	N/A
PCP Year of First Use	1,180	0	N/A	1,114	2	0.18	0	0.00
Pipe Lifetime Use	68,700	20,180	86.92	16	12	75.00	10	83.33
Pipe Past Month Use	5,809	239	73.64	3	0	0.00	N/A	N/A

Table I.2 Assessment of Feasibility of Using the Other Pair Member's Value in Imputation, by Drug Variable, 2009 NSDUH (continued)

Drug Variable	Number of Respondents in Domain	Number of Responding Pairs in Domain	Percentage of Pair Agreement	Number Missing in Domain	Number of Nonrespondent Paired with Respondents	Percentage of Nonrespondent Eligible for Edit	Number of Nonrespondent Paired with Respondents that Agree after PMN	Percentage of Eligible Nonrespondent with Pair Agreement after PMN
Sedatives Age at First Use	1,389	26	11.54	51	0	0.00	N/A	N/A
Sedatives Day of First Use	1,389	0	N/A	1,389	0	0.00	N/A	N/A
Sedatives 12-Month Frequency	292	1	0.00	50	1	2.00	0	0.00
Sedatives Lifetime Use	68,700	20,084	96.58	177	107	60.45	100	93.46
Sedatives Month of First Use	1,389	1	0.00	1,237	2	0.16	0	0.00
Sedatives Recency	1,389	26	76.92	33	0	0.00	N/A	N/A
Sedatives Year of First Use	1,389	1	100.0	1,227	2	0.16	0	0.00
Smokeless Tobacco Age at First Use	11,967	955	13.72	336	51	15.18	5	9.80
Smokeless Tobacco Day of First Use	11,967	0	N/A	11,967	0	0.00	N/A	N/A
Smokeless Tobacco Lifetime Use	68,700	20,174	75.22	35	18	51.43	13	72.22
Smokeless Tobacco Month of First Use	11,967	34	11.76	10,531	221	2.10	14	6.33
Smokeless Tobacco Recency	11,967	966	37.06	289	39	13.49	12	30.77
Smokeless Tobacco Year of First Use	11,967	39	53.85	10,434	234	2.24	11	4.70
Snuff Age at First Use	9,754	711	14.77	221	31	14.03	2	6.45
Snuff Day of First Use	9,754	0	N/A	9,754	0	0.00	N/A	N/A
Snuff 30-Day Frequency	2,605	113	24.78	50	2	4.00	1	50.00
Snuff Lifetime Use	68,700	20,173	78.76	37	19	51.35	11	57.89
Snuff Month of First Use	9,754	32	6.25	8,440	171	2.03	13	7.60
Snuff Recency	9,754	709	37.09	211	32	15.17	9	28.13

Table I.2 Assessment of Feasibility of Using the Other Pair Member's Value in Imputation, by Drug Variable, 2009 NSDUH (continued)

Drug Variable	Number of Respondents in Domain	Number of Responding Pairs in Domain	Percentage of Pair Agreement	Number Missing in Domain	Number of Nonrespondent Paired with Respondents	Percentage of Nonrespondent Eligible for Edit	Number of Nonrespondent Paired with Respondents that Agree after PMN	Percentage of Eligible Nonrespondent with Pair Agreement after PMN
Snuff Year of First Use	9,754	37	51.35	8,353	182	2.18	10	5.49
Any Other Stimulants Lifetime Use	68,700	20,123	91.52	124	68	54.84	64	94.12
Stimulants Age at First Use	4,887	257	11.67	163	12	7.36	0	0.00
Stimulants Day of First Use	4,887	0	N/A	4,887	0	0.00	N/A	N/A
Stimulants 12-Month Frequency	1,307	38	5.26	224	13	5.80	0	0.00
Stimulants Lifetime Use	68,700	20,119	89.14	131	72	54.96	68	94.44
Stimulants Month of First Use	4,887	13	15.38	4,326	44	1.02	3	6.82
Stimulants Recency	4,887	257	59.14	115	13	11.30	4	30.77
Stimulants Year of First Use	4,887	13	61.54	4,289	48	1.12	1	2.08
Tranquilizers Age at First Use	5,816	382	11.26	190	22	11.58	2	9.09
Tranquilizers Day of First Use	5,816	0	N/A	5,816	0	0.00	N/A	N/A
Tranquilizers 12-Month Frequency	2,140	82	9.76	267	13	4.87	0	0.00
Tranquilizers Lifetime Use	68,700	20,110	88.08	141	81	57.45	72	88.89
Tranquilizers Month of First Use	5,816	14	14.29	4,946	96	1.94	10	10.42
Tranquilizers Recency	5,816	393	56.23	126	12	9.52	8	66.67
Tranquilizers Year of First Use	5,816	18	72.22	4,873	96	1.97	3	3.13

N/A = not applicable; PMN = predictive mean neighborhood.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

Table I.3 Assessment of Feasibility of Using the Other Pair Member's Value in Imputation, by Health Insurance Variable, 2009 NSDUH

Health Insurance Variable	Number of Respondents in Domain	Number of Responding Pairs in Domain	Percentage of Pair Agreement	Number Missing in Domain	Number of Nonrespondent Paired with Respondents	Percentage of Nonrespondent Eligible for Edit	Number of Nonrespondent Paired with Respondents that Agree after PMN	Percentage of Eligible Nonrespondent with Pair Agreement after PMN
Health Insurance: CHAMPUS	68,700	20,070	97.58	235	115	48.94	104	90.43
Overall Health Insurance, as Defined by the 1999 Survey Method	68,700	19,894	81.09	539	268	49.72	144	53.73
Overall Health Insurance, as Defined by the 2001 Survey Method	68,700	19,882	83.13	558	278	49.82	170	61.15
Overall Health Insurance, as Defined by the Constituent Variables Method	68,700	19,885	83.12	552	275	49.82	171	62.18
Health Insurance: CAIDCHIP	68,700	19,899	86.52	568	267	47.01	178	66.67
Health Insurance: Medicare	68,700	20,052	96.47	256	135	52.73	126	93.33
Other Health Insurance	13,320	2,069	91.30	183	31	16.94	24	77.42
Private Health Insurance, Consistent with Pre-1999 Surveys	68,700	19,932	82.73	465	240	51.61	122	50.83
Private Health Insurance, as Defined by the Constituent Variables Method	68,700	19,932	82.73	465	240	51.61	141	58.75

PMN = predictive mean neighborhood.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

Table I.4 Assessment of Feasibility of Using the Other Pair Member's Value in Imputation, by Income Variable, 2009 NSDUH

Income Variable	Number of Respondents in Domain	Number of Responding Pairs in Domain	Percentage of Pair Agreement	Number Missing in Domain	Number of Nonrespondent Paired with Respondents	Percentage of Nonrespondent Eligible for Edit	Number of Nonrespondent Paired with Respondents that Agree after PMN	Percentage of Eligible Nonrespondent with Pair Agreement after PMN
Total Family Income > or < $20,000	43,949	9,722	93.30	2,255	564	25.01	355	62.94
Total Family Income (Finer Categories)	60,074	15,584	66.84	6,359	1,736	27.30	248	14.29
Family Received Public Assistance	68,700	19,924	97.75	492	243	49.39	196	80.66
Family Received Social Security or Railroad Retirement Payments	68,700	19,855	95.63	660	310	46.97	229	73.87
Family Received Supplemental Security Income	68,700	19,718	96.57	935	424	45.35	339	79.95
Family Received Welfare/ Job Placement/Child Care	68,700	19,986	96.61	371	190	51.21	161	84.74
Family Received Income from Job	68,700	20,101	93.54	193	81	41.97	60	74.07
Respondent/Other Family Member Received Food Stamps	68,700	20,056	95.69	262	125	47.71	88	70.40
Respondent's Total Income > or < $20,000	68,700	19,849	67.70	847	277	32.70	178	64.26
Respondent's Total Income (Finer Categories)	68,700	19,447	23.45	1,797	618	34.39	75	12.14
Number of Months on Welfare	4,587	854	78.81	218	45	20.64	13	28.89

PMN = predictive mean neighborhood.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

Table I.5 Assessment of Feasibility of Using the Other Pair Member's Value in Imputation, by Pair Variable, 2009 NSDUH

Pair Variable	Number of Respondents in Domain	Number of Responding Pairs in Domain	Percentage of Pair Agreement	Number Missing in Domain	Number of Nonrespondent Paired with Respondents	Percentage of Nonrespondent Eligible for Edit	Number of Nonrespondent Paired with Respondents that Agree after PMN	Percentage of Eligible Nonrespondent with Pair Agreement after PMN
Number of Spouse-Spouse Pairs in Household	68,700	20,153	100.0	94	0	0.00	N/A	N/A
Number of Spouse-Spouse Pairs with Children	68,700	20,165	100.0	61	0	0.00	N/A	N/A
Child-Parent, Child Focus, Child is 12-14	68,700	20,147	100.0	96	0	0.00	N/A	N/A
Child-Parent, Child Focus, Child is 12-17	68,700	20,128	100.0	137	0	0.00	N/A	N/A
Child-Parent, Child Focus, Child is 12-20	68,700	20,098	100.0	202	0	0.00	N/A	N/A
Child-Parent, Parent Focus, Child is 12-14	68,700	20,056	100.0	277	0	0.00	N/A	N/A
Child-Parent, Parent Focus, Child is 12-17	68,700	19,969	100.0	456	0	0.00	N/A	N/A
Child-Parent, Parent Focus, Child is 12-20	68,700	19,943	100.0	514	0	0.00	N/A	N/A
Sibling-Sibling (12-14/15-17), 15-17 Focus	68,700	20,167	100.0	69	0	0.00	N/A	N/A
Sibling-Sibling (12-17/18-25), 18-25 Focus	68,700	20,158	100.0	88	0	0.00	N/A	N/A
Multiplicity: Child-Parent, Child Focus, Child is 12-14	4,170	2,037	100.0	96	0	0.00	N/A	N/A
Multiplicity: Child-Parent, Child Focus, Child is 12-17	7,950	3,891	100.0	168	0	0.00	N/A	N/A
Multiplicity: Child-Parent, Child Focus, Child is 12-20	9,464	4,635	100.0	194	0	0.00	N/A	N/A
Multiplicity: Child-Parent, Child Focus, Child is 15-17	3,780	1,854	100.0	72	0	0.00	N/A	N/A
Multiplicity: Child-Parent, Parent Focus, Child is 12-14	4,170	2,085	100.0	0	0	0.00	N/A	N/A

Table I.5 Assessment of Feasibility of Using the Other Pair Member's Value in Imputation, by Pair Variable, 2009 NSDUH (continued)

Pair Variable	Number of Respondents in Domain	Number of Responding Pairs in Domain	Percentage of Pair Agreement	Number Missing in Domain	Number of Nonrespondent Paired with Respondents	Percentage of Nonrespondent Eligible for Edit	Number of Nonrespondent Paired with Respondents that Agree after PMN	Percentage of Eligible Nonrespondent with Pair Agreement after PMN
Multiplicity: Child-Parent, Parent Focus, Child is 12-17	7,950	3,975	100.0	0	0	0.00	N/A	N/A
Multiplicity: Child-Parent, Parent Focus, Child is 12-20	9,464	4,732	100.0	0	0	0.00	N/A	N/A
Multiplicity: Child-Parent, Parent Focus, Child is 15-17	3,780	1,890	100.0	0	0	0.00	N/A	N/A
Multiplicity: Sibling-Sibling (12-14/15-17), 15-17 Focus	4,382	2,189	100.0	4	0	0.00	N/A	N/A
Multiplicity: Sibling-Sibling (12-14/15-17), 12-14 Focus	4,382	2,191	100.0	0	0	0.00	N/A	N/A
Multiplicity: Sibling-Sibling (12-17/18-25), 18-25 Focus	5,106	2,550	100.0	6	0	0.00	N/A	N/A
Multiplicity: Sibling-Sibling (12-17/18-25), 12-17 Focus	5,106	2,544	100.0	18	0	0.00	N/A	N/A
Family Pair Relationship Indicator	40,384	20,000	100.0	384	0	0.00	N/A	N/A

N/A = not applicable; PMN = predictive mean neighborhood.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

Table I.6 Assessment of Feasibility of Using the Other Pair Member's Value in Imputation, by Roster Variable, 2009 NSDUH

Roster Variable	Number of Respondents in Domain	Number of Responding Pairs in Domain	Percentage of Pair Agreement	Number Missing in Domain	Number of Nonrespondent Paired with Respondents	Percentage of Nonrespondent Eligible for Edit	Number of Nonrespondent Paired with Respondents that Agree after PMN	Percentage of Eligible Nonrespondent with Pair Agreement after PMN
Presence of Family Members in Household	68,700	20,178	97.85	43	14	32.56	13	92.86
Number of Respondent's Family Members in Household excluding Foster Relationships	68,700	20,156	89.07	74	36	48.65	14	38.89
Number of Respondent's Family Members in Household including Foster Relationships	68,700	20,156	89.28	74	36	48.65	15	41.67
Number of People in Household Aged ≥ 65	68,700	19,994	99.43	370	168	45.41	143	85.12
Number of People in Household	68,700	20,181	94.71	35	11	31.43	3	27.27
Number of Respondent's Family Members in Household Aged < 18 including Foster Relationships	68,700	20,120	94.85	127	70	55.12	32	45.71
Number of Children in Household Aged < 18	68,700	20,091	96.96	190	90	47.37	39	43.33
Number of Respondent's Family Members in Household Aged < 18 excluding Foster Relationships	68,700	20,124	94.69	122	66	54.10	29	43.94

PMN = predictive mean neighborhood.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

Appendix J: Mental Health Variable Item Nonresponse Rates and Estimates by Imputation Cell Categories

This page intentionally left blank

Table J.1 Item Nonresponse Rates for Edited Past Month K6 Variables, 2010 NSDUH

K6 Variable	Description	Total	Not Missing	Missing	Unweighted Item Nonresponse Rate (Percent)	Weighted Item Nonresponse Rate (Percent)
DSTNRV30	How often felt nervous past 30 days	45,844	45,681	163	0.36	0.32
DSTHOP30	How often felt hopeless past 30 days	45,844	45,659	185	0.40	0.39
DSTRST30	How often felt restless/fidgety past 30 days	45,844	45,623	221	0.48	0.47
DSTCHR30	How often felt sad nothing could cheer you up past 30 days	45,844	45,681	163	0.36	0.36
DSTEFF30	How often felt everything effort in past 30 days	45,844	45,457	387	0.84	0.67
DSTNGD30	How often felt down/worthless/no good in past 30 days	45,844	45,667	177	0.39	0.36

K6 = Kessler-6, a psychological distress scale.

Note: The unweighted item nonresponse rates are defined as the total number of missing cases divided by the total number of cases. The weighted rates are defined similarly to the unweighted item nonresponse rates but with the survey weights applied to the percentages.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010 (Revised March 2012).

