2007 NATIONAL SURVEY ON DRUG USE AND HEALTH

IMPUTATION REPORT

Prepared for the 2007 Methodological Resource Book

Contract No. 283-2004-00022
RTI Project No. 0209009.377.007
Deliverable No. 39

Authors:

Kim Ault
Jeremy Aldworth
Kortnee Barnett-Walker
Lisa Carpenter
Elizabeth Copello
Peter Frechtel
Bing Liu
Peilan Martin

Project Director: Thomas G. Virag

Prepared for:

Substance Abuse and Mental Health Services Administration
Rockville, MD 20857

Prepared by:

RTI International
Research Triangle Park, NC 27709

February 2009

2007 NATIONAL SURVEY ON DRUG USE AND HEALTH

IMPUTATION REPORT

Prepared for the 2007 Methodological Resource Book

Contract No. 283-2004-00022
RTI Project No. 0209009.377.007
Deliverable No. 39

Authors:

Kim Ault
Jeremy Aldworth
Kortnee Barnett-Walker
Lisa Carpenter
Elizabeth Copello
Peter Frechtel
Bing Liu
Peilan Martin

Project Director:

Thomas G. Virag

Prepared for:

Substance Abuse and Mental Health Services Administration
Rockville, MD 20857

Prepared by:

RTI International
Research Triangle Park, NC 27709

February 2009

Acknowledgments

This report would not be possible without the guidance and input of staff from the Office of Applied Studies (OAS). In particular, Jonaki Bose, Joe Gfroerer, Art Hughes, and Michael Jones provided useful comments. Special thanks are also due to several current and former RTI International (a trade name of Research Triangle Institute) staff members. Avinash Singh and Ralph Folsom, along with Eric Grau, codeveloped the predictive mean neighborhood (PMN) methodology. Finally, Claudia Clark copyedited, with assistance from Michelle Pattie, and Joyce Clay-Brooks formatted the report in preparation for publication.

Table of Contents

Table of Contents (continued)

Table of Contents (continued)

Table of Contents (continued)

Table of Contents (continued)

Table of Contents (continued)

List of Tables

Page

List of Tables (continued)

Page

1. Introduction

The 2007 National Survey on Drug Use and Health (NSDUH)[1] was implemented using the 50-State multistage cluster design that was introduced in the 1999 survey. The survey was called the National Household Survey on Drug Abuse (NHSDA) until 2002. The 50-State design allows the Substance Abuse and Mental Health Services Administration (SAMHSA) to provide direct estimates for eight large States and to provide estimates based on small area estimation (SAE) methods for the remaining States and the District of Columbia. Other major changes in the 1999 survey from surveys in previous years included the introduction of computer-assisted interviewing (CAI) methods for both screening households and interviewing selected respondents.

The introduction of CAI technology was designed to produce more internally consistent data while still allowing the respondent to answer privately by using the audio computer-assisted self-interviewing (ACASI) method for the more sensitive parts of the interview, such as the drug use modules. Consequently, this ACASI approach allowed the respondent to enter answers to these sensitive questions directly into the computer away from the view of the field interviewer or any other household members. In addition, the questions were displayed on the screen for the respondent to read, and a recorded voice reading of the questions was provided to the respondent via earphones. Several alternatives to the CAI were evaluated in a field test in 1997, and a smaller pretest of a near-final CAI screening and individual questionnaires was conducted in 1998 (for details, see Office of Applied Studies [OAS], 2001; Penne, Lessler, Bieler, & Caspar, 1998).

Although the design of NSDUH has not changed significantly since the introduction of CAI in 1999, important methodological changes were introduced in the 2002 survey that affected the estimates from the survey years that followed. In addition to the name change, each NSDUH respondent has since received an incentive payment of $30, and, finally, information from the 2000 U.S. Decennial Census has been used in the NSDUH weighting procedures. Hence, the 2002 survey year is considered the "baseline year" from which all trends are measured.

This report focuses on the imputation procedures implemented for the 2007 survey. For more details on the editing procedures that were applied to the drug, nicotine dependence, income, and health insurance variables, as well as some of the demographic variables requiring imputation (marital status, education, employment status, and immigrant status), see Kroutil and Handley (2008); Kroutil, Handley, Felts, Bradshaw, and Chien (2008); and Kroutil and Chien (2008). However, the editing procedures for other demographic variables (age, interview date, birth date, gender, race, and Hispanicity), as well as all of the household composition and proxy variables, are discussed in this report. The criteria used for creating household-level and person-level files, along with eligibility and completeness rules, are discussed in Chapter 2, followed by a summary of the implemented imputation procedures in Chapter 3. Chapters 4 and 5 describe the imputation procedures applied to the core and noncore demographic variables, respectively.

[1] This report presents information from the 2007 National Survey on Drug Use and Health (NSDUH), an annual survey of the civilian, noninstitutionalized population of the United States aged 12 or older. Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA).

Chapter 4 also describes editing procedures for age, interview date, birth date, gender, race, and Hispanicity. The drug imputation procedures are discussed in Chapter 6. The imputation procedure for nicotine dependence differed from the procedures used for other variables and is described in Chapter 7. Chapter 8 describes the edits applied to the household roster, the creation and imputation of missing values in the roster-derived household composition variables, and the creation of respondent-level variables with individual roster information. Chapter 9 summarizes the editing and imputation procedures applied to the income variables. Procedures for the imputation of missing values in the health insurance variables are described in Chapter 10. Imputations in the 2007 survey also were conducted in the processing of pair relationships and their accompanying multiplicities for responding pairs, as well as household counts for all households. Although the procedures used for these imputations are similar to those discussed in this report, they are described in a separate report that focuses on the development of household and pair weights (Westlake, Barnett-Walker, Chen, Gordek, & Laufenberg, 2009).

This report also contains nine appendices, including three summaries of the various imputation methodologies used in the current sample. The hot deck is described in Appendix A; the general model used to adjust weights for item nonresponse is discussed in Appendix B; and the methodology developed specifically for NSDUH, the predictive mean neighborhood (PMN) procedure, is described in Appendix C. Respondents had the opportunity to write in responses to some of the drug and demographic questions if they felt the given responses did not apply to them. These responses, called "alpha-specify" or "other-specify" responses, were coded so that the data could be summarized in a meaningful way. A discussion of how this was done for race and Hispanicity is described in Appendix D. (Coding of alpha-specify responses for other variables is summarized by Kroutil and Chien (2008). Models used to assign a single race to multiple-race respondents are described in Appendix E. The covariates in each of the imputation models are listed in Appendix F. In this report, Appendix G has been modified to provide the details of the vector of predicted means used in the multivariate PMN procedure (for the various patterns of missing values) in combination with information on the number of respondents who met likeness constraints (i.e., flexible constraints that governed the similarity between donors and recipients) and logical constraints (i.e., fixed constraints to prevent logical inconsistencies). The quality control measures used in the imputation procedures are summarized in Appendix H. Reasons that interviewers gave for overriding consistency checks in the household roster are presented in Appendix I, along with evaluations of their legitimacy and the resulting actions in the editing of the roster. For the 2007 NSDUH questionnaire specifications for programming, refer to RTI International (2006).

2. Household-Level and Person-Level Files

2.1 Sample Design

The population of eligible respondents for the 2007 National Survey on Drug Use and Health (NSDUH)[2] was all civilian, noninstitutionalized residents of the United States (including the District of Columbia) aged 12 or older. As in other NSDUHs since 1999, this population included residents of noninstitutional group quarters (e.g., homeless shelters, rooming houses, dormitories, and group homes) and civilians residing on military bases. Persons excluded from the 2007 survey included those with no fixed household address (e.g., homeless transients not in shelters), residents of institutional group quarters (e.g., jails and hospitals), children younger than 12, and active military personnel.

The 2007 survey is the third NSDUH in a coordinated 5-year sample design from 2005 through 2009. Although there is no planned overlap with the 1999-2004 samples, a coordinated design facilitated 50 percent overlap in second-stage units (area segments) within each successive 2-year period from 2005 through 2009. For further details, refer to the 2007 NSDUH sample design report (Morton, Martin, Hirsch, & Chromy, 2008).

For the survey, a person was randomly selected for an interview through a four-stage sample selection process. States were first stratified into a total of 900 State sampling (SS) regions. The first stage of selection was census tracts. Within each of these SS regions, a sample of census tracts was selected with probabilities proportionate to a composite size measure and with minimum replacement. Within sampled census tracts, adjacent census blocks were combined to form the second-stage sampling units or area segments. One area segment was selected within each sampled census tract with probability proportional to population size.[3] Once the sample segments were selected, specially trained field staff visited areas and created lists of all eligible dwelling units (DUs) within the sample segment boundaries. These lists served as the frames for the third stage of sample selection. After the DUs were selected within each segment, an interviewer visited each selected DU to obtain a roster of all persons aged 12 or older. This roster information was then used to select zero, one, or two persons from the household at the fourth stage of sample selection.

At the end of the survey year, a household-level file and a person-level file were created to record the information obtained from the sampling processes. The person-level file was later subset into a smaller data file that contained only respondents who were considered "completed" cases—this file was used for analysis. Refer to Section 2.3 for the definition of a completed case. Also, the household-level and person-level files were utilized in the final creation of the person-level and pair-level analysis weights.

---

[2] This report presents information from the 2007 National Survey on Drug Use and Health (NSDUH), an annual survey of the civilian, noninstitutionalized population of the United States aged 12 or older. Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA).

[3] Segments consist of clusters of the geographic aggregated adjacent census blocks. SS regions were formed through geographically partitioning each State into roughly equal-sized regions based on a composite size measure. The 2007 NSDUH sample design report (Morton et al., 2008) contains more information regarding the sample design.

2.2 Dwelling Unit-Level Eligibility and Completeness Criteria

Before proceeding with the fourth stage of sample selection, a set of rules was used to determine whether a DU was eligible to be selected. Examples of ineligible DUs included units defined as "vacant" or "not a primary residency." Eligibility of the DU was recorded in the binary variable DUELIG, where a value of 1 indicated eligibility.

Occasionally, DUs were eligible but failed to complete the screening process. Reasons for not completing the screening process were recorded, including situations such as "language barrier," "refusal," and "denied access." Completeness of the screening process for the DU was recorded in the binary variable DUCOMP, where a value of 1 indicated completeness. For the segments where all the DUs were from denied-access areas, such as gated communities, an adjustment was made in the final household-level file. Although the field interviewers could not obtain an accurate count of DUs from denied-access areas, these DUs were considered eligible. Therefore, DU information from the U.S. Census Bureau for these areas was used in the household-level file.

During the second stage of sampling, it was possible to select a sample segment more than once because samples were selected with replacement. These duplicated segments had different segment IDs (SEGIDs) for each duplicate. However, one SEGID contained all the DU information and the other had none. The number of eligible DUs was split as evenly as possible between the two SEGIDs, and this information was updated in the household-level and person-level files.

2.3 Person-Level Eligibility and Completeness Criteria

During screening, respondents were asked to identify all eligible household members so that only eligible individuals were listed and, therefore, potentially selected. Eligibility was determined according to the criteria provided in Section 2.1. Eligible respondents at the time of screening were recorded in the binary variable PRELIG, which had a value of 1 if the household member was eligible. Respondents who were selected were recorded in the binary variable PRSEL, where 1 indicated a selected individual. It was possible for an individual to be selected, but at the time of the interview, to be determined ineligible. Examples of changes from eligibility to ineligibility included "the selected person turned out not to be a permanent resident in the DU" and "roster error." If this occurred, the value of PRELIG was changed from 1 to 0.

A summary of the number of selected, eligible, and completed dwelling units are shown in Table 2.1. The number of eligible persons also is summarized in Table 2.1.

Table 2.1 NSDUH Household and Person Eligibility and Completed Interview Counts: 2007

	Selected Dwelling Unit	Eligible Dwelling Units	Completed Screenings	Eligible Persons	Selected Persons	Inter-viewed Persons	Completed Cases
CAI	192,092	158,411	141,487	296,964	85,774	68,006	67,870

CAI = computer-assisted interviewing.

To be considered a completed case 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.[4] Unlike the paper-and-pencil interviewing (PAPI) questionnaire in 1999 and surveys prior to 1999, no logical inference could be made from information within a section if the gate question was not answered. This was because the computer-assisted interviewing (CAI) instrument routed respondents out of a section if the gate question was not answered. Completeness of the survey for eligible individuals was recorded in the binary variable PRCOMP, which had a value of 1 if the respondent was a completed case, and 0 if not. For a summary of the number of completed cases in the 2007 survey, see Table 2.1.

2.4 Variables in the Household-Level and Person-Level Files

This section documents some of the important person-level variables that were created for the household-level and person-level files.

Screener-level demographic variables were created from the screener roster information in the household-level and person-level files. XAGE was the screener age, which either could be "continuous" (single-year ages) or categorical. A respondent could choose to give an age category instead of the actual age. The age categories with their accompanying codes were 199 = 12 to 17 years old; 299 = 18 to 25 years old; 399 = 26 to 34 years old; 499 = 35 to 49 years old; and 599 = 50 years old or older. Screener race (XRACE1-XRACE6), screener Hispanicity (XHISP), and screener gender (XSEX) also were produced from the screener roster information. XRACE1 through XRACE6 were indicator variables representing white, black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, and other, respectively. The household-level variable PAIRSEL represented the number of persons within each age group selected from a DU. It was a 20-level variable indicating whether zero, one, or two individuals were selected from the five age groups (12 to 17, 18 to 25, 26 to 34, 35 to 49, and 50 or older) in a given household. (If two persons were selected from the household, this variable indicated the age groups of both pair members.) Similar to PAIRSEL, the household-level variable PAIRRESP had 20 levels, which indicated whether zero, one, or two persons completed the interviews from the five age groups within a household.