Table J.2 Item Nonresponse Rates for Edited Worst Month in Past Year K6 Variables, 2010 NSDUH

K6 Variable	Description	Total	Domain Status (DSTWORST)			"Not Worst Month" Cases		Unweighted Item Nonresponse Rate (Percent)	Weighted Item Nonresponse Rate (Percent)
			Past Month Was Worst Month	Past Month Was Not Worst Month	Missing	Not Missing	Missing		
DSTNRV12	How often felt nervous in worst month, past 12 months	45,844	29,936	15,662	246	15,621	41	0.26	0.30
DSTHOP12	How often felt hopeless in worst month, past 12 months	45,844	29,936	15,662	246	15,628	34	0.22	0.23
DSTRST12	How often felt restless in worst month, past 12 months	45,844	29,936	15,662	246	15,613	49	0.31	0.30
DSTCHR12	How often couldn't be cheered up in worst month, past 12 months	45,844	29,936	15,662	246	15,631	31	0.20	0.13
DSTEFF12	How often felt everything an effort in worst month, past 12 months	45,844	29,936	15,662	246	15,579	83	0.53	0.33
DSTNGD12	How often felt no good in worst month, past 12 months	45,844	29,936	15,662	246	15,627	35	0.22	0.18

K6 = Kessler-6, a psychological distress scale.

Note: The unweighted item nonresponse rates are defined as the total number of missing among the "Not Worst Month" cases divided by the total number of "Not Worst Month" cases (i.e., the denominator = 15,662). The weighted rates are defined similarly to the unweighted item nonresponse rates but with the survey weights applied to the percentages.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010 (Revised March 2012).

Table J.3 Item Nonresponse Rates for Edited WHODAS Variables, 2010 NSDUH

WHODAS Variables	Description	Total	Domain Status				"In Domain" Cases		Unweighted Item Nonresponse Rate (Percent)	Weighted Item Nonresponse Rate (Percent)
			Domain Variable	In Domain	Not in Domain	Domain Status Missing	Not Missing	Missing		
IMPREMEM	Difficulty remembering one month in past 12 months	45,844	DISTRESS	37,520	8,112	212	37,301	219	0.58	0.56
IMPCONCN	Difficulty concentrating one month in past 12 months	45,844	DISTRESS	37,520	8,112	212	37,317	203	0.54	0.58
IMPGOUT	Difficulty going out one month in past 12 months	45,844	DISTRESS	37,520	8,112	212	37,349	171	0.46	0.38
IMPPEOP	Difficulty dealing with strangers one month in past 12 months	45,844	DISTRESS	37,520	8,112	212	37,328	192	0.51	0.46
IMPSOC	Difficulty participating in social activities one month in past 12 months	45,844	DISTRESS	37,520	8,112	212	37,326	194	0.52	0.50
IMPHHLD	Difficulty taking care of household responsibilities one month in past 12 months	45,844	DISTRESS	37,520	8,112	212	37,344	176	0.47	0.48
IMPRESP	Difficulty taking care of work responsibilities one month in past 12 months	45,844	DISTRESS	37,520	8,112	212	37,318	202	0.54	0.66
IMPWORK	Difficulty doing daily work one month in past 12 months	45,844	DISTRESS, IMPRESP	35,866	9,564	414	35,839	27	0.08	0.09
IMPGOUTM	Emotional problems keep you from leaving house	45,844	DISTRESS, IMPGOUT	699	44,762	383	695	4	0.57	0.16
IMPEOPM	Emotional problems keep you from dealing with strangers	45,844	DISTRESS, IMPEOP	1,123	44,317	404	1,115	8	0.71	0.50
IMPSOCM	Emotional problems keep you from participating in social activities	45,844	DISTRESS, IMPSOC	1,382	44,056	406	1,374	8	0.58	0.56
IMPHHLDM	Emotional problems keep you from taking care of household responsibilities	45,844	DISTRESS, IMPHHLD	474	44,982	388	471	3	0.63	0.03
IMPRESPM	Emotional problems keep you from taking care of work responsibilities	45,844	DISTRESS, IMPRESP	1,452	43,978	414	1,439	13	0.90	0.33

WHODAS = World Health Organization Disability Assessment Schedule.

Note: The unweighted item nonresponse rates are defined as the total number of missing among the "In Domain" cases divided by the total number of "In Domain" cases. Refer to Section 9.2 for further details regarding the domains (i.e., the conditions under which the questions related to these items were asked). The weighted rates are defined similarly to the unweighted item nonresponse rates but with the survey weights applied to the percentages.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010 (Revised March 2012).

Table J.4 Model Summaries for WSHD, Mental Health Variables, Imputation Set 1

Variable	Variable Description	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
MHSUITHK	Seriously thought about killing self in past year	Nothing additional	AMDEYR NGD30_M
NRV30_M	How often felt nervous past 30 days		HOP30_M RST30_M
CHR30_M	How often felt sad nothing could cheer you up past 30 days		NGD30_M HOP30_M
NGD30_M	How often felt down/worthless/no good past 30 days		CHR30_M HOP30_M
HOP30_M	How often felt hopeless past 30 days		NGD30_M CHR30_M NRV30_M
RST30_M	How often felt restless/fidgety past 30 days		NRV30_M HOP30_M CHR30_M
AMDEYR	Past year major depressive episode (MDE)		NGD30_M MHSUITHK CHR30_M
EFF30_M	How often felt everything effort past 30 days		CHR30_M NGD30_M RST30_M

WSHD = weighted sequential hot deck.

Note: The starting list of predictor variables always includes age, gender, race/Hispanicity recode (7 levels), poverty level (3 levels), education level (4 levels), and employment status (4 levels). It also always includes all other variables in the same imputation set. The domain includes all respondents aged 18 or older.

Table J.5 Model Summaries for WSHD, Mental Health Variables, Imputation Set 2

Variable	Variable Description	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
DSTWORST	Was there a worse month in the past year than the last month?	Base set; imputation-revised variables from Imputation Set 1	NGD30_M AMDEYR RST30_M

WSHD = weighted sequential hot deck.

Note: The starting list of predictor variables always includes age, gender, race/Hispanicity recode (7 levels), poverty level (3 levels), education level (4 levels), and employment status (4 levels). It also always includes all other variables in the same imputation set. The domain includes all respondents aged 18 and older.

Table J.6 Model Summaries for WSHD, Mental Health Variables, Imputation Set 3

Variable	Variable Description	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
NRV12_M	How often felt nervous worst month	Base set; imputation-revised variables from Imputation Set 1	RST12_M HOP12_M NRV30_M
CHR12_M	How often felt sad nothing could cheer you up worst month		HOP12_M NGD12_M CHR30_M
NGD12_M	How often felt down/worthless/no good worst month		HOP12_M NGD30_M CHR12_M
HOP12_M	How often felt hopeless worst month		NGD12_M CHR12_M HOP30_M
RST12_M	How often felt restless/fidgety worst month		HOP12_M RST30_M NRV12_M CHR12_M
EFF12_M	How often felt everything effort worst month		CHR12_M EFF30_M NGD12_M

WSHD = weighted sequential hot deck.

Note: The starting list of predictor variables always includes age, gender, race/Hispanicity recode (7 levels), poverty level (3 levels), education level (4 levels), and employment status (4 levels). It also always includes all other variables in the same imputation set. The domain includes all respondents whose imputation-revised value of DSTWORST is 1 (yes, there was a worse month in the past year than the last month).

Table J.7 Model Summaries for WSHD, Mental Health Variables, Imputation Set 4

Variable	Variable Description	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
GOUT_M	Difficulty going out one month in past 12 months	Base set; imputation-revised variables from Imputation Sets 1 and 3; imputation-revised versions of K6SCMAX and WSPDSC2	SOC_M K6SCMAX RESP_M
HHLD_M	Difficulty household responsibilities one month in past 12 months		RESP_M SOC_M
PEOP_M	Difficulty dealing with strangers one month in past 12 months		SOC_M GOUT_M
SOC_M	Difficulty participating in social activities one month in past 12 months		GOUT_M PEOP_M HHLD_M
RESP_M	Difficulty with work/school responsibilities one month in past 12 months		HHLD_M SOC_M CONCN_M GOUT_M
CONCN_M	Difficulty concentrating one month in past 12 months		REMEM_M RESP_M
REMEM_M	Difficulty remembering one month in past 12 months		CONCN_M HHLD_M

WSHD = weighted sequential hot deck.

Note: The starting list of predictor variables always includes age, gender, race/Hispanicity recode (7 levels), poverty level (3 levels), education level (4 levels), and employment status (4 levels). It also always includes all other variables in the same imputation set. The domain includes all respondents whose imputation-revised value of K6SCMAX is greater than zero.

Table J.8 Model Summaries for WSHD, Mental Health Variables, Imputation Sets 5-9

Imputation Set	Variable	Variable Description	Domain	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
5	GOUTM_M	Did emotional problems keep you from going out?	Imputation-revised GOUT_M = 5	Base set; imputation-revised variables from Imputation Sets 1, 3, and 4; imputation-revised versions of K6SCMAX and WSPDSC2	K6SCMAX HHLD_M SOC_M
6	HHLDM_M	Did emotional problems keep you from household responsibilities?	Imputation-revised HHLD_M = 5		K6SCMAX AMDEYR
7	PEOPM_M	Did emotional problems keep you from dealing with strangers?	Imputation-revised PEOP_M = 5		K6SCMAX AMDEYR GOUT_M SOC_M HHLD_M
8	RESPM_M	Did emotional problems keep you from work/school responsibilities?	Imputation-revised RESP_M = 5		AMDEYR GOUT_M K6SCMAX REMEM_M
9	SOCM_M	Did emotional problems keep you from social activities?	Imputation-revised SOC_M = 5		K6SCMAX AMDEYR GOUT_M

WSHD = weighted sequential hot deck.

Note: The starting list of predictor variables always includes age, gender, race/Hispanicity recode (7 levels), poverty level (3 levels), education level (4 levels), and employment status (4 levels). It also always includes all other variables in the same imputation set.

Table J.9 Model Summaries for WSHD, Mental Health Variables, Imputation Set 10

Variable	Variable Description	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
WORK_M	Difficulty with daily work one month in past 12 months	Base set; imputation-revised variables from Imputation Sets 1 and 3-9; imputation-revised versions of K6SCMAX and WSPDSC2	RESP_M HHLD_M

WSHD = weighted sequential hot deck.

Note: The starting list of predictor variables always includes age, gender, race/Hispanicity recode (7 levels), poverty level (3 levels), education level (4 levels), and employment status (4 levels). It also always includes all other variables in the same imputation set. The domain includes all respondents whose imputation-revised value of RESP_M is 1, 2, 3, or 4.

Table J.10 SMI, AMI, and SMI Predictor Variable Estimates,[1] by Imputation Cell Categories for Current Versus WSHD Imputation Methods, 2010 NSDUH

	SMI		AMI		WSPDSC2_M		WHODASC3_M		MHSUITHK		AMDEYR	
Imputation Class	**CM**	**WSHD**	**CM**	**WSHD**	**CM**	**WSHD**	**CM**	**WSHD**	**CM**	**WSHD**	**CM**	**WSHD**
Total	4.1	4.1	18.1	18.2	1.4	1.5	0.9	0.9	3.8	3.8	6.8	6.9
Age												
18-25	3.9	3.9	18.1	18.2	2.3	2.4	1.2	1.2	6.6	6.7	8.2	8.3
26-34	5.1	5.1	21.6	21.8	1.9	1.9	1.0	1.0	4.1	4.1	7.2	7.2
35-49	5.3	5.3	20.6	20.8	1.5	1.5	0.9	0.9	4.0	4.1	7.6	7.7
50+	3.0	3.0	15.1	15.2	0.9	0.9	0.7	0.7	2.6	2.6	5.6	5.7
Gender												
Male	3.0	3.0	14.8	15.0	1.2	1.2	0.7	0.7	3.8	3.8	5.0	5.1
Female	5.1	5.1	21.1	21.2	1.7	1.7	1.0	1.0	3.9	3.9	8.4	8.5
Hispanic Origin and Race												
Not Hispanic or Latino												
White	4.3	4.3	19.1	19.2	1.4	1.5	0.9	0.9	4.0	4.1	7.3	7.4
Black or African American	3.9	3.9	17.0	17.2	1.5	1.5	0.8	0.8	4.1	4.1	5.8	5.9
Other or Multiple Races	3.6	3.6	15.9	16.0	1.3	1.3	0.8	0.8	4.0	4.0	5.5	5.5
Hispanic or Latino	3.2	3.3	15.2	15.4	1.5	1.5	0.8	0.8	2.4	2.4	5.6	5.7

AMI = any mental illness, CM = current method, WSHD = weighted sequential hot deck, K6 = Kessler-6, a 6-item psychological distress scale, SMI = serious mental illness, WHODAS = World Health Organization Disability Assessment Schedule.

Note: Estimates have been rounded to the nearest tenth to ensure respondent confidentiality.

[1] SMI, AMI, MHSUITHK, and AMDEYR estimates are prevalence estimates expressed as percentages; WSPDSC2_M and WHODASC3_M estimates are means of K6 and WHODAS total scores, respectively.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010 (Revised March 2012).

Table J.11 SMI, AMI, and SMI Predictor Variable Standard Errors, by Imputation Cell Categories for Current Versus WSHD Imputation Methods, 2010 NSDUH

	SMI		AMI		WSPDSC2_M		WHODASC3_M		MHSUITHK		AMDEYR	
Imputation Class	**CM**	**WSHD**	**CM**	**WSHD**	**CM**	**WSHD**	**CM**	**WSHD**	**CM**	**WSHD**	**CM**	**WSHD**
Total	0.0016	0.0016	0.0030	0.0030	0.0252	0.0254	0.0147	0.0147	0.0014	0.0014	0.0019	0.0019
Age												
18-25	0.0017	0.0017	0.0035	0.0035	0.0390	0.0390	0.0194	0.0193	0.0022	0.0022	0.0025	0.0025
26-34	0.0037	0.0037	0.0069	0.0069	0.0641	0.0641	0.0318	0.0318	0.0035	0.0035	0.0042	0.0042
35-49	0.0029	0.0029	0.0052	0.0052	0.0444	0.0449	0.0250	0.0249	0.0026	0.0026	0.0035	0.0035
50+	0.0027	0.0027	0.0055	0.0055	0.0405	0.0407	0.0264	0.0265	0.0022	0.0022	0.0034	0.0034
Gender												
Male	0.0020	0.0020	0.0042	0.0042	0.0313	0.0315	0.0193	0.0192	0.0021	0.0021	0.0025	0.0025
Female	0.0023	0.0023	0.0043	0.0043	0.0371	0.0372	0.0206	0.0206	0.0018	0.0018	0.0028	0.0028
Hispanic Origin and Race												
Not Hispanic or Latino												
White	0.0019	0.0019	0.0036	0.0036	0.0287	0.0288	0.0175	0.0176	0.0017	0.0017	0.0023	0.0023
Black or African American	0.0036	0.0036	0.0084	0.0084	0.0695	0.0701	0.0366	0.0368	0.0038	0.0038	0.0047	0.0047
Other or Multiple Races	0.0066	0.0066	0.0119	0.0119	0.0940	0.0978	0.0516	0.0518	0.0055	0.0055	0.0072	0.0072
Hispanic or Latino	0.0039	0.0039	0.0077	0.0078	0.0748	0.0753	0.0420	0.0422	0.0024	0.0024	0.0051	0.0051

AMI = any mental illness, CM = current method, WSHD = weighted sequential hot deck, K6 = Kessler-6, a 6-item psychological distress scale, SMI = serious mental illness, WHODAS = World Health Organization Disability Assessment Schedule.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2010 (Revised March 2012).