As described in the 2007 NSDUH sample design report (Morton et al., 2008), States were partitioned into SS regions, which were further partitioned into clusters of adjacent blocks called "segments." The variable SEGID (segment ID number) was a two-letter State abbreviation followed by a two-digit SS region and a two-digit segment identifier, which uniquely identified each segment. Census region (REGION) was a four-level geographic variable recoded from the respondent's State of residence. The four levels were Northeast, Midwest, South, and West. The population density variable PDEN2 classified respondents according to their living situation, whether it be in a rural or urban area, and, if urban, the size of the urban area. It was used to categorize segments where the respondents lived according to the modified 2000 census data,

[4] For more details on editing rules regarding the drug usage gate questions, please refer to the 2007 NSDUH editing and coding reports (Kroutil, Handley, Felts, Bradshaw, & Chien, 2008; Kroutil & Handley, 2008).

which was adjusted to more recent data from Claritas, Inc.[5] This variable had five levels: segment in core-based statistical area (CBSA)[6] with 1 million or more persons; segment in CBSA with 250,000 to 999,999 persons; segment in CBSA with fewer than 250,000 persons; segment in urban area but not in CBSA; and segment in rural area (not in CBSA and not in urban area). The variable PLACNAME was the census place name associated with each segment. According to the census documentation, this variable was defined as places, for the reporting of decennial census data, which includes census-designated places, consolidated cities, and incorporated places. If duplicate place names existed within the same county, the places were distinguished by their legal description (e.g., "city" or "village"). However, because the variable PLACNAME was used to help the field interviewers locate the segment and was limited by the number of characters printed on the map, identifiers like "city" or "village" have been removed from the place name. The variable STATE represented the Federal Information Processing Standards (FIPS) State codes for the 50 States and the District of Columbia. The variable STATE was created at the sampling stage and did not contain any missing values.

The variables VESTR and VEREP were created to capture the sampling design structure. Each SS region appeared in a different variance estimation stratum (VESTR) every quarter. Two replicates (VEREP) were defined within each variance stratum. Each replicate consisted of four segments, one for each quarter of data collection. Other sampling variables such as DIVISION, SSREGION, GQTYPE, ID, RURORURB, STNAME, STUSAB, and QUARTER[7] also were included in the household-level and person-level files.

[5] Claritas, Inc., is a market research firm headquartered in San Diego, California.

[6] CBSAs, developed in response to standards put forth by the Office of Management and Budget (OMB), are metropolitan and micropolitan areas that were designated using data from the 2000 census. More information about CBSAs can be retrieved from http://www.census.gov/hhes/www/housing/resseg/cbsa.html.

[7] For more details on these sampling variables, please refer to the 2007 NSDUH sample design report (Morton et al., 2008).

3. Overview of Item Imputation Procedures

3.1 Introduction

As with most large-scale sample surveys, the 2007 National Survey on Drug Use and Health (NSDUH)[8] faced the problem of analyzing datasets that contained missing responses for some items. In association with this, there were other issues such as inconsistent or invalid responses and violation of skip patterns. Although the instrument was designed to enforce skip patterns, which has reduced inconsistencies relative to paper-and-pencil interviewing (PAPI), and to perform some consistency checks, inconsistent and invalid responses still occurred. These response errors were an obvious source of bias that was considered in the analysis of NSDUH data (Cox & Cohen, 1985).

Editing to correct erroneous and inconsistent responses and to replace missing values is appropriate when a unique association exists between predictor variables and the variable to be predicted (Cox & Cohen, 1985). For instance, gender often can be inferred from the respondent's relationship to the head of a household (e.g., son, daughter). However, even when good predictor variables are present, a prediction may not be possible for every record having missing or faulty data (e.g., "cousin" does not clarify the gender of a respondent). The remaining faulty and missing data often are replaced with statistically imputed data.

Since the 1999 survey, NSDUH has been conducted using computer-assisted interviewing (CAI) methods, and the CAI instrument has been the only method used since the 2000 survey. To maintain consistency with surveys since 1999, most of the procedures in the 2007 sample were identical to those used in the previous survey years since 1999 (excluding the 1999 PAPI sample). Each year, however, minor modifications were made to the CAI instrument, which subsequently required adjustments to the imputation procedures, and the 2007 survey was no exception. As in the 2006 survey, the procedure developed specifically for the 1999 survey—the predictive mean neighborhood (PMN) procedure—was applied to most of the variables requiring imputation in the 2007 survey. The only imputations that did not incorporate the PMN method were those used for the nicotine dependence variables, which also were handled differently in the 2006 survey. Table 3.1 provides a brief summary of the types of imputation procedures used for each of the variables imputed in the samples in the 1999 to 2007 surveys.

The vast majority of imputation-revised variables were identified by their names, which were given the prefix "IR." The imputation-revised employment status variables EMPSTAT4 and EMPSTATY were exceptions to this rule. Although no missing data were possible for gender, the "IR" prefix for IRSEX was maintained for continuity with past years. Associated indicator variables, which were identified by the prefix "II," were created to tell the user which values were imputed and which ones were not. For some imputation-revised variables, additional imputation indicators were created with the prefix "II2." These indicators gave more details about the source of the imputed or logically assigned value.

[8] This report presents information from the 2007 National Survey on Drug Use and Health (NSDUH), an annual survey of the civilian, noninstitutionalized population of the United States aged 12 or older. Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA).

Table 3.1 Summary of Item Imputation Procedure Used, by Variable and NSDUH Survey Year

Variable	1999[1]	2000	2001	2002/2003	2004-2007
Interview Date	Random[2]	Random	None	None	None
Age	None[3]	None	None	None	None
Birth Date	None	Random	Random	Random	Random
Gender	None	None	None	None	None
Race	USHD[4]	MPMN[5]	MPMN	MPMN	MPMN
Hispanic or Latino-Origin Indicator	USHD	UPMN[6]	UPMN	UPMN	UPMN
Marital Status	USHD	MPMN	MPMN	MPMN	MPMN
Hispanic or Latino-Origin Group	USHD	MPMN	MPMN	MPMN	MPMN
Education	USHD	USHD	MPMN	MPMN	MPMN
Employment Status	USHD	USHD	MPMN	MPMN	MPMN
Immigrant	Not imputed	Not imputed	Not imputed	WSHD[7]	UPMN
Health Insurance	MPMN	MPMN	MPMN	MPMN[8]	MPMN
Lifetime Drug Usage	UPMN	MPMN	MPMN	MPMN	MPMN
Recency and Frequency of Use[9]	MPMN	MPMN	MPMN	MPMN	MPMN
Age at First Use	UPMN	UPMN	UPMN	UPMN	UPMN
Age at First Daily Cigarette Use	UPMN	UPMN	UPMN	UPMN	UPMN
Personal and Family Income (Binary)	MPMN	MPMN	MPMN	MPMN	MPMN
Personal and Family Income (Finer Categories)	UPMN	UPMN	UPMN	UPMN	UPMN
Nicotine Dependence	Not imputed	Not imputed	Regression	Regression	Regression
Household Size (Roster-Derived)	UPMN	UPMN	UPMN	UPMN	UPMN
Other Household Composition (Roster-Derived)	UPMN	UPMN	UPMN	UPMN	UPMN

[1] The 1999 survey year also included a paper-and-pencil interviewing (PAPI) sample. The procedures listed here are from the computer-assisted interviewing (CAI) sample.

[2] "Random" refers to a random assignment within a quarter for the interview date and a random assignment using age and interview date for the birth date.

[3] "None" means that no missing values were encountered after editing, and thus no imputation was necessary. For gender (from the 2002 survey onward) and age, missing values were precluded by design (see Chapter 4).

[4] "USHD" refers to the unweighted sequential hot-deck method of item imputation described in this report (see Appendix A).

[5] "MPMN" refers to the procedure based on the multivariate predictive mean neighborhood model described in this report (see Appendix C).

[6] "UPMN" refers to the procedure based on the univariate predictive mean neighborhood model described in this report (see Appendix C).

[7] "WSHD" refers to the weighted sequential hot-deck method of item imputation described in this report (see Appendix A).

[8] Although MPMN was the method used for health insurance in all years since the 1999 survey, imputation also was applied to more detailed health insurance variables in the surveys from 2002 onward.

[9] "Recency and Frequency of Use" included variables measuring recency of use, 12-month frequency of use, 30-day frequency of use, and binge drinking frequency in past 30 days. "Binge drinking" was defined as having five or more drinks on the same occasion on a given day.

This chapter provides a brief description of PMN, the imputation procedure most used in the 2007 survey, followed by a description of the other procedures used in the survey and a summary of the changes in imputation procedures that occurred between the 2006 and 2007 surveys.

3.2 Overview of the Predictive Mean Neighborhood Imputation Procedure for the NSDUH Sample

PMN was developed specifically for the 1999 survey. A combination of model-assisted imputation and a random nearest neighbor hot-deck imputation, PMN was implemented for nearly all variables requiring imputation in the 2007 survey (exceptions are shown in Table 3.1).

In general, when large nonresponse occurs, limited donor sets can be used for imputation. For the 2007 survey, to adjust for this sparseness of data, predictive mean modeling was used for the imputation of many of the variables (Table 3.1). The models incorporated sampling design weights[9] with a response propensity adjustment computed to make the item respondent weights representative of the entire sample. The item response propensity model is a special case of the generalized exponential model (GEM),[10] which was developed for weighting procedures. The macro for this model was used to apply the item response propensity model and is described in detail in Appendix B. Predicted values (predicted means) were obtained from the models for both item respondents and item nonrespondents. The means of a particular outcome variable were modeled as a function of the predictors (covariates), where these means gave a summary of the effects of covariates on the outcome variable. Unlike the sequential hot-deck imputation method, where values for the covariates were matched through a sorting procedure, the model-based approach used the predicted mean to convert the covariates' effects into a single number. The predicted means, along with other constraints, were used to define the neighborhoods from which donors were randomly selected for the final assignment of imputed values. This assignment was done with either a single predicted mean or several predicted means at once. The method associated with the single predicted mean is called the univariate predictive mean neighborhood (UPMN) method. The multivariate predictive mean neighborhood (MPMN) method is the name associated with the assignment using several predicted means.[11] More details regarding these UPMN and MPMN imputation procedures are provided in Appendix C. For the types of regression models used for each variable that underwent the PMN imputation procedure, see Table 3.2.

[9] In the 2007 survey, the final analysis weights were not available in time for imputation processing of all variables. The person-level sample design weights were therefore adjusted, using a simple ratio adjustment, to account for nonresponse at the household level.

[10] The GEM macro, which was written in SAS/IML® software, was developed at RTI International (a trade name for Research Triangle Institute) for weighting procedures.

[11] Although it was often the case that one predicted mean corresponded to one response variable and a vector of predicted means corresponded to several response variables, it was also common practice to (1) assign several values from a single predicted mean (univariate matching, multivariate assignment) or (2) assign a single response value from a vector of predicted means (multivariate matching, univariate assignment). The latter occurred when the response variable was categorical with three or more levels, resulting in a vector of predicted multinomial probabilities, even though only one cell would have a response assigned to it.

Table 3.2 Regression Models Used for Each Variable Imputed with PMN

Variable	Domain[1]	Type of Regression Model	SAS/SUDAAN Procedure[2,3]
Demographics			
Marital Status	15 years or older	Multinomial Logistic	MULTILOG
Race	All	Multinomial Logistic	MULTILOG
Hispanic or Latino Indicator	All	Binomial Logistic	RLOGIST
Hispanic or Latino Group	Hispanics	Multinomial Logistic	MULTILOG
Education Level	All	Multinomial Logistic	MULTILOG
Employment Status	15 years or older	Multinomial Logistic	MULTILOG
Immigrant Status: Born-in-U.S. Indicator	All	Binomial Logistic	RLOGIST
Immigrant Status: Age of Entry	Not born in U.S.	Simple Linear	REGRESS
Drugs			
Lifetime Drug Use	All	Binomial Logistic	RLOGIST
Recency of Drug Use, "Hierarchical" Drugs	All lifetime users for past year vs. not past year; all past year users for past month vs. not past month	Binomial Logistic	RLOGIST
Recency of Drug Use, Pipes	All lifetime users	Binomial Logistic	RLOGIST
Recency of Drug Use, All Other Drugs	All lifetime users	Multinomial Logistic	MULTILOG
12-Month Frequency of Drug Use	All past year users	Simple Linear	REGRESS
Daily Drug Use over Past 30 Days, Cigarettes, Chewing Tobacco, and Snuff	All past month users	Binomial Logistic	RLOGIST
30-Day Frequency of Drug Use, Cigarettes, Chewing Tobacco, and Snuff	All past month users except those who used daily over the past 30 days	Simple Linear	REGRESS
30-Day Frequency of Drug Use, All Other Drugs	All past month users	Simple Linear	REGRESS
Age at First Drug Use	All lifetime users	Simple Linear	REGRESS

(continued)

Table 3.2 Regression Models Used for Each Variable Imputed with PMN (continued)

Variable	Domain[1]	Type of Regression Model	SAS/SUDAAN Procedure[2,3]
Household Composition			
Total Number of Rostered Persons	All	Poisson	LOGLINK
Total Number of Children Younger than 18	All	Poisson	LOGLINK
Total Number of Persons Aged 65 or Older	All	Poisson	LOGLINK
Indicator of Whether the Respondent Has Family Members in Household	All	Binomial Logistic	RLOGIST
Total Number of Respondent's Family Members in the Household (Excludes Foster Relationships)	All	Poisson	LOGLINK
Total Number of Respondent's Family Members in the Household Younger than 18 (Excludes Foster Relationships)	All	Poisson	LOGLINK
Total Number of Respondent's Family Members in the Household (Includes Foster Relationships)	All	Poisson	LOGLINK
Total Number of Respondent's Family Members in the Household Younger than 18 (Includes Foster Relationships)	All	Poisson	LOGLINK

(continued)

Table 3.2 Regression Models Used for Each Variable Imputed with PMN (continued)

Variable	Domain[1]	Type of Regression Model	SAS/SUDAAN Procedure[2,3]
Income			
Source of Income	All	Binomial Logistic	RLOGIST
Months on Welfare	All respondents who received welfare payments or welfare services in the past year	Simple Linear	REGRESS
Total Income (Binary)	All	Binomial Logistic	RLOGIST
Finer Income Categories	All	Time-to-Event (Survival)	LIFEREG
Health Insurance			
Health Insurance (Old Method)	All	Binomial Logistic	RLOGIST
Health Insurance (Constituent Variables Method)	All	Binomial Logistic	RLOGIST

[1] The set of respondents who were included in the model and for whom predicted means were calculated.
[2] SAS® software is a registered trademark of SAS Institute, Inc. SUDAAN® is a registered trademark of Research Triangle Institute.
[3] See RTI International (2007) for more information on all procedures except PROC LIFEREG. See SAS Institute (1999) for more information on PROC LIFEREG. PROC LIFEREG is the only SAS procedure in this table. All other procedures are SAS-callable SUDAAN procedures.