This page intentionally left blank

Appendix K: Substance Dependence and Abuse Item Nonresponse Rates and Patterns

This page intentionally left blank

Table K.1 Item Nonresponse Rates for Nicotine Dependence Syndrome Scale Score (NDSSANSP) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status		Past Month Use Cases			Unweighted Item Nonresponse Rate[2]	Weighted Item Nonresponse Rate[3]
			Never Used/Not Used in Past Month	Used in Past Month	Not Missing	Logically Assigned	Missing[1]		
Need to smoke to feel less irritable	CIGIRTBL	70,109	54,132	15,977	15,908	2	67	0.42	0.58
Start to crave cigarettes when don't smoke for few hours	CIGCRAVE	70,109	54,132	15,977	15,931	2	44	0.28	0.34
Craving of cigarette like strong force can't control	CIGCRAGP	70,109	54,132	15,977	15,925	2	50	0.31	0.39
Feel a sense of control over your smoking	CIGINCTL	70,109	54,132	15,977	15,914	2	61	0.38	0.41
Tend to avoid places that don't allow smoking	CIGAVOID	70,109	54,132	15,977	15,890	2	85	0.53	0.47
Rather not travel by airplane because no smoking	CIGPLANE	70,109	54,132	15,977	15,896	2	79	0.49	0.56
Sometimes worry that you will run out of cigarettes	CIGRNOUT	70,109	54,132	15,977	15,931	2	44	0.28	0.29
Smoke cigarettes fairly regularly throughout the day	CIGREGDY	70,109	54,132	15,977	15,928	2	47	0.29	0.33
Smoke same amount on weekends as on weekdays	CIGREGWK	70,109	54,132	15,977	15,900	2	75	0.47	0.41
Smoke same number of cigarettes from day to day	CIGREGNM	70,109	54,132	15,977	15,906	2	69	0.43	0.32
Number of cigarettes smoked per day often changes	CIGNMCHG	70,109	54,132	15,977	15,886	2	89	0.56	0.51
Have many cigarettes in hour, then no cigarettes until hours later	CIGSVLHR	70,109	54,132	15,977	15,877	2	98	0.61	0.54
Number of cigarettes smoked per day influenced by other things	CIGINFLU	70,109	54,132	15,977	15,902	2	73	0.46	0.46
Smoking not affected by other things	CIGNOINF	70,109	54,132	15,977	15,868	2	107	0.67	0.64
Amount of smoking has increased since started smoking	CIGINCRS	70,109	54,132	15,977	15,906	2	69	0.43	0.53
Need to smoke a lot more to be satisfied	CIGSATIS	70,109	54,132	15,977	15,912	2	63	0.39	0.52
Smoke much more now before feel anything	CIGLOTMR	70,109	54,132	15,977	15,863	2	112	0.70	0.84
How soon after waking do you have your first cigarette	CIGWAKE	70,109	54,132	15,977	15,409	2	566	3.54	2.56

Note: The NDSS score, NDSSANSP, is calculated as the average score of these 17 variables pertaining to five aspects of dependence.

[1] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."

[2] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past month users.

[3] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past month users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.2 Item Nonresponse Rates for Alcohol Dependence in Past Year (DEPNDALC) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status		Past Year Use Cases			Unweighted Item Nonresponse Rate[2]	Weighted Item Nonresponse Rate[3]
			Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[1]		
Spent month/more getting/drinking alcohol in past 12 months	ALCLOTTM	70,109	36,406	33,703	33,452	0	251	0.74	0.32
Month/more spent getting over alcohol effects in past 12 months	ALCGTOVR*	70,109	36,406	33,703	33,454	0	249	0.74	0.31
Able to keep limits or drank more in past 12 months	ALCKPLMT*	70,109	36,406	33,703	33,398	0	305	0.90	0.50
Needed more alcohol to get same effect in past 12 months	ALCNDMOR	70,109	36,406	33,703	33,423	0	280	0.83	0.40
Drinking same amount of alcohol has less effect in past 12 months	ALCLSEFX*	70,109	36,406	33,703	33,330	0	373	1.11	0.81
Able to cut/stop drinking every time in past 12 months	ALCCUTEV*	70,109	36,406	33,703	33,317	0	386	1.15	0.75
Continued to drink alcohol despite emotional problems	ALCEMCTD*	70,109	36,406	33,703	33,424	0	279	0.83	0.40
Continued to drink alcohol despite physical problems	ALCPHCTD*	70,109	36,406	33,703	33,441	0	262	0.78	0.37
Less activities because of alcohol use in past 12 months	ALCLSACT	70,109	36,406	33,703	33,432	0	271	0.80	0.38
Had 2+ alcohol withdrawal symptoms at same time in past 12 months	ALCWDSMT*	70,109	36,406	33,703	33,277	0	426	1.26	0.96

Note: Alcohol dependence in past year (DEPNDALC) is determined by seven criteria that were computed from the 10 variables shown in this table.

*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."

[2] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.

[3] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.3 Item Nonresponse Rates for Prescription Pain Reliever Dependence in Past Year (DEPNDANL) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status		Past Year Use Cases			Unweighted Item Nonresponse Rate[2]	Weighted Item Nonresponse Rate[3]
			Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[1]		
Spent month/more getting/using pain relievers in past 12 months	ANLLOTTM	70,109	65,425	4,684	4,489	0	195	4.16	3.14
Month/more spent getting over pain reliever effects in past 12 months	ANLGTOVR*	70,109	65,425	4,684	4,487	0	197	4.21	3.15
Able to keep limits or use pain relievers more in past 12 months	ANLKPLMT*	70,109	65,425	4,684	4,468	0	216	4.61	3.45
Needed more pain relievers to get same effect in past 12 months	ANLNDMOR	70,109	65,425	4,684	4,474	0	210	4.48	3.25
Using same amount of pain relievers has less effect in past 12 months	ANLLSEFX*	70,109	65,425	4,684	4,464	0	220	4.70	3.63
Able to cut/stop using pain relievers every time in past 12 months	ANLCUTEV*	70,109	65,425	4,684	4,434	0	250	5.34	4.03
Continued to use pain relievers despite emotional problems	ANLEMCTD*	70,109	65,425	4,684	4,462	0	222	4.74	3.37
Continued to use pain relievers despite physical problems	ANLPHCTD*	70,109	65,425	4,684	4,465	0	219	4.68	3.50
Less activities because of pain reliever use in past 12 months	ANLLSACT	70,109	65,425	4,684	4,461	0	223	4.76	3.77
Had 3+ pain reliever withdrawal symptoms at same time in past 12 months	ANLWDSMT*	70,109	65,425	4,684	4,420	0	264	5.64	4.45

Note: Prescription pain reliever dependence in past year (DEPNDANL) is determined by seven criteria that were computed from the 10 variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."
[2] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.
[3] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.4 Item Nonresponse Rates for Cocaine Dependence in Past Year (DEPNDCOC) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status[1]			Past Year Use Cases			Unweighted Item Nonresponse Rate[3]	Weighted Item Nonresponse Rate[4]
			Unknown	Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[2]		
Spent month/more getting/using cocaine in past 12 months	COCLOTTM	70,109	4	68,524	1,581	1,541	1	39	2.47	1.66
Month/more spent getting over cocaine effects in past 12 months	COCGTOVR*	70,109	4	68,524	1,581	1,540	0	41	2.59	1.89
Able to keep limits or use cocaine more in past 12 months	COCKPLMT*	70,109	4	68,524	1,581	1,535	1	45	2.85	1.92
Needed more cocaine to get same effect in past 12 months	COCNDMOR	70,109	4	68,524	1,581	1,535	1	45	2.85	1.92
Using same amount of cocaine has less effect in past 12 months	COCLSEFX*	70,109	4	68,524	1,581	1,524	0	57	3.61	2.45
Able to cut/stop using cocaine every time in past 12 months	COCCUTEV*	70,109	4	68,524	1,581	1,530	1	50	3.16	2.00
Continued to use cocaine despite emotional problems	COCEMCTD*	70,109	4	68,524	1,581	1,538	1	42	2.66	1.67
Continued to use cocaine despite physical problems	COCPHCTD*	70,109	4	68,524	1,581	1,538	0	43	2.72	1.92
Less activities because of cocaine use in past 12 months	COCLSACT	70,109	4	68,524	1,581	1,535	1	45	2.85	1.93
When cut down on cocaine, felt blue in past 12 months	COCFLBLU*	70,109	4	68,524	1,581	1,531	1	49	3.10	2.03
Had 2+ cocaine withdrawal symptoms at same time in past 12 months	COCWDSMT*	70,109	4	68,524	1,581	1,531	1	49	3.10	2.03

Note: Cocaine dependence in past year (DEPNDCOC) is determined by seven criteria that were computed from the 11 variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] The imputation-revised domain for cocaine dependence includes four cases whose value was unknown. These cases have a recency value of not used in past year (IRCOCRC=3) and time since last used needle to inject cocaine (CONDLREC) values of "Refused (97)" or "Blank (98)."

[2] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."

[3] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.

[4] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.5 Item Nonresponse Rates for Hallucinogen Dependence in Past Year (DEPNDHAL) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status		Past Year Use Cases			Unweighted Item Nonresponse Rate[2]	Weighted Item Nonresponse Rate[3]
			Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[1]		
Spent month/more getting/using hallucinogens in past 12 months	HALLOTTM	70,109	67,806	2,303	2,227	1	75	3.26	3.18
Month/more spent getting over hallucinogen effects in past 12 months	HALGTOVR*	70,109	67,806	2,303	2,226	0	77	3.34	3.44
Able to keep limits or use hallucinogens more in past 12 months	HALKPLMT*	70,109	67,806	2,303	2,219	1	83	3.61	3.41
Needed more hallucinogens to get same effect in past 12 months	HALNDMOR	70,109	67,806	2,303	2,222	1	80	3.48	3.27
Using same amount of hallucinogens has less effect in past 12 months	HALLSEFX*	70,109	67,806	2,303	2,216	0	87	3.78	3.73
Able to cut/stop using hallucinogens every time in past 12 months	HALCUTEV*	70,109	67,806	2,303	2,208	1	94	4.08	3.64
Continued to use hallucinogens despite emotional problems	HALEMCTD*	70,109	67,806	2,303	2,216	1	86	3.74	3.38
Continued to use hallucinogens despite physical problems	HALPHCTD*	70,109	67,806	2,303	2,220	0	83	3.60	3.52
Less activities because of hallucinogen use in past 12 months	HALLSACT	70,109	67,806	2,303	2,215	1	87	3.78	3.41

Note: Hallucinogen dependence in past year (DEPNDHAL) is determined by six criteria that were computed from the nine variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."
[2] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.
[3] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.6 Item Nonresponse Rates for Heroin Dependence in Past Year (DEPNDHER) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status[1]			Past Year Use Cases			Unweighted Item Nonresponse Rate[3]	Weighted Item Nonresponse Rate[4]
			Unknown	Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[2]		
Spent month/more getting/using heroin in past 12 months	HERLOTTM	70,109	1	69,841	267	263	0	4	1.50	0.58
Month/more spent getting over heroin effects in past 12 months	HERGTOVR*	70,109	1	69,841	267	262	0	5	1.87	0.68
Able to keep limits or use heroin more in past 12 months	HERKPLMT*	70,109	1	69,841	267	262	0	5	1.87	0.68
Needed more heroin to get same effect in past 12 months	HERNDMOR	70,109	1	69,841	267	262	0	5	1.87	0.68
Using same amount of heroin has less effect in past 12 months	HERLSEFX*	70,109	1	69,841	267	261	0	6	2.25	0.72
Able to cut/stop using heroin every time in past 12 months	HERCUTEV*	70,109	1	69,841	267	262	0	5	1.87	0.68
Continued to use heroin despite emotional problems	HEREMCTD*	70,109	1	69,841	267	261	0	6	2.25	0.89
Continued to use heroin despite physical problems	HERPHCTD*	70,109	1	69,841	267	262	0	5	1.87	1.68
Less activities because of heroin use in past 12 months	HERLSACT	70,109	1	69,841	267	263	0	4	1.50	0.67
Had 2+ heroin withdrawal symptoms at same time in past 12 months	HERWESMT*	70,109	1	69,841	267	259	0	8	3.00	1.94

Note: Heroin dependence in past year (DEPNDHER) is determined by seven criteria that were computed from the 10 variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] The imputation-revised domain for heroin dependence includes one case whose value was unknown. This case had a past year use value of not used in past year (HERYR=0), time since last smoked heroin value (HRSMKREC) of "Blank (98)," time since last sniffed heroin value (HRSNFREC) of "Blank (98)," and time since last used needle to inject heroin (HRNDLREC) value of "Blank (98)."

[2] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."

[3] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.

[4] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.7 Item Nonresponse Rates for Inhalant Dependence in Past Year (DEPNDINH) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status		Past Year Use Cases			Unweighted Item Nonresponse Rate[2]	Weighted Item Nonresponse Rate[3]
			Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[1]		
Spent month/more getting/using inhalants in past 12 months	INHLOTTM	70,109	68,984	1,125	1,036	0	89	7.91	5.34
Month/more spent getting over inhalant effects in past 12 months	INHGTOVR*	70,109	68,984	1,125	1,036	0	89	7.91	5.25
Able to keep limits or use inhalants more in past 12 months	INHKPLMT*	70,109	68,984	1,125	1,022	0	103	9.16	6.33
Needed more inhalants to get same effect in past 12 months	INHNDMOR	70,109	68,984	1,125	1,026	0	99	8.80	5.82
Using same amount of inhalants has less effect in past 12 months	INHLSEFX*	70,109	68,984	1,125	1,013	0	112	9.96	7.02
Able to cut/stop using inhalants every time in past 12 months	INHCUTEV*	70,109	68,984	1,125	1,000	0	125	11.11	8.05
Continued to use inhalants despite emotional problems	INHEMCTD*	70,109	68,984	1,125	1,017	0	108	9.60	6.71
Continued to use inhalants despite physical problems	INHPHCTD*	70,109	68,984	1,125	1,020	0	105	9.33	6.59
Less activities because of inhalant use in past 12 months	INHLSACT	70,109	68,984	1,125	1,019	0	106	9.42	6.62