Wherever necessary and feasible, additional restrictions were placed on the membership in the hot-deck neighborhoods. These constraints were implemented to make imputed values consistent with preexisting, nonmissing values of the item nonrespondent and to make candidate donors as much like the recipients (the item nonrespondents) as possible. The former are called "logical constraints" and could not be loosened. The latter, called "likeness constraints," could be loosened if insufficient donors were available to meet the restriction. If more than one likeness constraint was placed on a neighborhood, the restrictions were loosened in a priority order deemed appropriate for the response variable in question.

In the 2007 survey, the variables related to drug use, household composition, income, and health insurance were highly correlated with age. This, along with the desire to expedite the implementation of procedures, made it necessary to separate the model building and final assignments of imputed values for these variables into three distinct age groups. The drug use variables were imputed within each of three age groups: 12 to 17, 18 to 25, and 26 or older. The household composition (roster-derived), income, and health insurance variables were imputed within the following four age groups: 12 to 17, 18 to 25, 26 to 64, and 65 or older. The age group restriction on the neighborhoods could be considered a likeness constraint. However, the models also were built separately within the age groups, so this restriction was not loosened unless no other options were available. Although the demographic variables did not always show a high

correlation with age, the imputation of missing values in the demographic variables also was performed within age groups. This was done to maintain consistency with how the other variables were imputed, and it facilitated processing. The same three age groups that were used for drugs were also used for demographics. Occasionally, small sample sizes necessitated the aggregation of age groups at the modeling stage. In particular, the models for education level (highest grade completed) were fit within the age groups of 12 to 17 and 18 or older. In the employment status models, the 15-to-17 and 18-to-25 age groups were aggregated. Finally, all age groups were aggregated for the Hispanic or Latino group, marital status, and immigrant age-of-entry models.

For the drug variables, there was originally some interest in requiring the donor to be from the same State as the recipient. However, this could not be implemented because of insufficient pools of donors. A different approach was adopted, which was also applied in the 2007 survey: information about the State of residence of each respondent was incorporated into the modeling and hot-deck steps of the PMN procedure by grouping respondents into three State usage-level categories for each drug, depending on the response variable of interest. Respondents from States with high usage of a given drug were placed into one category, respondents from medium usage States into another, and the remainder into a third category. This categorical "State rank" variable was used as one set of covariates in the imputation models. In addition, as another likeness constraint, eligible donors for each item nonrespondent were restricted to be from States with the same level of usage (the same State rank) as the item nonrespondent. A similar State-rank variable was used in the income imputations, but only in the modeling step (not in the hot-deck step). The three State-rank categories were defined in terms of the income level of the States: high-income States, middle-income States, and low-income States. No State-rank variables were created for any other variables.

3.3 Other Imputation Procedures Used in the 2007 Survey

Each respondent had a valid age (AGE) and interview date (INTDATE). No imputation was required for these variables. However, sometimes the availability of several alternative values required rules, as outlined in Chapter 4, for selecting the most appropriate values. Missing values for birth date (BRTHDATE) were imputed using a random imputation within the bounds determined by AGE and INTDATE.

The exact date of first drug use was imputed using a random assignment within an interval of possible dates of first use. Each day in the interval was equally likely to be selected. The interval could be up to a year in length. The date was imputed for almost all lifetime users of each drug because no respondents were asked for an exact date of first use (though many were asked for the year and month of first use). Chapter 6 provides more details on the algorithm.

The imputation-revised versions of the nicotine dependence variables differed from other imputation-revised variables in three ways: (1) as stated earlier in this chapter, PMN was not used to impute missing values; (2) imputed values did not resemble preexisting nonmissing values; and (3) not all missing values were imputed. Weighted least squares regressions were used to obtain continuous predicted means, which were used directly as imputed values. Whereas the nonimputed values were limited to integer values between 1 and 5, imputed values fell anywhere on the continuous scale. Imputations were performed only if the respondent answered

at least 16 of the 17 nicotine dependence questions. If the respondent was eligible to answer the nicotine dependence questions, but answered 15 or fewer of them, no attempt was made to replace missing values by imputed values. For these respondents, in the imputation-revised versions of the variables, missing values were still represented as missing values.

3.4 Changes in Procedures from the 2006 Survey to the 2007 Survey

Overall, the changes implemented between the 2006 and 2007 surveys were minor, both in number and in type. One of these changes was the result of modifications to the CAI instrument. Other changes, however, were procedure enhancements involving both editing and imputation.

3.4.1 Differences between Instruments in the 2006 and 2007 Surveys Affecting Variables Requiring Imputation

In the survey years from 1999 to 2005, all respondents received the same set of income questions. However, in the 2006 survey, about 5 percent of respondents received a reduced set of income questions. In the original (full) set of income questions, 10 source-of-income variables were covered as follows: Social Security, Supplemental Security Income, welfare cash assistance, welfare noncash assistance, wages, food stamps, child support, interest/investment income, other income, and the number of months receiving welfare. Except in households with no other family members, separate questions were asked to ascertain personal and other family-level responses, which were then combined to create family-level variables. In the reduced set of income variables, only 6 of the 10 source-of-income variables were covered; questions covering Social Security, child support, interest/investment income, and other income were omitted. In addition, for the 6 remaining source-of-income variables, separate questions to ascertain personal and other family-level responses were no longer asked; all questions were asked at the family level only.

In the 2007 survey, for the subsample of respondents who received the reduced set of income questions, the family-level Social Security question was asked. Therefore, in 2007, there were 7 source-of-income variables included in the reduced set of questions instead of 6. The set of logical constraints that was used in the hot-deck program for the respondents who received the reduced set of questions was modified accordingly. Additional details are provided in Chapter 9.

3.4.2 Improvements in Imputation Procedures from the 2006 Survey to the 2007 Survey

Although there were no major changes to imputation procedures implemented between the 2006 and 2007 surveys, there were several minor changes. In survey year 2007, some respondents entered variants of "Spanish" as a write-in response to the main race question, did not respond affirmatively to the Hispanic origin question, and did not select one of the listed racial categories in the race questions. For these respondents, the donor in the hot-deck step for race was required to have mentioned "Spanish" (either as a write-in response or as a listed response to the Hispanic group question) but did not have to be Hispanic. This response pattern did not appear in earlier years.

In addition, a new pattern of item nonresponse occurred for smokeless tobacco recency and frequency in 2007. The hot-deck program was modified to construct the predictive mean vector and to apply constraints for this new pattern. The corresponding documentation was added to Appendix G.

4. Core Demographics

4.1 Introduction

Several demographic characteristics were needed for all respondents in the 2007 National Survey on Drug Use and Health (NSDUH).[12] Core demographic data were collected on both the screener[13] and the questionnaire. Missing values in screener and questionnaire demographic variables were imputed separately for the set of all eligible rostered individuals and for the set of completed respondents (i.e., screener data and questionnaire data were edited and imputed independently).[14] As an initial step, prior to any processing of the data, completed cases were identified. Only these completed cases were included in the subsequent editing, imputation, and analysis of questionnaire data.

The core demographics in the 2007 survey discussed in this report are age, birth date, gender, race, Hispanicity, marital status, and education level (highest grade completed). The only noncore demographic variables imputed were the immigrant variables and employment status. Although the interview date was not classified as a core demographic variable, its editing procedures also are included in this chapter.

Prior to imputation, logical editing was performed on all of these variables. Through the editing process, some missing values were replaced with coded information from the "other-specify" questionnaire responses, thus reducing the amount of statistical imputation required. Noncore information was not used to edit core variables.

After editing, the variables were handled using one of three procedures. For interview date, age, and gender, no statistical imputation was required, because no values were missing after editing. For birth date, 62 respondents had missing values, which were imputed using a random assignment from all possible birth dates that were consistent with the interview date and the age. The missing values in the marital status, race, Hispanicity, and education level variables were imputed using the predictive mean neighborhood (PMN) method. This procedure is described in Appendix C. Missing values for the noncore demographic variables, which are discussed in the next chapter, also were imputed using the PMN method.

This chapter describes the editing and imputation procedures used to create the final core demographic variables and interview date for all respondents who were considered "completed cases."[15] A summary of item nonresponse is included for each variable described here.

---

[12] This report presents information from the 2007 National Survey on Drug Use and Health (NSDUH), an annual survey of the civilian, noninstitutionalized population of the United States aged 12 or older. Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA).

[13] The "screener" refers to the information about household members obtained at the second stage of sampling in NSDUH, the selection of dwelling units within segments (groups of U.S. Census Bureau blocks). The screener information was obtained independently of the questionnaire information.

[14] See the weighting report for the 2007 survey (Chen et al., 2009) for a description of the imputation procedures used for screener demographics for the set of all eligible rostered individuals.

[15] See Chapter 2 for a definition of a "completed case."

4.2 Editing of Demographic Variables

The editing procedures for some of the core demographic variables (marital status and education level) are described by Kroutil and Chien (2008) and are briefly summarized in this chapter. However, the editing procedures for other core demographic variables (age, birth date, gender, race, and Hispanicity) and the interview date are described only in this chapter. For interview date, age, and gender, no imputation was required and the edited variable was considered the final variable to be used for analysis. There were missing values for birth date, but these values were imputed using a random number, a process that is described in this section. The variable for birth date also was considered "final." However, the edited variables for marital status, race, Hispanicity, and education level were intermediate variables because a final imputation, as described in Section 4.3, was used to allocate values when data were missing. When a respondent was known to belong to one of several races based on a write-in response[16] indicating a country of origin, randomly generated numbers were used to allocate the respondent to a particular race. In these cases, the "edited variable" included these imputed values.

Although editing was performed separately for screener and questionnaire data, there were some extraordinary circumstances where the screener respondent age was compared with questionnaire age for editing purposes. Segment-level screener information was used in imputation models, which are described in Section 4.3.

4.2.1 Interview Date (INTDATE)

After each questionnaire module was complete, the time was automatically saved by the computer-assisted interviewing (CAI) instrument. The time for each module was called a "time stamp," and the date portion of the time stamp was called a "date stamp." This information was used to help determine the value for the interview date.

The specific date stamps used to determine the edited interview date (INTDATE) were indicated in the variable EIIDATE. If the label for EIIDATE indicated that the interview date was set to a particular date stamp, then that date stamp was made to be consistent with all subsequent date stamps, unless otherwise indicated. If the interview date was set to the end-of-interview date stamp, then that date stamp was made to be consistent with all preceding date stamps, except for those indicated.

In some cases, the respondent's birthday occurred between the beginning and the end of the interview. In these cases, the interview date was set to the end-of-interview date stamp, which was consistent with the first date stamp after the respondent's birthday. (This date stamp was indicated in the CAI.)

A date stamp was not used to set the interview date if any of the following conditions were true:

[16] In the section of the questionnaire where the respondent (through the interviewer) selects a race, a respondent can reject the options given and direct the interviewer to provide an alternative answer, also known as a "write-in response." See Section 4.2.6 for details.

1. The date stamp was more than 14 days outside the quarter in which the interview was supposed to take place.
2. The date stamp was later in time than a subsequent date stamp.
3. The date stamp occurred before a birthday, which in turn occurred before the end of the interview.

For a summary of the editing of interview dates, see Table 4.1. As stated above, this information was recorded in the editing indicator variable EIIDATE.

Table 4.1 Interview Date Editing Summary

Value of EIIDATE	Assignment of Interview Date	Frequency	Percent
1	Begin date stamp (all date stamps exist)	67,836	99.95
1.01	Begin date stamp (all date stamps exist except last one)	6	0.01
1.02	Begin date stamp (all date stamps exist through sedatives)	19	0.03
1.03	Begin date stamp (all date stamps exist through stimulants)	1	0.00
1.04	Begin date stamp (all date stamps exist through tranquilizer)	1	0.00
1.05	Begin date stamp (all date stamps exist through pain relievers)	1	0.00
1.06	Begin date stamp (all date stamps exist through inhalants)	1	0.00
3	Tutorial date stamp (begin date stamp is outside quarter)	1	0.00
8	End date stamp (tutorial date stamp is the first occurrence of new date stamp; birthday is between the begin and end date stamp)	1	0.00
8.13	End date stamp (tranquilizers date stamp is the first occurrence of new date stamp; birthday is between the begin and end date stamp)	1	0.00
8.16	End date stamp (noncore demographics date stamp is the first occurrence of new date stamp; birthday is between the begin and end date stamp)	2	0.00

4.2.2 Age

4.2.2.1 Final Edited Age (AGE)

After a respondent had entered his or her birth date in the first part of the questionnaire, he or she had multiple opportunities to change his or her age in response to consistency checks throughout the questionnaire. Therefore, it was possible for the age recorded by the respondent at the beginning of the questionnaire (CALCAGE) to be different from the age at the end of the questionnaire (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 (INTDATE) and the raw birth date (AGE1), the age corresponding to the "self" in the questionnaire household roster (if it existed), and the pre-interview screener age. In most cases, when determining the final edited continuous age, priority was given to CALCAGE, NEWAGE, and the age calculated from AGE1 and INTDATE. There were occasions, however, where the

age corresponding to the "self" in the household roster was used even if it did not agree with CALCAGE and NEWAGE. If the final age (AGE) did not agree with the originally entered raw birth date (AGE1), the birth date also was edited. An intermediate value for age was determined in the following manner:

Intermediate value for age =

- NEWAGE, if nonmissing and exactly equal to CALCAGE, where TBEG_TUT (the interview date time stamp at the beginning of the tutorial) = INTDATE (the edited interview date) (age indicator = 1); else
- NEWAGE, if nonmissing, TBEG_TUT and INTDATE were not equal, but NEWAGE was exactly equal to CALCAGE (adjusted by Blaise[17] to a changed interview date if the interview date was changed within the questionnaire), and the respondent's birthday did not fall between the dates corresponding to TBEG_TUT and INTDATE (age indicator = 1); else
- NEWAGE, if nonmissing, TBEG_TUT and INTDATE were not equal, the respondent's birthday fell between the dates corresponding to TBEG_TUT and INTDATE, the given value of CALCAGE agreed with what it should be based on INTDATE and the given birth date (i.e., EIIDATE not equal to 6), and NEWAGE and CALCAGE were exactly equal (age indicator = 1); else
- age calculated from INTDATE and the reported birth date, if the birth date was nonmissing, TBEG_TUT and INTDATE were not equal, the respondent's birthday fell between the dates corresponding to TBEG_TUT and INTDATE, and the given value of CALCAGE did not agree with what it should be based on INTDATE and the given birth date (EIIDATE = 6), where the newly calculated age based on INTDATE was exactly equal to the screener age and/or the roster age (if it existed) (age indicator = 2); else
- NEWAGE, if NEWAGE differed from CALCAGE and NEWAGE = screener age and NEWAGE = roster age (if it existed), and the interview date at the beginning of the interview (TBEGINTR) was within the appropriate quarter (age indicator = 3); else
- CALCAGE, if CALCAGE differed from NEWAGE and CALCAGE = screener age and CALCAGE = roster age (if it existed), and the interview date at the beginning of the interview (TBEGINTR) was within the appropriate quarter (age indicator = 4); else
- age calculated from reported birth date and INTDATE, if EIIDATE = 5 and NEWAGE = CALCAGE (but neither was equal to the correct age) (age indicator = 5); else
- NEWAGE, if NEWAGE differed from CALCAGE, but NEWAGE = roster age, provided roster age existed (age indicator = 6); else

[17] Blaise is the computer program within the CAI instrument that was used to direct the respondent and interviewer through the questionnaire.