Note: Inhalant dependence in past year (DEPNDINH) is determined by six criteria that were computed from the nine variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."
[2] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.
[3] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.8 Item Nonresponse Rates for Marijuana Dependence in Past Year (DEPNDMRJ) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status		Past Year Use Cases			Unweighted Item Nonresponse Rate[2]	Weighted Item Nonresponse Rate[3]
			Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[1]		
Spent month/more getting/using marijuana in past 12 months	MRJLOTTM	70,109	60,556	9,553	9,405	0	148	1.55	1.14
Month/more spent getting over marijuana effects in past 12 months	MRJGTOVR*	70,109	60,556	9,553	9,407	0	146	1.53	1.11
Able to keep limits or use marijuana more in past 12 months	MRJKPLMT*	70,109	60,556	9,553	9,393	0	160	1.67	1.19
Needed more marijuana to get same effect in past 12 months	MRJNDMOR	70,109	60,556	9,553	9,402	0	151	1.58	1.23
Using same amount of marijuana has less effect in past 12 months	MRJLSEFX*	70,109	60,556	9,553	9,386	0	167	1.75	1.39
Able to cut/stop using marijuana every time in past 12 months	MRJCUTEV*	70,109	60,556	9,553	9,391	0	162	1.70	1.28
Continued to use marijuana despite emotional problems	MRJEMCTD*	70,109	60,556	9,553	9,391	0	162	1.70	1.32
Continued to use marijuana despite physical problems	MRJPHCTD*	70,109	60,556	9,553	9,397	0	156	1.63	1.18
Less activities because of marijuana use in past 12 months	MRJLSACT	70,109	60,556	9,553	9,407	0	146	1.53	1.11

Note: Marijuana dependence in past year (DEPNDMRJ) is determined by six criteria that were computed from the nine variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."
[2] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.
[3] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.9 Item Nonresponse Rates for Sedative Dependence in Past Year (DEPNDSED) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status		Past Year Use Cases			Unweighted Item Nonresponse Rate[2]	Weighted Item Nonresponse Rate[3]
			Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[1]		
Spent month/more getting/using sedatives in past 12 months	SEDLOTTM	70,109	69,886	223	207	0	16	7.17	6.47
Month/more spent getting over sedative effects in past 12 months	SEDGTOVR*	70,109	69,886	223	207	0	16	7.17	6.47
Able to keep limits or use sedatives more in past 12 months	SEDKPLMT*	70,109	69,886	223	203	0	20	8.97	7.80
Needed more sedatives to get same effect in past 12 months	SEDNDMOR	70,109	69,886	223	207	0	16	7.17	5.99
Using same amount of sedatives has less effect in past 12 months	SEDLSEFX*	70,109	69,886	223	204	0	19	8.52	7.64
Able to cut/stop using sedatives every time in past 12 months	SEDCUTEV*	70,109	69,886	223	203	0	20	8.97	7.81
Continued to use sedatives despite emotional problems	SEDEMCTD*	70,109	69,886	223	204	0	19	8.52	6.87
Continued to use sedatives despite physical problems	SEDPHCTD*	70,109	69,886	223	205	0	18	8.07	6.39
Less activities because of sedative use in past 12 months	SEDLSACT	70,109	69,886	223	206	0	17	7.62	6.02
Had 2+ sedative withdrawal symptoms at same time in past 12 months	SEDWDSMT*	70,109	69,886	223	201	0	22	9.87	7.88

Note: Sedative dependence in past year (DEPNDSED) is determined by seven criteria that were computed from the 10 variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."

[2] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.
[3] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.10 Item Nonresponse Rates for Stimulant Dependence in Past Year (DEPNDSTM) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status[1]			Past Year Use Cases			Unweighted Item Nonresponse Rate[3]	Weighted Item Nonresponse Rate[4]
			Unknown	Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[2]		
Spent month/more getting/using stimulants in past 12 months	STMLOTTM	70,109	10	68,888	1,211	1,174	0	37	3.06	2.88
Month/more spent getting over stimulant effects in past 12 months	STMGTOVR*	70,109	10	68,888	1,211	1,175	0	36	2.97	2.60
Able to keep limits or use stimulants more in past 12 months	STMKPLMT*	70,109	10	68,888	1,211	1,172	0	39	3.22	2.69
Needed more stimulants to get same effect in past 12 months	STMNDMOR	70,109	10	68,888	1,211	1,172	0	39	3.22	2.59
Using same amount of stimulants has less effect in past 12 months	STMLSEFX*	70,109	10	68,888	1,211	1,171	0	40	3.30	2.65
Able to cut/stop using stimulants every time in past 12 months	STMCUTEV*	70,109	10	68,888	1,211	1,171	0	40	3.30	2.64
Continued to use stimulants despite emotional problems	STMEMCTD*	70,109	10	68,888	1,211	1,172	0	39	3.22	2.74
Continued to use stimulants despite physical problems	STMPHCTD*	70,109	10	68,888	1,211	1,172	0	39	3.22	2.74
Less activities because of stimulant use in past 12 months	STMLSACT	70,109	10	68,888	1,211	1,173	0	38	3.14	2.65
When cut down on stimulants, felt blue in past 12 months	STMFLBLU*	70,109	10	68,888	1,211	1,166	0	45	3.72	3.18
Had 2+ stimulant withdrawal symptoms at same time in past 12 months	STMWDSMT*	70,109	10	68,888	1,211	1,166	0	45	3.72	3.18

Note: Stimulant dependence in past year (DEPNDSTM) is determined by seven criteria that were computed from the 11 variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] The imputation-revised domain for stimulant dependence includes 10 cases whose value was unknown. These cases have a positive value for ever used (STMFLAG=1), a past year use value of not used in past year (STMYR=0), and a time since last used needle to inject methamphetamine (MTNDLREC) value of used at some point in lifetime – logically assigned (9), "Refused (97)," or "Blank (98)," or a time since last used needle to inject other stimulant (OSTNLREC) value of "Refused (97)" or "Blank (98)." There were also four cases that were not in the domain and required hard coding to "never used/not used in past year."

[2] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."

[3] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.

[4] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.11 Item Nonresponse Rates for Tranquilizer Dependence in Past Year (DEPNDTRN) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status		Past Year Use Cases			Unweighted Item Nonresponse Rate[2]	Weighted Item Nonresponse Rate[3]
			Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[1]		
Spent month/more getting/using tranquilizers in past 12 months	TRNLOTTM	70,109	68,194	1,915	1,866	1	48	2.51	1.98
Month/more spent getting over tranquilizer effects in past 12 months	TRNGTOVR*	70,109	68,194	1,915	1,863	0	52	2.72	2.26
Able to keep limits or use tranquilizers more in past 12 months	TRNKPLMT*	70,109	68,194	1,915	1,850	1	64	3.34	2.65
Needed more tranquilizers to get same effect in past 12 months	TRNNDMOR	70,109	68,194	1,915	1,856	1	58	3.03	2.21
Using same amount of tranquilizers has less effect in past 12 months	TRNLSEFX*	70,109	68,194	1,915	1,844	0	71	3.71	3.16
Able to cut/stop using tranquilizers every time in past 12 months	TRNUTEV*	70,109	68,194	1,915	1,840	1	74	3.87	3.04
Continued to use tranquilizers despite emotional problems	TRNEMCTD*	70,109	68,194	1,915	1,850	1	64	3.34	2.55
Continued to use tranquilizers despite physical problems	TRNPHCTD*	70,109	68,194	1,915	1,852	0	63	3.29	2.72
Less activities because of tranquilizer use in past 12 months	TRNLSACT	70,109	68,194	1,915	1,853	1	61	3.19	2.53

Note: Tranquilizer dependence in past year (DEPNDTRN) is determined by six criteria that were computed from the nine variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."

[2] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.
[3] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.12 Item Nonresponse Rates for Alcohol Abuse in Past Year (ABUSEALC) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status		Past Year Use Cases			Unweighted Item Nonresponse Rate[2]	Weighted Item Nonresponse Rate[3]
			Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[1]		
Alcohol causes serious problems at home/work/school in past 12 months	ALCSERPB	70,109	36,406	33,703	33,438	0	265	0.79	0.37
Drinking alcohol and doing dangerous activities in past 12 months	ALCPDANG	70,109	36,406	33,703	33,436	0	267	0.79	0.35
Drinking alcohol causes problems with law in past 12 months	ALCLAWTR	70,109	36,406	33,703	33,452	0	251	0.74	0.32
Drinking alcohol causes problems with family/friends in past 12 months	ALCFMFPB	70,109	36,406	33,703	33,450	0	253	0.75	0.34
Continued to drink alcohol despite problems with family/friends	ALCFMCTD*	70,109	36,406	33,703	33,447	0	256	0.76	0.34

Note: Alcohol abuse in past year (ABUSEALC) is determined by four criteria that were computed from the five variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."

[2] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.

[3] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.13 Item Nonresponse Rates for Prescription Pain Reliever Abuse in Past Year (ABUSEANL) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status		Past Year Use Cases			Unweighted Item Nonresponse Rate[2]	Weighted Item Nonresponse Rate[3]
			Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[1]		
Pain relievers cause serious problems at home/work/school in past 12 months	ANLSERPB	70,109	65,425	4,684	4,463	0	221	4.72	3.55
Using pain relievers and doing dangerous activities in past 12 months	ANLPDANG	70,109	65,425	4,684	4,463	0	221	4.72	3.45
Using pain relievers causes problems with law in past 12 months	ANLLAWTR	70,109	65,425	4,684	4,470	0	214	4.57	3.39
Using pain relievers causes problems with family/friends in past 12 months	ANLFMFPB	70,109	65,425	4,684	4,472	0	212	4.53	3.44
Continued to use pain relievers despite problems with family/friends	ANLFMCTD*	70,109	65,425	4,684	4,472	0	212	4.53	3.44

Note: Prescription pain reliever abuse in past year (ABUSEANL) is determined by four criteria that were computed from the five variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."
[2] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.
[3] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.14 Item Nonresponse Rates for Cocaine Abuse in Past Year (ABUSECOC) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status[1]			Past Year Use Cases			Unweighted Item Nonresponse Rate[3]	Weighted Item Nonresponse Rate[4]
			Unknown	Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[2]		
Cocaine causes serious problems at home/work/school in past 12 months	COCSERPB	70,109	4	68,524	1,581	1,535	1	45	2.85	1.84
Using cocaine and doing dangerous activities in past 12 months	COCPDANG	70,109	4	68,524	1,581	1,535	1	45	2.85	1.96
Using cocaine causes problems with law in past 12 months	COCLAWTR	70,109	4	68,524	1,581	1,536	1	44	2.78	1.78
Using cocaine causes problems with family/friends in past 12 months	COCFMFPB	70,109	4	68,524	1,581	1,536	1	44	2.78	1.78
Continued to use cocaine despite problems with family/friends	COCFMCTD*	70,109	4	68,524	1,581	1,536	1	44	2.78	1.78

Note: Cocaine abuse in past year (ABUSECOC) is determined by four criteria that were computed from the five variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] The imputation-revised domain for cocaine dependence includes four cases whose value was unknown. These cases have a recency value of not used in past year (IRCOCRC=3) and time since last used needle to inject cocaine (CONDLREC) values of "Refused (97)" or "Blank (98)."

[2] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."

[3] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.

[4] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.15 Item Nonresponse Rates for Hallucinogen Abuse in Past Year (ABUSEHAL) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status		Past Year Use Cases			Unweighted Item Nonresponse Rate[2]	Weighted Item Nonresponse Rate[3]
			Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[1]		
Hallucinogens cause serious problems at home/work/school in past 12 months	HALSERPB	70,109	67,806	2,303	2,217	1	85	3.69	3.36
Using hallucinogens and doing dangerous activities in past 12 months	HALPDANG	70,109	67,806	2,303	2,220	1	82	3.56	3.30
Using hallucinogens causes problems with law in past 12 months	HALLAWTR	70,109	67,806	2,303	2,221	1	81	3.52	3.27
Using hallucinogens causes problems with family/friends in past 12 months	HALFMFPB	70,109	67,806	2,303	2,221	1	81	3.52	3.24
Continued to use hallucinogens despite problems with family/friends	HALFMCTD*	70,109	67,806	2,303	2,221	1	81	3.52	3.24

Note: Hallucinogen abuse in past year (ABUSEHAL) is determined by four criteria that were computed from the five variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."
[2] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.
[3] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.16 Item Nonresponse Rates for Heroin Abuse in Past Year (ABUSEHER) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status[1]			Past Year Use Cases			Unweighted Item Nonresponse Rate[3]	Weighted Item Nonresponse Rate[4]
			Unknown	Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[2]		
Heroin causes serious problems at home/work/school in past 12 months	HERSERPB	70,109	1	69,841	267	262	0	5	1.87	0.88
Using heroin and doing dangerous activities in past 12 months	HERPDANG	70,109	1	69,841	267	261	0	6	2.25	1.15
Using heroin causes problems with law in past 12 months	HERLAWTR	70,109	1	69,841	267	261	0	6	2.25	1.15
Using heroin causes problems with family/friends in past 12 months	HERFMFPB	70,109	1	69,841	267	262	0	5	1.87	1.10
Continued to use heroin despite problems with family/friends	HERFMCTD*	70,109	1	69,841	267	262	0	5	1.87	1.10

Note: Heroin abuse in past year (ABUSEHER) is determined by four criteria that were computed from the five variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] The imputation-revised domain for heroin dependence includes one case whose value was unknown. This case had a past year use value of not used in past year (HERYR=0), time since last smoked heroin value (HRSMKREC) of "Blank (98)," time since last sniffed heroin value (HRSNFREC) of "Blank (98)," and time since last used needle to inject heroin (HRNDLREC) value of "Blank (98)."

[2] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."

[3] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.

[4] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.17 Item Nonresponse Rates for Inhalant Abuse in Past Year (ABUSEINH) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status		Past Year Use Cases			Unweighted Item Nonresponse Rate[2]	Weighted Item Nonresponse Rate[3]
			Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[1]		
Inhalants cause serious problems at home/work/school in past 12 months	INHSERPB	70,109	68,984	1,125	1,020	0	105	9.33	6.56
Using inhalants and doing dangerous activities in past 12 months	INHPDANG	70,109	68,984	1,125	1,022	0	103	9.16	6.05
Using inhalants causes problems with law in past 12 months	INHLAWTR	70,109	68,984	1,125	1,022	0	103	9.16	6.38
Using inhalants causes problems with family/friends in past 12 months	INHFMFPB	70,109	68,984	1,125	1,025	0	100	8.89	5.99
Continued to use inhalants despite problems with family/friends	INHFMCTD*	70,109	68,984	1,125	1,025	0	100	8.89	5.99

Note: Inhalant abuse in past year (ABUSEINH) is determined by four criteria that were computed from the five variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."
[2] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.
[3] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.18 Item Nonresponse Rates for Marijuana Abuse in Past Year (ABUSEMRJ) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status		Past Year Use Cases			Unweighted Item Nonresponse Rate[2]	Weighted Item Nonresponse Rate[3]
			Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[1]		
Marijuana causes serious problems at home/work/school in past 12 months	MRJSERPB	70,109	60,556	9,553	9,404	0	149	1.56	1.13
Using marijuana and doing dangerous activities in past 12 months	MRJPDANG	70,109	60,556	9,553	9,405	0	148	1.55	1.09
Using marijuana causes problems with law in past 12 months	MRJLAWTR	70,109	60,556	9,553	9,406	0	147	1.54	1.09
Using marijuana causes problems with family/friends in past 12 months	MRJFMFPB	70,109	60,556	9,553	9,409	0	144	1.51	1.08
Continued to use marijuana despite problems with family/friends	MRJFMCTD*	70,109	60,556	9,553	9,408	0	145	1.52	1.08

Note: Marijuana abuse in past year (ABUSEMRJ) is determined by four criteria that were computed from the five variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."
[2] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.
[3] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.19 Item Nonresponse Rates for Sedative Abuse in Past Year (ABUSESED) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status		Past Year Use Cases			Unweighted Item Nonresponse Rate[2]	Weighted Item Nonresponse Rate[3]
			Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[1]		
Sedatives cause serious problems at home/work/school in past 12 months	SEDSERPB	70,109	69,886	223	207	0	16	7.17	6.00
Using sedatives and doing dangerous activities in past 12 months	SEDPDANG	70,109	69,886	223	206	0	17	7.62	6.23
Using sedatives causes problems with law in past 12 months	SEDLAWTR	70,109	69,886	223	206	0	17	7.62	6.08
Using sedatives causes problems with family/friends in past 12 months	SEDFMFPB	70,109	69,886	223	206	0	17	7.62	6.04
Continued to use sedatives despite problems with family/friends	SEDFMCTD*	70,109	69,886	223	206	0	17	7.62	6.04

Note: Sedative abuse in past year (ABUSESED) is determined by four criteria that were computed from the five variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."
[2] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.
[3] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.20 Item Nonresponse Rates for Stimulant Abuse in Past Year (ABUSESTM) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status[1]			Past Year Use Cases			Unweighted Item Nonresponse Rate[3]	Weighted Item Nonresponse Rate[4]
			Unknown	Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[2]		
Stimulants cause serious problems at home/work/school in past 12 months	STMSERPB	70,109	10	68,888	1,211	1,173	0	38	3.14	2.65
Using stimulants and doing dangerous activities in past 12 months	STMPDANG	70,109	10	68,888	1,211	1,176	0	35	2.89	2.18
Using stimulants causes problems with law in past 12 months	STMLAWTR	70,109	10	68,888	1,211	1,175	0	36	2.97	2.21
Using stimulants causes problems with family/friends in past 12 months	STMFMFPB	70,109	10	68,888	1,211	1,175	0	36	2.97	2.21
Continued to use stimulants despite problems with family/friends	STMFMCTD*	70,109	10	68,888	1,211	1,175	0	36	2.97	2.21

Note: Stimulant abuse in past year (ABUSESTM) is determined by four criteria that were computed from the five variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] The imputation-revised domain for stimulant dependence includes 10 cases whose value was unknown. These cases have a positive value for ever used (STMFLAG=1), a past year use value of not used in past year (STMYR=0), and a time since last used needle to inject methamphetamine (MTNDLREC) value of used at some point in lifetime – logically assigned (9), "Refused (97)," or "Blank (98)," or a time since last used needle to inject other stimulant (OSTNLREC) value of "Refused (97)" or "Blank (98)." There were also four cases that were not in the domain and required hard coding to "never used/not used in past year."

[2] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."

[3] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.

[4] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.21 Item Nonresponse Rates for Tranquilizer Abuse in Past Year (ABUSETRN) Variables, 2011 NSDUH

Variable Description	Variable Name	Total	Imputation-Revised Domain Status		Past Year Use Cases			Unweighted Item Nonresponse Rate[2]	Weighted Item Nonresponse Rate[3]
			Never Used/Not Used in Past Year	Used in Past Year	Not Missing	Logically Assigned	Missing[1]		
Tranquilizers cause serious problems at home/work/school in past 12 months	TRNSERPB	70,109	68,194	1,915	1,853	1	61	3.19	2.67
Using tranquilizers and doing dangerous activities in past 12 months	TRNPDANG	70,109	68,194	1,915	1,856	1	58	3.03	2.35
Using tranquilizers causes problems with law in past 12 months	TRNLAWTR	70,109	68,194	1,915	1,857	1	57	2.98	2.54
Using tranquilizers causes problems with family/friends in past 12 months	TRNFMFPB	70,109	68,194	1,915	1,857	1	57	2.98	2.53
Continued to use tranquilizers despite problems with family/friends	TRNFMCTD*	70,109	68,194	1,915	1,857	1	57	2.98	2.53

Note: Tranquilizer abuse in past year (ABUSETRN) is determined by four criteria that were computed from the five variables shown in this table.
*Cases coded as "Legitimate Skip (99)" are included in the "Not Missing" column.

[1] Missing values are defined as those values coded "Don't Know (94)," "Refused (97)," and "Blank (98)."
[2] The unweighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users.
[3] The weighted item nonresponse rate is the number of missing values (excluding the number of logically assigned) divided by the number of past year users using the final analytic weight.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.22 Item Nonresponse Patterns for Nicotine Dependence Syndrome Scale Score (NDSSANSP) and Fagerstrom Test of Nicotine Dependence Scale Score (FTNDDNSP), 2011 NSDUH

Nicotine Dependence Syndrome Scale Score (NDSSANSP)	Frequency		
Total	70,109		
Never used/not used in past month	54,132		
Used in past month	15,977		
Item Nonresponse Patterns among Past Month Users	**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past month	15,977	100.00	100.00
Missing none of 18 variables[1] in NDSS and FTND measures	15,203	95.16	96.17
Missing 1 variable in NDSS measure (imputed cases)	151	0.95	0.82
Dependence regardless of missing data	27	0.17	0.26
No dependence regardless of missing data	20	0.13	0.10
Item Nonresponse Patterns that Affect Determination of Dependence Status			
Missing either FTND or NDSS measure	576	3.62	2.66
Missing in FTND measure and "No" in NDSS measure	466	2.92	2.03
Missing 2 variables in NDSS measure	19	0.12	0.06
Missing 3 variables in NDSS measure	18	0.11	0.08
Missing 4 variables in NDSS measure	10	0.06	0.09
Missing 5 variables in NDSS measure	9	0.06	0.04
Missing 6 variables in NDSS measure	6	0.04	0.02
Missing 7 variables in NDSS measure	5	0.03	0.08
Missing 8 variables in NDSS measure	3	0.02	0.01
Missing 9 variables in NDSS measure	2	0.01	0.00
Missing 10 variables in NDSS measure	1	0.01	0.00
Missing 11 variables in NDSS measure	1	0.01	0.00
Missing 12 variables in NDSS measure	3	0.02	0.05
Missing 13 variables in NDSS measure	2	0.01	0.01
Missing 14 variables in NDSS measure	2	0.01	0.00
Missing 15 variables in NDSS measure	1	0.01	0.00
Missing 16 variables in NDSS measure	1	0.01	0.00
Missing 17 variables in NDSS measure	27	0.17	0.19

NDSS = Nicotine Dependence Syndrome; FTND = Fagerstrom Test of Nicotine Dependence.

Note: The weighted percentage is computed using the final analytic weight.

[1] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.23 Item Nonresponse Patterns for Alcohol Dependence (DEPNDALC), 2011 NSDUH

Alcohol Dependence (DEPNDALC)		Frequency		
Total		70,109		
Never used/not used in past year/not used on at least six days in past year		36,406		
Used in past year and used on at least six days in past year		33,703		
Item Nonresponse Patterns among Past Year Users		**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year and used on at least six days in past year		33,703	100.00	100.00
Missing none of 7 criteria[1]		33,042	98.04	98.28
Dependence regardless of missing data		20	0.06	0.04
No dependence regardless of missing data		297	0.88	1.03
Item Nonresponse Patterns that Affect Determination of Dependence Status				
	Total	344	1.02	0.65
Number of criteria true	Number of criteria missing			
0	3 or more	305	0.90	0.53
1	2 or more	16	0.05	0.03
2	1 or more	23	0.07	0.10

Note: The weighted percentage is computed using the final analytic weight.

[1] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.24 Item Nonresponse Patterns for Pain Reliever Dependence (DEPNDANL), 2011 NSDUH

Pain Reliever Dependence (DEPNDANL)		Frequency		
Total		70,109		
Never used/not used in past year		65,425		
Used in past year		4,684		
Item Nonresponse Patterns among Past Year Users		**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year		4,684	100.00	100.00
Missing none of 7 criteria[1]		4,388	93.68	94.74
Dependence regardless of missing data		3	0.06	0.03
No dependence regardless of missing data		53	1.13	1.35
Item Nonresponse Patterns that Affect Determination of Dependence Status				
	Total	240	5.12	3.88
Number of criteria true	Number of criteria missing			
0	3 or more	229	4.89	3.76
1	2 or more	5	0.11	0.04
2	1 or more	6	0.13	0.08

Note: The weighted percentage is computed using the final analytic weight.

[1] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.25 Item Nonresponse Patterns for Cocaine Dependence (DEPNDCOC), 2011 NSDUH

Cocaine Dependence (DEPNDCOC)		Frequency		
Total		70,109		
Never used/not used in past year		68,524		
Used in past year/unknown[1]		1,585		
Item Nonresponse Patterns among Past Year Users		**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year/unknown		1,585	100.00	100.00
Missing none of 7 criteria[2]		1,515	95.58	96.67
Dependence regardless of missing data		2	0.13	0.04
No dependence regardless of missing data		17	1.07	0.51
Item Nonresponse Patterns that Affect Determination of Dependence Status				
	Total	51	3.22	2.77
Number of criteria true	Number of criteria missing			
0	3 or more	49	3.09	2.74
1	2 or more	2	0.13	0.02
2	1 or more	0	0.00	0.00

Note: The weighted percentage is computed using the final analytic weight.

[1] The imputation-revised domain for cocaine dependence includes four cases whose value was unknown.

[2] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.26 Item Nonresponse Patterns for Hallucinogen Dependence (DEPNDHAL), 2011 NSDUH

Hallucinogen Dependence (DEPNDHAL)		Frequency		
Total		70,109		
Never used/not used in past year		67,806		
Used in past year		2,303		
Item Nonresponse Patterns among Past Year Users		**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year		2,303	100.00	100.00
Missing none of 6 criteria[1]		2,193	95.22	95.64
Dependence regardless of missing data		1	0.04	0.01
No dependence regardless of missing data		21	0.91	0.71
Item Nonresponse Patterns that Affect Determination of Dependence Status				
	Total	88	3.82	3.63
Number of criteria true	Number of criteria missing			
0	3 or more	86	3.73	3.62
1	2 or more	1	0.04	0.00
2	1 or more	1	0.04	0.01

Note: The weighted percentage is computed using the final analytic weight.

[1] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.27 Item Nonresponse Patterns for Heroin Dependence (DEPNDHER), 2011 NSDUH

Heroin Dependence (DEPNDHER)		Frequency		
Total		70,109		
Never used/not used in past year		69,841		
Used in past year/unknown[1]		268		
Item Nonresponse Patterns among Past Year Users		**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year/unknown		268	100.00	100.00
Missing none of 7 criteria[2]		259	96.64	97.77
Dependence regardless of missing data		0	0.00	0.00
No dependence regardless of missing data		1	0.37	0.21
Item Nonresponse Patterns that Affect Determination of Dependence Status				
	Total	8	2.99	2.02
Number of criteria true	Number of criteria missing			
0	3 or more	6	2.24	0.97
1	2 or more	2	0.75	1.05
2	1 or more	0	0.00	0.00

Note: The weighted percentage is computed using the final analytic weight.

[1] The imputation-revised domain for heroin dependence includes one case whose value was unknown.

[2] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.28 Item Nonresponse Patterns for Inhalant Dependence (DEPNDINH), 2011 NSDUH

Inhalant Dependence (DEPNDINH)		Frequency		
Total		70,109		
Never used/not used in past year		68,984		
Used in past year		1,125		
Item Nonresponse Patterns among Past Year Users		**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year		1,125	100.00	100.00
Missing none of 6 criteria[1]		983	87.38	90.80
Dependence regardless of missing data		0	0.00	0.00
No dependence regardless of missing data		34	3.02	2.54
Item Nonresponse Patterns that Affect Determination of Dependence Status				
	Total	108	9.60	6.66
Number of criteria true	Number of criteria missing			
0	3 or more	104	9.24	6.34
1	2 or more	2	0.18	0.22
2	1 or more	2	0.18	0.11

Note: The weighted percentage is computed using the final analytic weight.

[1] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.29 Item Nonresponse Patterns for Marijuana Dependence (DEPNDMRJ), 2011 NSDUH

Marijuana Dependence (DEPNDMRJ)		Frequency		
Total		70,109		
Never used/not used in past year/not used on at least six days in past year		60,556		
Used in past year and used on at least six days in past year		9,553		
Item Nonresponse Patterns among Past Year Users		**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year and used on at least six days in past year		9,553	100.00	100.00
Missing none of 6 criteria[1]		9,328	97.64	98.02
Dependence regardless of missing data		9	0.09	0.07
No dependence regardless of missing data		52	0.54	0.57
Item Nonresponse Patterns that Affect Determination of Dependence Status				
	Total	164	1.72	1.34
Number of criteria true	Number of criteria missing			
0	3 or more	147	1.54	1.15
1	2 or more	4	0.04	0.02
2	1 or more	13	0.14	0.17

Note: The weighted percentage is computed using the final analytic weight.

[1] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.30 Item Nonresponse Patterns for Sedative Dependence (DEPNDSED), 2011 NSDUH

Sedative Dependence (DEPNDSED)		Frequency		
Total		70,109		
Never used/not used in past year		69,886		
Used in past year		223		
Item Nonresponse Patterns among Past Year Users		**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year		223	100.00	100.00
Missing none of 7 criteria[1]		199	89.24	91.70
Dependence regardless of missing data		0	0.00	0.00
No dependence regardless of missing data		2	0.90	0.39
Item Nonresponse Patterns that Affect Determination of Dependence Status				
	Total	22	9.87	7.91
Number of criteria true	Number of criteria missing			
0	3 or more	21	9.42	7.83
1	2 or more	0	0.00	0.00
2	1 or more	1	0.45	0.08

Note: The weighted percentage is computed using the final analytic weight.

[1] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.31 Item Nonresponse Patterns for Stimulant Dependence (DEPNDSTM), 2011 NSDUH

Stimulant Dependence (DEPNDSTM)		Frequency		
Total		70,109		
Never used/not used in past year		68,888		
Used in past year/unknown[1]		1,221		
Item Nonresponse Patterns among Past Year Users		**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year/unknown		1,221	100.00	100.00
Missing none of 7 criteria[2]		1,161	95.09	93.54
Dependence regardless of missing data		1	0.08	0.04
No dependence regardless of missing data		6	0.49	0.21
Item Nonresponse Patterns that Affect Determination of Dependence Status				
Total		53	4.34	6.20
Number of criteria true	Number of criteria missing			
0	3 or more	50	4.10	5.81
1	2 or more	2	0.16	0.28
2	1 or more	1	0.08	0.11

Note: The weighted percentage is computed using the final analytic weight.

[1] The imputation-revised domain for stimulant dependence includes 10 cases whose value was unknown.

[2] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.32 Item Nonresponse Patterns for Tranquilizer Dependence (DEPNDTRN), 2011 NSDUH

Tranquilizer Dependence (DEPNDTRN)		Frequency		
Total		70,109		
Never used/not used in past year		68,194		
Used in past year		1,915		
Item Nonresponse Patterns among Past Year Users		**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year		1,915	100.00	100.00
Missing none of 6 criteria[1]		1,829	95.51	96.48
Dependence regardless of missing data		0	0.00	0.00
No dependence regardless of missing data		20	1.04	0.71
Item Nonresponse Patterns that Affect Determination of Dependence Status				
Total		66	3.45	2.82
Number of criteria true	Number of criteria missing			
0	3 or more	64	3.34	2.79
1	2 or more	2	0.10	0.03
2	1 or more	0	0.00	0.00

Note: The weighted percentage is computed using the final analytic weight.