- CALCAGE, if CALCAGE differed from NEWAGE, but CALCAGE = roster age, provided roster age existed (age indicator = 7); else
- NEWAGE, if NEWAGE differed from age calculated from reported birth date and INTDATE, but NEWAGE = CALCAGE, screener age, and roster age (if it existed) (age indicator = 8); else
- CALCAGE, if CALCAGE differed from NEWAGE, but CALCAGE = age calculated from INTDATE and the reported birth date, and CALCAGE was within 1 year of screener age and roster age (age indicator = 9).

After the rules above were applied, this intermediate age value was compared with the age corresponding to the "self" in the household roster. In most cases, the final edited value for the age variable (AGE) was set to this intermediate age value. There were exceptions, however, as detailed in the following paragraph.

By the time that the interviewer reached the roster part of the questionnaire, he or she had multiple opportunities to change the respondent's age stored in the computer in response to consistency checks involving age. This value of age was called CURNTAGE by the Blaise program. One of the consistency checks in the questionnaire household roster was to verify the value of the respondent's own entry for age in the household roster (the "self" entry) against the value of CURNTAGE. If the self age differed from CURNTAGE, then the interviewer could either change the respondent's age entered in the roster or override the consistency check and provide an explanation as to why the roster age did not match CURNTAGE. If the consistency check for age was overridden, then the value for age corresponding to the self may not match the intermediate age value described above. However, if the explanations given for overriding the age consistency check were sufficient and other evidence pointed to the veracity of the roster age, and if the difference between CURNTAGE and the roster age for self was less than 2 years, then AGE was set to the roster age, even if it disagreed with both NEWAGE and CALCAGE. In particular, all of the following conditions had to be met for this to occur:

1. The interviewer specifically indicated that the roster age was the correct one.
2. The pre-interview screener age matched the roster age.
3. The other household member's roster supported the roster age value, if another member of the household completed the interview.

Table 4.2 provides a summary of the editing procedures used to create AGE for the 2007 survey. This information was recorded in the editing indicator variable EIAGE.

Table 4.2 Age Editing Summary

Value of EIAGE	Assignment of Age	Frequency	Percent
1	NEWAGE (consistent with CALCAGE and INTDATE—AGE1)	67,863	99.99
4	CALCAGE (consistent with screener age)	1	0.00
6	NEWAGE (consistent with roster age)	2	0.00
10	Roster age; disagrees with NEWAGE and CALCAGE by at least 2 years, but consistent with screener age, and interviewer specifically indicates that roster age was correct and NEWAGE and CALCAGE were incorrect	4	0.01

4.2.2.2 Recoded Age Categorical Variables (CATAGE, CATAG2, CATAG3)

Three age category variables were created from the final age: CATAGE with four levels (12 to 17, 18 to 25, 26 to 34, and 35 or older), CATAG2 with three levels (12 to 17, 18 to 25, and 26 or older), and CATAG3 with five levels (12 to 17, 18 to 25, 26 to 34, 35 to 49, and 50 or older). These variables were used instead of the continuous age variables in some subsequent imputations and analysis.

4.2.3 Birth Date (BRTHDATE)

To continue with the questionnaire, respondents were required to provide their date of birth and/or current age at the beginning of the interview. Each complete case respondent possessed a current age, although a number of cases had missing birth dates. If the birth date was nonmissing but was inconsistent with AGE and INTDATE (either in the raw data or as a result of editing age and/or interview date), then the reported birth month and day were preserved, and the birth year was adjusted according to the interview date and age.

In cases with missing birth dates, a birth date was randomly selected from all possible birth dates, given the final age and interview date. Each date in this period (365 or 366 days, depending on whether the period includes February 29 in a leap year) had an equal probability of selection.

See Table 4.3 for a summary of the birth date editing. This information was recorded in the editing indicator variable EIBDATE.

Table 4.3 Birth Date Editing Summary

Value of EIBDATE	Assignment of Birth Date	Frequency	Percent
1	Reported birth date	67,802	99.90
2	Reported birthday, year from AGE and INTDATE	6	0.01
3	Randomly assigned using AGE and INTDATE	62	0.09

4.2.4 Gender (IRSEX)

As with previous surveys since 2002, it was mandatory in the 2007 survey that an interviewer enters the respondent's gender in QD01. As a result, it was not possible to have missing values for this question. To maintain continuity with previous surveys (1999-2001), the variable name IRSEX was used to describe gender in the 2007 survey. However, it was not necessary to create an imputation indicator, because IRSEX and QD01 were exactly equivalent.

4.2.5 Marital Status (MARITAL, EDMARIT)

In the 2007 questionnaire, a single core question (QD07) asked about the respondent's marital status, among respondents aged 15 or older. The exact phrasing of the question follows:

QD07: Are you now married, widowed, divorced or separated, or have you never married?

1 MARRIED
2 WIDOWED
3 DIVORCED OR SEPARATED
4 HAVE NEVER MARRIED

The creation of the edited variable derived from QD07, MARITAL, is described in Kroutil and Chien (2008). The base variable for creating an imputation-revised version of marital status was called EDMARIT. This variable was equivalent to MARITAL, with the exception that all legitimate skips were collapsed into a single legitimate skip code (99), and missing values were set to the SAS®[18] missing code (.) so that they could be properly handled by the modeling programs.

4.2.6 Race, Hispanic/Latino Indicator, Hispanic/Latino Group

4.2.6.1 Introduction

In the 2007 questionnaire, two core questions focused on the respondent's race (QD05 and QD05ASIA) and two focused on the respondent's ethnicity[19] (QD03 and QD04). For those questions with multiple categories (QD04, QD05, and QD05ASIA), the respondent had the opportunity to select more than one category. Two more Hispanic/Latino group categories were added to QD04 since the 2004 survey: Dominican (from Dominican Republic) and Spanish (from Spain). These new categories were added to the survey because of the large number of other-specify responses in previous NSDUHs that mapped to these categories. The questions are presented below.

QD03: Are you of Hispanic, Latino, or Spanish origin or descent?

1 YES
2 NO

[18] SAS® software is a registered trademark of SAS Institute, Inc.

[19] The questions about ethnicity were limited to determining whether a respondent was Hispanic/Latino or not, and the specific Hispanic/Latino group to which a Hispanic/Latino respondent belonged.

QD04: (Asked only if QD03 = 1) Which of these Hispanic, Latino, or Spanish groups best describes you?