[1] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.33 Item Nonresponse Patterns for Alcohol Abuse (ABUSEALC), 2011 NSDUH

Alcohol Abuse (ABUSEALC)	Frequency		
Total	70,109		
Never used/not used in past year/not used on at least six days in past year	36,406		
Used in past year and used on at least six days in past year	33,703		
Item Nonresponse Patterns among Past Year Users	**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year and used on at least six days in past year	33,703	100.00	100.00
Missing none of 4 criteria[1]	33,395	99.09	99.57
Abuse regardless of missing data	3	0.01	0.00
No abuse regardless of missing data	7	0.02	0.01
One or more criteria missing when no criteria is true	298	0.88	0.42

Note: The weighted percentage is computed using the final analytic weight.

[1] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.34 Item Nonresponse Patterns for Prescription Pain Reliever Abuse (ABUSEANL), 2011 NSDUH

Prescription Pain Reliever Abuse (ABUSEANL)	Frequency		
Total	70,109		
Never used/not used in past year	65,425		
Used in past year	4,684		
Item Nonresponse Patterns among Past Year Users	**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year	4,684	100.00	100.00
Missing none of 4 criteria[1]	4,449	94.98	96.11
Abuse regardless of missing data	0	0.00	0.00
No abuse regardless of missing data	2	0.04	0.02
One or more criteria missing when no criteria is true	233	4.97	3.88

Note: The weighted percentage is computed using the final analytic weight.

[1] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.35 Item Nonresponse Patterns for Cocaine Abuse (ABUSECOC), 2011 NSDUH

Cocaine Abuse (ABUSECOC)	Frequency		
Total	70,109		
Never used/not used in past year	68,524		
Used in past year/unknown[1]	1,585		
Item Nonresponse Patterns among Past Year Users	**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year/unknown	1,585	100.00	100.00
Missing none of 4 criteria[2]	1,533	96.72	97.06
Abuse regardless of missing data	0	0.00	0.00
No abuse regardless of missing data	0	0.00	0.00
One or more criteria missing when no criteria is true	52	3.28	2.94

Note: The weighted percentage is computed using the final analytic weight.

[1] The imputation-revised domain for cocaine dependence includes four cases whose value was unknown.

[2] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.36 Item Nonresponse Patterns for Hallucinogen Abuse (ABUSEHAL), 2011 NSDUH

Hallucinogen Abuse (ABUSEHAL)	Frequency		
Total	70,109		
Never used/not used in past year	67,806		
Used in past year	2,303		
Item Nonresponse Patterns among Past Year Users	**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year	2,303	100.00	100.00
Missing none of 4 criteria[1]	2,214	96.14	96.35
Abuse regardless of missing data	0	0.00	0.00
No abuse regardless of missing data	1	0.04	0.03
One or more criteria missing when no criteria is true	88	3.82	3.62

Note: The weighted percentage is computed using the final analytic weight.

[1] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.37 Item Nonresponse Patterns for Heroin Abuse (ABUSEHER), 2011 NSDUH

Heroin Abuse (ABUSEHER)	Frequency		
Total	70,109		
Never used/not used in past year	69,841		
Used in past year/unknown[1]	268		
Item Nonresponse Patterns among Past Year Users	**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year/unknown	268	100.00	100.00
Missing none of 4 criteria[2]	259	96.64	97.34
Abuse regardless of missing data	0	0.00	0.00
No abuse regardless of missing data	1	0.37	0.43
One or more criteria missing when no criteria is true	8	2.99	2.23

Note: The weighted percentage is computed using the final analytic weight.

[1] The imputation-revised domain for heroin dependence includes one case whose value was unknown.

[2] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.38 Item Nonresponse Patterns for Inhalant Abuse (ABUSEINH), 2011 NSDUH

Inhalant Abuse (ABUSEINH)	Frequency		
Total	70,109		
Never used/not used in past year	68,984		
Used in past year	1,125		
Item Nonresponse Patterns among Past Year Users	**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year	1,125	100.00	100.00
Missing none of 4 criteria[1]	1,018	90.49	93.31
Abuse regardless of missing data	0	0.00	0.00
No abuse regardless of missing data	0	0.00	0.00
One or more criteria missing when no criteria is true	107	9.51	6.69

Note: The weighted percentage is computed using the final analytic weight.

[1] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.39 Item Nonresponse Patterns for Marijuana Abuse (ABUSEMRJ), 2011 NSDUH

Marijuana Abuse (ABUSEMRJ)	Frequency		
Total	70,109		
Never used/not used in past year/not used on at least six days in past year	60,556		
Used in past year and used on at least six days in past year	9,553		
Item Nonresponse Patterns among Past Year Users	**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year and used on at least six days in past year	9,553	100.00	100.00
Missing none of 4 criteria[1]	9,392	98.31	98.71
Abuse regardless of missing data	1	0.01	0.00
No abuse regardless of missing data	7	0.07	0.04
One or more criteria missing when no criteria is true	153	1.60	1.25

Note: The weighted percentage is computed using the final analytic weight.

[1] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.40 Item Nonresponse Patterns for Sedative Abuse (ABUSESED), 2011 NSDUH

Sedative Abuse (ABUSESED)	Frequency		
Total	70,109		
Never used/not used in past year	69,886		
Used in past year	223		
Item Nonresponse Patterns among Past Year Users	**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year	223	100.00	100.00
Missing none of 4 criteria[1]	203	91.03	93.51
Abuse regardless of missing data	0	0.00	0.00
No abuse regardless of missing data	0	0.00	0.00
One or more criteria missing when no criteria is true	20	8.97	6.49

Note: The weighted percentage is computed using the final analytic weight.

[1] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.41 Item Nonresponse Patterns for Stimulant Abuse (ABUSESTM), 2011 NSDUH

Stimulant Abuse (ABUSESTM)	Frequency		
Total	70,109		
Never used/not used in past year	68,888		
Used in past year/unknown[1]	1,221		
Item Nonresponse Patterns among Past Year Users	**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year/unknown	1,221	100.00	100.00
Missing none of 4 criteria[2]	1,171	95.90	93.83
Abuse regardless of missing data	0	0.00	0.00
No abuse regardless of missing data	1	0.08	0.04
One or more criteria missing when no criteria is true	49	4.01	6.12

Note: The weighted percentage is computed using the final analytic weight.

[1] The imputation-revised domain for stimulant dependence includes 10 cases whose value was unknown.

[2] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.42 Item Nonresponse Patterns for Tranquilizer Abuse (ABUSETRN), 2011 NSDUH

Tranquilizer Abuse (ABUSETRN)	Frequency		
Total	70,109		
Never used/not used in past year	68,194		
Used in past year	1,915		
Item Nonresponse Patterns among Past Year Users	**Frequency**	**Unweighted Percentage**	**Weighted Percentage**
Used in past year	1,915	100.00	100.00
Missing none of 4 criteria[1]	1,850	96.61	97.07
Abuse regardless of missing data	0	0.00	0.00
No abuse regardless of missing data	0	0.00	0.00
One or more criteria missing when no criteria is true	65	3.39	2.93

Note: The weighted percentage is computed using the final analytic weight.

[1] Nonmissing variables include values equal to 1 (Yes), 2 (No), or 99 (Legitimate Skip).

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, Restricted-Use Data, 2011.

Table K.43 Model Summaries for WSHD, Pain Reliever Dependence and Abuse

Variable	Variable Description	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
ANLLOTTM	Spent month/more getting/using pain relievers in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLGTOVR
ANLGTOVR	Month/more spent getting over pain reliever effects in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLLOTTM
ANLLIMIT	Able to keep limits or use pain relievers more in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLKPLMT
ANLKPLMT	Set limits on pain reliever use in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLLIMIT
ANLNDMOR	Needed more pain relievers to get same effect in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLLSEFX

Table K.43 Model Summaries for WSHD, Pain Reliever Dependence and Abuse (continued)

Variable	Variable Description	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
ANLLSEFX	Using same amount of pain relievers has less effect in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLNDMOR
ANLCUTDN	Want or try to cut down using pain relievers in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLCUTEV
ANLCUTEV	Able to cut/stop using pain relievers every time in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLCUTDN
ANLCUT1X	Cut down or stop using pain relievers at least one time in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLCUTEV
ANLWD3SX	Had 3+ pain reliever withdrawal symptoms in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLCUT1X

Table K.43 Model Summaries for WSHD, Pain Reliever Dependence and Abuse (continued)

Variable	Variable Description	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
ANLWDSMT	Had 3+ pain reliever withdrawal symptoms at same time in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLWD3SX
ANLEMOPB	Pain relievers cause problems with emotions or nerves in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLEMCTD
ANLEMCTD	Continued to use pain relievers despite emotional problems	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLEMOPB
ANLPHLPB	Any physical problems caused or worsened by pain relievers in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLEMCTD
ANLPHCTD	Continued to use pain relievers despite physical problems	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLEMCTD

Table K.43 Model Summaries for WSHD, Pain Reliever Dependence and Abuse (continued)

Variable	Variable Description	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
ANLLSACT	Less activities because of pain reliever use in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLSERPB, Imputation-Revised ANLNDMOR, Imputation-Revised ANLPHLPB
ANLSERPB	Pain relievers cause serious problems at home/work/school in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLLSACT, Imputation-Revised ANLFMFPB
ANLPDANG	Using pain relievers and doing dangerous activities in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLFMCTD, Imputation-Revised ANLPHLPBR_I
ANLLAWTR	Using pain relievers causes problems with law in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLFMCTD, Imputation-Revised ANLPDANG
ANLFMFPB	Using pain relievers causes problems with family/friends in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLFMCTD

Table K.43 Model Summaries for WSHD, Pain Reliever Dependence and Abuse (continued)

Variable	Variable Description	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
ANLFMCTD	Continued to use pain relievers despite problems with family/friends	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Pain Reliever Recency, Imputation-Revised Pain Reliever Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised ANLFMFPB

Table K.44 Model Summaries for WSHD, Stimulant Dependence and Abuse

Variable	Variable Description	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
STMLOTTM	Spent month/more getting/using stimulants in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised Stimulant Frequency of Use
STMGTOVR	Month/more spent getting over stimulant effects in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised STMLOTTM
STMLIMIT	Set limits on stimulant use in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised Stimulant Frequency of Use
STMKPLMT	Able to keep limits or use stimulants more in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised STMLIMIT
STMNDMOR	Needed more stimulants to get same effect in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised Stimulant Frequency of Use

Table K.44 Model Summaries for WSHD, Stimulant Dependence and Abuse (continued)

Variable	Variable Description	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
STMLSEFX	Using same amount of stimulants has less effect in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised STMNDMOR
STMCUTDN	Want or try to cut down using stimulants in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised Stimulant Frequency of Use
STMCUTEV	Able to cut/stop using stimulants every time in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised STMCUTDN
STMCUT1X	Cut down or stop using stimulants at least one time in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised STMCUTEV
STMFLBLU	Felt blue in past 12 months when cut down on stimulants	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised STMCUT1X

Table K.44 Model Summaries for WSHD, Stimulant Dependence and Abuse (continued)

Variable	Variable Description	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
STMWD2SX	Had 2+ stimulant withdrawal symptoms in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised STMFLBLU
STMWDSMT	Had 2+ stimulant withdrawal symptoms at same time in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised STMWD2SX
STMEMOPB	Stimulants cause problems with emotions or nerves in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised Stimulant Frequency of Use
STMEMCTD	Continued to use stimulants despite emotional problems	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised STMEMOPB
STMPHLPB	Any physical problems caused or worsened by stimulants in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised STMEMCTD

Table K.44 Model Summaries for WSHD, Stimulant Dependence and Abuse (continued)

Variable	Variable Description	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
STMPHCTD	Continued to use stimulants despite physical problems	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised STMPHLPB
STMLSACT	Less activities because of stimulant use in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised Stimulant Frequency of Use
STMSERPB	Stimulants cause serious problems at home/work/school in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised Stimulant Frequency of Use
STMPDANG	Using stimulants and doing dangerous activities in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised Stimulant Frequency of Use
STMLAWTR	Using stimulants causes problems with law in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised Stimulant Frequency of Use

Table K.44 Model Summaries for WSHD, Stimulant Dependence and Abuse (continued)

Variable	Variable Description	Starting List of Predictor Variables	Predictor Variables Used in Imputation Classes
STMFMFPB	Using stimulants causes problems with family/friends in past 12 months	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised Stimulant Frequency of Use
STMFMCTD	Continued to use stimulants despite problems with family/friends	Gender; Age; Imputation-Revised Age at First Use, Imputation-Revised Stimulant Recency, Imputation-Revised Stimulant Frequency of Use, All Dependence and Abuse Item-Level Variables	Imputation-Revised STMFMFPB

Appendix L: Income Item Nonresponse Patterns

This page intentionally left blank

L.1 Introduction

Low response rates to income questions and resulting nonresponse bias are well documented in the survey research literature (Bollinger, Hirsch, Hokayem, & Ziliak, 2014; Tourangeau & Yan, 2007; Pleis & Dahlhamer, 2004; Moore, Stinson, & Welniak, Jr., 2000; Juster & Smith, 1997). Like many other household surveys, the family income variables measured in the National Survey on Drug Use and Health (NSDUH) have much lower response rates than the vast majority of other questionnaire items. Typically, approximately 90 percent of variables that underwent statistical imputation required less than 5 percent of their records to be logically assigned or statistically imputed (Center for Behavioral Health Statistics and Quality, 2015, Appendix A). In 2014, 6,589 cases were missing for finer categories of total family income resulting in a weighted nonresponse rate of 10.46 percent. This relatively high item nonresponse rate is of interest because the NSDUH family income variables and their recodes are used in many analyses, and the distribution of public health variables by income levels has implications for policy decisions. Any steps that could be taken to reduce the impact of item nonresponse, to improve the imputation method, or to improve the questionnaire would be helpful with reducing nonresponse bias and improving overall data quality. By improving the understanding of the mechanisms of item nonresponse for income, this appendix describes solutions and recommendations for further addressing and reducing item nonresponse and/or nonresponse bias in variables measuring total family income and poverty through better imputation methods or by making changes to the questionnaire to reduce item nonresponse.

L.2 Measurement of Total Income in NSDUH

NSDUH estimates respondents' total income for adults and youths aged 12 to 17 by asking about total personal income and total family income, based on two questions:

1. Of these income groups, which category best represents (your/SAMPLE MEMBER's) total personal income during [the previous calendar year]?
2. Of these income groups, which category best represents (your/SAMPLE MEMBER's) total combined family income during [the previous calendar year]?

Respondents receive these questions after being routed through an unfolding bracket of related income questions in order to minimize income item nonresponse as much as possible.

Family is defined as any related member in the household roster, including all foster relationships and unmarried partners (including same-sex partners). Roommates, boarders, and other nonrelatives are excluded from the definition of family for total family income. Responses from proxies are accepted for items of health insurance and income from a family member living in the same household who is identified as being better able to give the correct information. The NSDUH questionnaire allows respondents to decline to answer any question (except age) by entering "Don't know" (DK) or "Refused" (REF) as a response.

Total family income is also used to establish a respondent's poverty level using the NSDUH data. Poverty level is determined by comparing a respondent's total family income with the U.S. Census Bureau's poverty thresholds (both measured in dollar amounts), with a respondent's family size and composition (i.e., number of children) taken into consideration. The

resulting variables indicating levels of poverty are often used in NSDUH analyses. When total family income is missing, the poverty level is also unknown.

L.2.1 Questionnaire Skip Logic

Total income is measured in NSDUH through an unfolding bracket of questionnaire items that can be understood as steps. These steps are illustrated in the flowchart in Figure L.1. There are 29 finer categories of personal and family income that are captured by the variables PINC2 and FINC2, respectively. Table L.1 shows the binary and finer categories of income.

Figure L.1 2014 NSDUH Questionnaire Measurement of Total Income

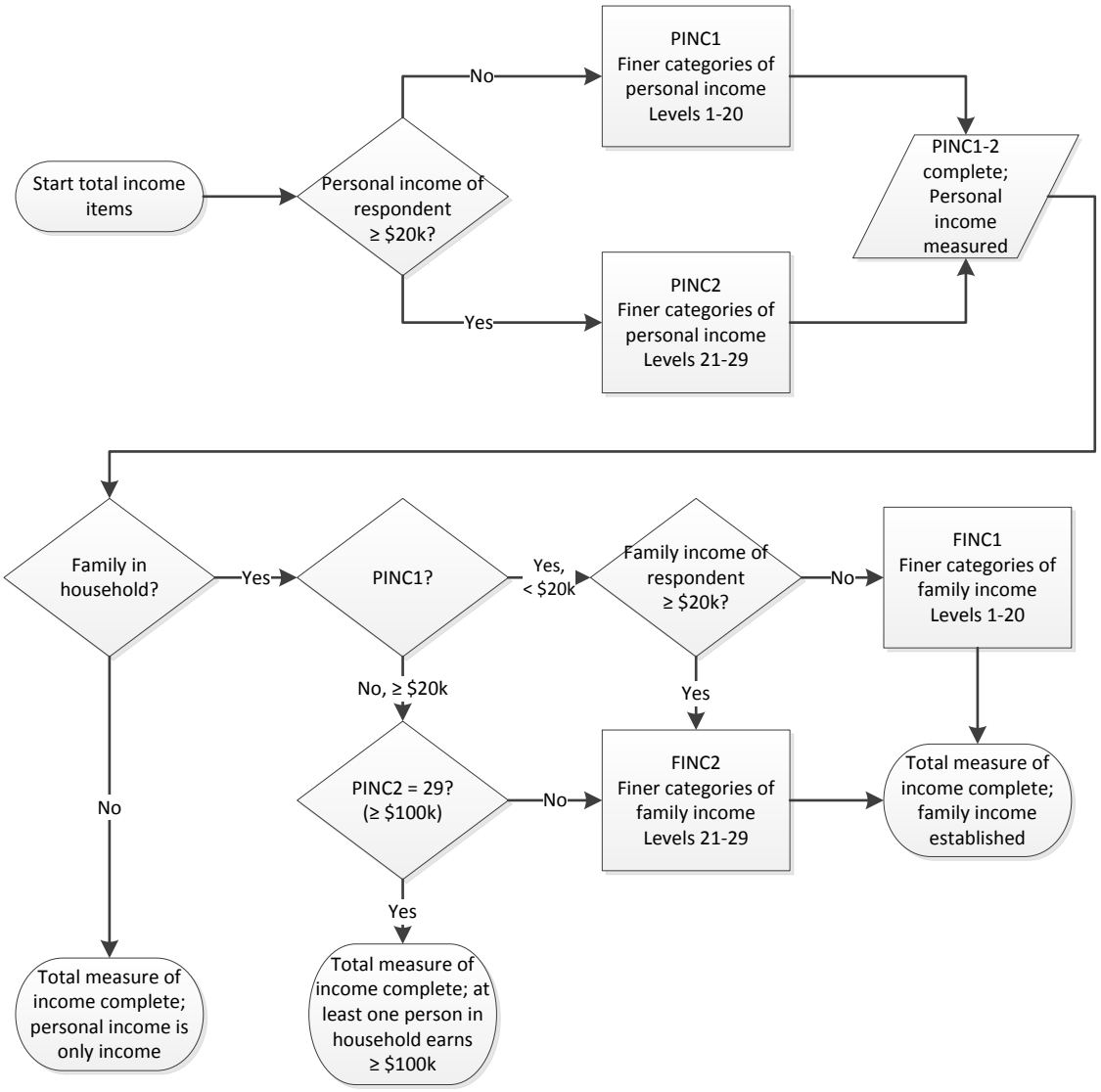


Table L.1 Binary Categories of Personal and Family Total Income: 2014 NSDUH

Binary Category (PINC1, FINC1)		Finer Category (PINC2, FINC2)	
Level	**Description**	**Level**	**Description**
2	< $20,000	1	< $1,000
		2	$1,000-$1,999
		⋮	
		20	$19,000-$19,999
1	≥ $20,000	21	$20,000-$24,999
		⋮	
		26	$45,000-$49,999
		27	$50,000-$75,999
		28	$75,000-$99,999
		29	≥ $100,000[1]

[1] This final measurement category of income was measured in the 2014 and previous iterations of NSDUH. Beginning in 2015, this category became $100,000-$149,999 and a 30th measurement category of ≥ $150,000 was added.

The first step of the income questionnaire section establishes whether the personal income of the respondent is greater than or equal to $20,000 with a binary questionnaire item. This response is captured by PINC1. The next step in measuring total personal income is to determine the finer category of personal income for each respondent.

Respondents reporting personal incomes that are less than $20,000 are directed to the finer personal income categories with levels ranging from less than $1,000 (PINC2 = 1) to $19,000-$19,999 (PINC2 = 20). Respondents reporting personal incomes that are greater than or equal to $20,000 are directed to the finer personal income categories, ranging from $20,000-$24,999 (PINC2 = 21) to $100,000 or more (PINC2 = 29). This response is captured by PINC2.

After measuring finer categories of total personal income, respondents are routed to questionnaire items measuring total family income depending on the presence of family members in the household. This information is reported in the household roster. If the respondent did not report any family members in his or her household, then the questionnaire item on total family income is not asked and the personal income response (PINC2) is used for total family income for that respondent. If the respondent reported family members in the household and the personal income level that is reported in PINC1 is greater than or equal to $20,000, then the respondent is directly routed to the finer categories of family income question, FINC2, skipping the binary family income questionnaire item FINC1.

If the respondent reports the highest level of response available for personal income, $100,000 or more (PINC2 = 29), total family income is automatically completed as $100,000 or more (FINC2 = 29). If reported personal income is less than $20,000, the questionnaire then asks respondents whether the total family income is greater than or equal to $20,000 (FINC1). Based on this response, respondents are directed to one of the two questionnaire items measuring finer categories of family income (captured by FINC2). FAMINC2 is the final resulting variable measuring the finer categories of family income. It is equal to PINC2 for those with no other

family members in the household roster, and it is equal to FINC2 for those with other family members in the household roster.

L.2.2 Paths to Income Item Nonresponse

Because of the skip logic and questionnaire routing, there are six ways, or "paths," in which the respondent could have a valid value for total family income, and there are nine opportunities for respondents to become defined as missing in the total family income measure (see Table L.2 and Figure L.2 for paths of nonresponse). The flowchart in Figure L.2 illustrates how and where these valid or missing values occur based on the structure of the income questionnaire items. The numbered missing nodes correspond with the paths of nonresponse shown in Table L.2.

Table L.2 Paths of Finer Categories of Family Income Item Nonresponse: 2014 NSDUH

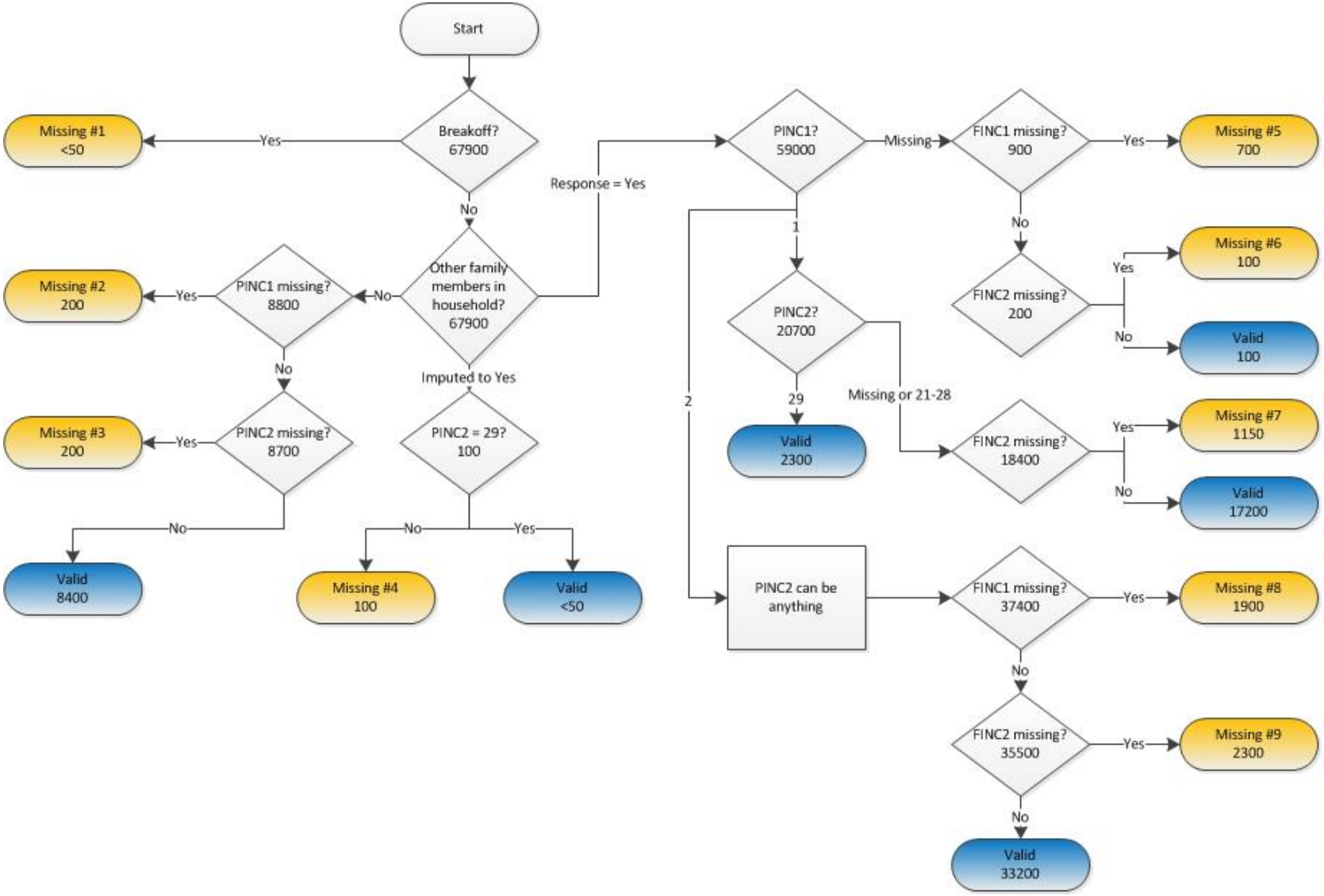
	Path of Income Item Nonresponse	Frequency	Percent	Rank by Frequency Missing
1	Breakoff	27	0.41	9
2	No Family in HH, PINC1 Missing	182	2.76	6
3	No Family in HH, PINC2 Missing	241	3.66	5
4	Imputed to Family in HH, PINC2 ≠ 29	56	0.85	7 (tie)
5	Family in HH, PINC1 Missing, FINC1 Missing	686	10.41	4
6	Family in HH, PINC1 Missing, FINC1 Valid, FINC2 Missing	56	0.85	7 (tie)
7	Family in HH, PINC1 = 1, PINC2 ≠ 29, FINC2 Missing	1,150	17.45	3
8	Family in HH, PINC1 = 2, FINC1 Missing	1,929	29.28	2
9	Family in HH, PINC1 = 2, FINC1 Valid, FINC2 Missing	2,262	34.33	1
Total Income Item Nonresponse		6,589	100.00	N/A

HH = household.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2014.

The two most common paths of income item nonresponse occur for sample members with personal incomes that are less than $20,000 living in households with family members who either (1) responded to the binary total family income questionnaire item but did not respond to the finer categories of total family income question (nonrespondents = 2,262) or (2) did not respond to the binary total family income questionnaire item and thus did not receive the finer categories of total family income question (nonrespondents = 1,929). The third most common path to item nonresponse for total family income occurs for sample members with personal incomes that are greater than $20,000 but less than $100,000 living in households with family members who do not answer the finer categories of family income question (nonrespondents = 1,150). This is followed by the fourth most common path of family income item nonresponse for respondents living with family members in the household who did not answer the binary personal or family income items at all (nonrespondents = 686). The next most common paths of family income item nonresponse occur for respondents without any family members living in the same household and either do not respond to the binary personal income questionnaire item (nonrespondents = 182) or respond to the binary personal income item and do not respond to the finer categories of personal income (nonrespondents = 241). The three remaining paths of family income item nonresponse are less common with fewer than 60 nonrespondents each.

Figure L.2 Income Item Response and Nonresponse Paths with Frequencies: 2014 NSDUH



Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2014.

L.3 Potential Solutions and Recommendations

In NSDUH, income item nonresponse has historically been handled using imputation. Some recent papers (Bollinger et al., 2014; Pleis & Dahlhamer, 2004) suggest that income nonresponse depends, in part, on the unknown values of the variable that are not actually observed, so income item nonresponse can be described as not missing at random (NMAR). Because income item nonresponse is NMAR, imputation involving only auxiliary variables does not completely correct for the present nonresponse bias and resulting measurement error (Little & Rubin 1987; Frechtel & Copello, 2007). However, NSDUH has certain survey features that can be used to more accurately decrease income item nonresponse. This section presents potential solutions and recommendations for the future based on reclaiming[1] missing income responses from family pair members.

L.3.1 Reclaiming with Other Family Pair Member

In each household selected for NSDUH, zero, one, or two household members are selected for interviewing. When two members of the same household are selected and both complete an interview, a "responding pair" is formed. In the 2014 NSDUH, 53.7 percent of the unit respondents were members of a responding pair. The pair relationship can be parent-child, sibling-sibling, spouse-spouse, or some other relationship.

Chapter 8 of this report discusses the use of the other pair member as a donor in imputation when exactly one pair member is missing for all variables that currently undergo imputation. In that chapter, family income is considered a good candidate for this sort of logical imputation whenever the pair members are members of the same family. Because most of the questions in the income section ask about the family income of those in the household, given the case where the pair members are members of the same family, the only source of disagreement should be measurement error (Frechtel, Scott, Couzens, Moore, & Bose, 2012). This section describes reclaiming values of family income that sometimes involves the data reported by the respondent and sometimes involves the data reported by the other family pair member (OFPM) about the family. The following two key recoded variables involving family income are considered:

1. INCOME5, a straightforward five-level recode of finer income; and
2. POVERTY2, a three-level recode of not only family income but also of roster information and type of household.

Although recodes on NSDUH typically do not have imputation indicators associated with them, it may be beneficial to add them because the imputation indicators associated with INCOME5 and POVERTY2 would differ from the imputation indicator for the family finer categories income variable IRFAMIN2, called IIFAMIN2.

[1] To "reclaim" in this context is to assign a value for a recode using logic when some or all of the parent variables for the recode have missing values. It might also be called "logical imputation."

L.3.1.1 Reclaiming Missing Values for INCOME5

Table L.3 shows how the 29 levels of IRFAMIN2 map to the 5 levels of INCOME5. The highest three levels are the same for both variables, but the lower levels are heavily aggregated.

Table L.3 Mapping of IRFAMIN2 Levels to INCOME5 Levels

IRFAMIN2		INCOME5	
Level	**Description**	**Level**	**Description**
1	< $1,000	1	< $20,000
2	[$1,000, $2,000)		
⋮	⋮		
20	[$19,000, $20,000)		
21	[$20,000, $25,000)	2	[$20,000, $50,000)
⋮	⋮		
26	[$45,000, $50,000)		
27	[$50,000, $75,000)	3	[$50,000, $75,000)
28	[$75,000, $100,000)	4	[$75,000, $100,000)
29	≥ $100,000	5	≥ $100,000

There are exactly 661 respondents in the 2014 data whose missing values for IRFAMIN2 can be reclaimed using only the self's data. For these types of respondents, it is known that INCOME5 = 1 even though IRFAMIN2 is missing. This can occur in two ways:

1. The respondent has no other family members in the household, reported having a personal income of less than $20,000 (PINC1 = 2), and had a missing value for the personal finer categories income variable (PINC2).
2. The respondent has other family members in the household, reported having a family income of less than $20,000 (FINC1 = 2), and had a missing value for the personal finer categories income variable (FINC2).

Using the self, 661 (10.03 percent) of the 6,589 missing values for IRFAMIN2 can be reclaimed. Of the remaining 5,928 cases with missing values, 1,267 (21.37 percent) cases had an OFPM with a nonmissing value for FAMINC2, and 60 (1.01 percent) other cases had an OFPM with family binary income of less than $20,000. In total, the reclaiming process can reduce the item nonresponse rate for INCOME5 from 9.70 percent to 6.78 percent (Table L.4).

Table L.4 Reclaiming of Missing Values of INCOME5

	Count	**Nonresponse for INCOME5 (%)**
Cases with Missing FAMINC2	6,589	9.70
Minus Cases with FAMINC1 = 2	661	8.73
Minus Cases Where OFPM Has Nonmissing FAMINC2	1,267	6.86
Minus Cases Where OFPM Has FAMINC1 = 2	60	6.78
Final Nonresponse Rate for INCOME5	**4,601**	**6.78**

L.3.1.2 Reclaiming Missing Values for POVERTY2

The variable POVERTY2 is a complex recode involving the imputation-revised family finer categories income variable IRFAMIN2, plus the respondent's age (AGE) and the roster variables IRFMLYSZ (imputation-revised household size including fosters) and IRKDFMLY (imputation-revised number of children in household including fosters). The variable is created in two steps. In the first step, the poverty threshold is calculated based on a formula from the U.S. Census Bureau involving values that are captured by the variables AGE, IRFMLYSZ (imputation-revised family size), and IRKDFMLY (imputation-revised number of children younger than 18 in the household).[2] In the second step, the family income variable IRFAMIN2 (actually the midpoint of the interval associated with the value of IRFAMIN2) is compared with the poverty threshold. POVERTY2 has four levels:

1. The respondent is 18 to 22 years old and lives in a college dorm (POVERTY2 = missing).
2. The family income is less than the poverty threshold (POVERTY2 = 1).
3. The family income is greater than or equal to the poverty threshold but less than twice the poverty threshold (POVERTY2 = 2).
4. The family income is greater than twice the poverty threshold (POVERTY2 = 3).

Note that the pre-imputation versions of IRFMLYSZ and IRKDFMLY have missing values as well, but the item nonresponse rates for these (less than 1 percent) are much smaller in comparison with the item nonresponse rate associated with the pre-imputation version of IRFAMIN2 (9.70 percent weighted, 10.46 percent unweighted). Because of the relatively low item nonresponse rates associated with these household size variables, it is assumed that there are no missing data for the two household composition variables, and reclaiming efforts for POVERTY2 are focused on obtaining valid responses from income variables for this investigation.

In total, 1,964 cases with missing values for FAMINC2 can be reclaimed if an imputation indicator were created for POVERTY2. The first step in reclaiming is to calculate bounds for the family income based on both the person's responses and on the OFPM's responses. The possible lower bounds for FAMINC2 are the following:

- (L1) 21, if PINC1 = 1;
- (L2) PINC2, if $2 \leq$ PINC2 ≤ 28; and
- (L3) 21, if FINC1 = 1.

The possible upper bounds for FAMINC2 are the following:

- (U1) 20, if PINC1 = 2, PINC2 is missing, and the person lives with no other family members; and
- (U2) 20, if FINC1 = 2.

[2] See https://www.census.gov/topics/income-poverty/poverty.html.

All of these bounds were calculated for both the self and the OFPM, except U1, which was only calculated for the self. If the self has no other family members in the household, then it is assumed that the OFPM's data would not be available.

Table L.5 shows the reclaiming of missing values for POVERTY2 relative to FAMINC2. Probably the most interesting category is the 612 cases that can be reclaimed using the self data. Of the 612 cases, 277 (45.26 percent) can be reclaimed because the family income is definitely greater than twice the poverty threshold, 320 (52.29 percent) can be reclaimed because the family income is definitely less than the poverty threshold, and 15 (2.45 percent) can be reclaimed because the family income is definitely greater than the poverty threshold and less than twice the poverty threshold. Overall, the reclaiming process can reduce the item nonresponse rate for POVERTY2 from 9.70 percent to 6.81 percent.

Table L.5 Reclaiming of Missing Values of POVERTY2

	Count	Nonresponse for INCOME5 (%)
Cases with Missing FAMINC2	6,589	9.70
Minus Cases with College Students in Dorms	8	9.69
Minus Cases Where POVERTY2 Value Can Be Determined Based on the Self Data	612	8.79
Minus Cases Where OFPM Has Nonmissing FAMINC2	1,288	6.89
Minus Cases Where POVERTY2 Value Can Be Determined Using Bounds from the OFPM Data	56	6.81
Final Nonresponse Rate for POVERTY2		6.81

L.3.1.3 Next Steps for Reclaiming Income with Other Family Pair Member

One problem with editing income using the OFPM data is that a direct assignment can create an inconsistent record. There are 130 cases in the 2014 data where a direct assignment of the OFPM's value for FAMINC2 would create an inconsistent record. These are mostly cases where the self has FINC1 = 1 and the OFPM has FINC1 = 2, or the self has FINC1 = 2 and the OFPM has FINC1 = 1. There are also 24 cases in the 2014 data where the OFPM has a missing value for FAMINC2, but the bounds based on the OFPM's responses are inconsistent with bounds based on the self's responses. For example, the self might have FINC1 = 2, the OFPM might have FINC1 = 1, and both pair members might have FINC2 missing.

The direct-assignment approach rests on an implicit assumption that these pair members agree when they both respond. This is not always the case. When both family pair members responded in the 2014 survey, they agreed 68.64 percent of the time, suggesting that there is non-negligible measurement error present (Table L.6). Sometimes the responses of the family pair members are quite different. For example, they disagreed by greater than four or more levels of FAMINC2 8.59 percent of the time.

Table L.6 Disagreement in Family Income among Family Pair Members: 2014 NSDUH

Number of Income Levels between Responses	Number of Pairs	Percentage of Pairs
0	9,441	68.64
1	2,013	14.63
2	754	5.48
3	366	2.66
4+	1,181	8.59

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2014.

The response agreement between the responding family pair members across pair types shows that the agreement rates for grandparent-grandchild pairs (56.99 percent) and spouse-spouse pairs (57.36 percent) are lowest, and the agreement rates among sibling-sibling pairs where one of the pair members is aged 12 to 14 are highest (84.51 percent) (Table L.7). The low agreement rate for spouse-spouse pairs may suggest measurement error associated with a lack of knowledge. The relatively high agreement rates among parent-child pairs where one of the pair members is aged 12 to 14 (77.66 percent) or 15 to 17 (77.65 percent) may be a result of the higher number of proxy responses for the family income item in this age group, and it is possible that the parent pair member gave responses as a proxy for the child pair member.

Table L.7 Agreement by Pair Type among Responding Pairs for Family Finer Categories Income: 2014 NSDUH

Pair Type	Responding Pairs		Percent Agreement among Pairs
	Number	Percent	
Parent-Child, Child Aged 12-14	2,596	41.63	77.66
Parent-Child, Child Aged 15-17	2,139	34.40	77.65
Parent-Child, Child Aged 18-20	711	11.40	59.92
Parent-Child, Child Aged 21+	790	12.67	56.33
Parent-Child Total	**6,236**	**100.00**	**72.93**
Sibling-Sibling, Youngest 12-14/Oldest 15-17	1,291	31.00	84.51
Sibling-Sibling, Youngest 12-17/Oldest 18-25	1,330	31.94	63.53
Sibling-Sibling, Other Age Pairings	1,543	37.06	66.88
Sibling-Sibling Total	**4,164**	**100.00**	**71.28**
Spouse-Spouse with Children	1,631	51.55	56.59
Spouse-Spouse without Children	1,511	47.76	58.37
Spouse-Spouse, Children Not Clear	22	0.70	45.45
Spouse-Spouse Total	**3,164**	**100.00**	**57.36**
Grandparent-Grandchild	**193**	**100.00**	**56.99**
Overall Total	**13,757**	**100.00**	**68.63**

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2014.

A caveat to using OFPM data is that sometimes the pair members disagree on whether they are part of a family. Other-pair-member roster editing has revealed that, especially for blended families and families involving unmarried partners, the pair members occasionally disagree on the nature of their relationship. The imputation-revised family-skip variable IRFAMSKP does not have to be consistent among the pair members (though the imputation-revised pair relationship variable IRPRREL does), and those with IRFAMSKP = 1 are

automatically assigned skip codes for edited and imputation-revised income variables related to the family.

For these reasons, if these methods were to be put into practice, it is recommended that the OFPM's values be used only if all three of the following conditions are met:

1. According to IRPRREL, the pair members are members of the same family.
2. Both pair members have IRFAMSKP = 0.
3. The values of the other pair member are consistent with the nonmissing responses given by the respondent.

L.3.2 Proxy Respondents

Respondents' selection of a proxy tends to improve the response rate for income variables. The next step is to examine whether proxy responses tend to improve agreement between the pair members. Table L.8 shows the agreement among pairs with at least one proxy respondent compared with no proxy respondents, by pair type. For every pair type, the agreement is higher, usually *much* higher, for pairs with at least one proxy respondent. Pair types with high levels of agreement in the previous table (Table L.7) are the pair types that tend to use proxies frequently: namely, the pairs with children aged 12 to 17.

Table L.8 Influence of Proxy Respondents on Agreement by Pair Type among Responding Pairs for Family Finer Categories Income: 2014 NSDUH

Pair Type	Responding Pairs		Percent Agreement among Pairs	
	Number	Percent with at Least One Proxy	At Least One Proxy	No Proxies
Parent-Child, Child Aged 12-14	2,596	97.34	77.92	68.12
Parent-Child, Child Aged 15-17	2,139	92.01	79.98	50.88
Parent-Child, Child Aged 18-20	711	49.93	75.49	44.38
Parent-Child, Child Aged 21+	790	30.25	78.66	46.64
Parent-Child Total	**6,236**	**81.61**	**78.58**	**47.86**
Sibling-Sibling, 12-14/15-17	1,291	98.22	85.02	56.52
Sibling-Sibling, 12-17/18-25	1,330	91.05	66.31	35.29
Sibling-Sibling, Other	1,543	69.09	80.21	37.11
Sibling-Sibling Total	**4,164**	**85.13**	**77.18**	**37.48**
Spouse-Spouse with Children	1,631	20.48	74.55	51.97
Spouse-Spouse without Children	1,511	16.68	75.79	54.88
Spouse-Spouse, Children Not Clear	22	4.55	100.00	42.86
Spouse-Spouse Total	**3,164**	**18.55**	**75.13**	**53.32**
Grandparent-Grandchild	193	75.65	60.96	44.68
Overall Total	**13,757**	**68.09**	**77.56**	**49.57**

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2014.

These results suggest that consistency would be increased by having only one member of the household respond to the family income questions, but this may not increase the accuracy of the family income measures. The next logical next step would be to use the pair data to estimate

the measurement error. The family income questions are among the few NSDUH questions that are asked of two individuals who should be giving the same response. For these variables, it is possible that the measurement error is larger than the sampling error.

L.3.3 Promising Potential of Introducing Nonresponse-Specific Probe Items

The final recommended next step for further reducing income item nonresponse is to explore the addition of probe items to the questionnaire.

There is evidence from drug measures in NSDUH that the inclusion of probe items asking for coarser estimates has some impact on nonresponse. Past month users of certain drugs are asked to report the number of days in the past 30 days in which they have used the drug, and the resulting measure is referred to as 30-day frequency of use. If they respond "Don't know" or refuse this question, they receive a probe asking for an estimate of their drug use within the past 30 days from the following categories:

- 1 or 2 days;
- 3 to 5 days;
- 6 to 9 days;
- 10 to 19 days;
- 20 to 29 days; and
- all 30 days.

The 30-day frequency-of-use probe questions have surprisingly high response rates, often greater than 80 percent (Table L.9). If this approach is applied to income, perhaps the probe would include response levels matching the categories of the INCOME5 variable (Table L.3), or even the levels of the POVERTY2 variable, based on the age and the responses to the household roster questions (Section L.3.1.2).

Table L.9 Item Response Rates of 30-Day Frequency Probe Questions: 2014 NSDUH

Drug	Number of Nonrespondents to Original 30-Day Frequency Question	Number of Respondents to Probe	Response Rate to Probe
Cigarettes	103	94	91.26
Snuff	20	16	80.00
Chewing Tobacco	14	11	78.57
Cigars	28	26	92.86
Alcohol	234	196	83.76
Inhalants	14	11	78.57
Marijuana	81	69	85.19
Hallucinogens	4	3	75.00
Cocaine	1	1	100.00
Crack	1	1	100.00
Heroin	1	1	100.00

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2014.

Because it has been established that there are fundamental differences between nonresponse types, the current unfolding bracket structure of the NSDUH income questions may potentially be improved by adding probe items based on types of nonresponse (i.e., "Don't know" and refusals). Further work on the nature of the probe item for nonrespondents refusing to answer the income items and whether reassuring the respondent of anonymity and the importance in the accuracy of the survey estimates would be effective is needed. The next step for examining the effectiveness of such probes in reducing income item nonresponse would require further refinement and field testing.

This page intentionally left blank