1 MEXICAN / MEXICAN AMERICAN / MEXICANO / CHICANO
2 PUERTO RICAN
3 CENTRAL OR SOUTH AMERICAN
4 CUBAN / CUBAN AMERICAN
5 DOMINICAN (FROM DOMINICAN REPUBLIC)
6 SPANISH (FROM SPAIN)
7 OTHER (SPECIFY)

QD05: Which of these groups describes you?

1 WHITE
2 BLACK / AFRICAN AMERICAN
3 AMERICAN INDIAN OR ALASKA NATIVE (AMERICAN INDIAN INCLUDES NORTH AMERICAN, CENTRAL AMERICAN, AND SOUTH AMERICAN INDIANS)
4 NATIVE HAWAIIAN
5 OTHER PACIFIC ISLANDER
6 ASIAN (FOR EXAMPLE: ASIAN INDIAN, CHINESE, FILIPINO, JAPANESE, KOREAN, AND VIETNAMESE)
7 OTHER (SPECIFY)

QD05ASIA: (Asked only if level 6 of QD05 was selected) Which of these groups describes you?

1 ASIAN INDIAN
2 CHINESE
3 FILIPINO
4 JAPANESE
5 KOREAN
6 VIETNAMESE
7 OTHER (SPECIFY)

As stated in the guidelines from the Office of Management and Budget (OMB),[20] "Hispanic/Latino" was considered an ethnicity, not a race. However, when given the opportunity to enter a race and when the given choices did not apply, many respondents entered "Hispanic" or some Hispanic/Latino group, resulting in a considerable amount of missing data for the race question. The final drug use tables were cross-classified with a variable that combined race and ethnicity. Nevertheless, separate variables were initially created for race and ethnicity, and the race/ethnicity variables used in the tables were derived from these separate variables.

[20] In October 1997, the OMB released a notice, "Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity" (OMB, 1997) that provides new standards for maintaining, collecting, and presenting Federal data on race and ethnicity.

As a result of the confusion between Hispanicity and race, Hispanicity was used in the editing of race, and vice versa. In the process of editing race, the other-specify response to the Hispanic/Latino group question (QD04) was consulted (if it existed) if no race information was identified in QD05 or QD05ASIA. Similarly, in the process of editing the Hispanic/Latino group, the other-specify responses to the race questions (QD05 and QD05ASIA) were consulted (if they existed) if no Hispanic/Latino group information was identified in QD04. Because of the interdependence of race and Hispanicity, the editing of these variables will be discussed in a single section (Section 4.2.6.2).

The procedures used to edit the race and Hispanicity variables in the surveys since 2003 differed in several ways from the procedures used in previous surveys. One of the major differences was in the handling of race for multiple-race respondents. The procedural changes were triggered by the elimination of the QD06 question, which appeared in the survey from 1999 to 2002. QD06 asked respondents who selected more than one racial category from QD05 and QD05ASIA combined to choose the race with which they identified the most. Without this question, it was impossible to determine (directly) the single race that a given multiple-race respondent would most closely identify himself or herself.

4.2.6.2 Categories Used in Race and Hispanicity Variables

4.2.6.2.1 Racial Categories

For editing purposes, the 5 specific categories in QD05 (white, black/African American, American Indian/Alaska Native, Native Hawaiian, and Other Pacific Islander) and the 6 specific categories in QD05ASIA (Asian Indian, Chinese, Filipino, Japanese, Korean, and Vietnamese) were combined to produce 11 racial categories. Two other categories also were created: "Other Asian" and "Asian nonspecific." Respondents could choose almost any subset of these categories. The only impossible subsets were those that included "Asian nonspecific" in combination with one or more specific Asian categories. Combining the information from QD05 and QD05ASIA, as well as QD04 when necessary, allowed the creation of all the edited and imputation-revised race variables.

Two types of race variables were created after editing: one that included levels for more than one race, and another that did not. In addition to the 13 edited single-race categories given above, respondents also could identify themselves as belonging to a combination of racial categories. For some of the variables that accounted for multiple-race responses, these responses were recorded in three levels: more than one race, more than one Asian race, and Native Hawaiian/Other Pacific Islander. Other variables were created that recorded the specific combination of races that was entered. For the variables that did not account for multiple-race responses, multiple-race respondents were allocated to one of the races they selected. This was easily done in survey years prior to 2003 because the response to QD06 (when nonmissing) provided this value. However, with the absence of QD06 since the 2003 survey, a single race was selected from the multiple races chosen in some other manner. The method used for doing this is discussed in Appendix E. A discussion of why this type of variable was needed is given in Section 4.2.6.4.3.

4.2.6.2.2 Hispanic/Latino Categories

With the addition of two Hispanic/Latino categories since the 2004 survey, respondents were given the choice of seven categories in QD04 (Mexican/Mexican American/Mexicano/Chicano, Puerto Rican, Central or South American, Cuban/Cuban American, Dominican (from Dominican Republic), Spanish (from Spain), or some other Hispanic/Latino group),[21] and they could choose more than one category. As with QD05, interviewers could manually enter the alternative to the choices given, which would be either coded to some subset of the existing seven categories or set to missing. The other-specify responses to QD05 and/or QD05ASIA, if nonmissing, were consulted if no Hispanic/Latino origin group information was available from QD04. The final imputation-revised Hispanic/Latino group variable, IRHOGRP4, included all seven Hispanic/Latino group levels and a legitimate skip code (99) for respondents who were not Hispanic/Latino.

4.2.6.3 Classification of Other-Specify Codes

All other-specify responses from QD04, QD05, and QD05ASIA were assigned both a race code and a Hispanic/Latino code. Each of these codes was mapped to at least one of the categories described in Section 4.2.6.2 and this section, or to some other code that was informative in the final imputation described in Section 4.3. A summary of categories of other-specify codes and how they were handled is given in the following sections. Appendix D provides the individual other-specify codes and more details about how they were handled.

4.2.6.3.1 Mapping of Race Other-Specify Codes

In general, race codes were of four types: (1) directly mapped codes; (2) indirectly mapped codes (these required a quick imputation using a randomly generated number); (3) informative codes for formal imputation procedures; and (4) noninformative codes. The edits following either directly or indirectly mapped codes resulted in values that were considered "final." The other two types of codes resulted in incomplete values requiring imputation, and were either informative or noninformative for the formal imputation procedures as described in Section 4.3.

Directly Mapped Codes. The directly mapped codes were mapped to one or more of the categories given in the questionnaire (see Section 4.2.6.2). There were two types of directly mapped codes: (1) racial category codes; and (2) geographic category codes. Racial category codes were exactly equivalent to one or more categories in QD05 or QD05ASIA, and were mapped directly to those categories regardless of whether the write-in response was in QD05 or QD05ASIA. (Respondents were still considered at least part Asian even if the write-in response in QD05ASIA was non-Asian. The racial makeup of a respondent who entered a non-Asian racial category in QD05ASIA was determined on a case-by-case basis.) For example, a response such as "Han" mapped directly to a category in QD05ASIA ("Chinese"), and a response such as "mestizo" mapped directly to two categories in QD05, "white" and "Native American." Geographic category codes corresponded to a country, where census data indicated a racially homogeneous society. For example, an entry of "Polish" in QD05 mapped to white because the

[21] When listing the six Hispanic/Latino defined categories in QD04, they shall henceforth be listed in this chapter as Mexican, Puerto Rican, Central or South American, Cuban, Dominican, and Spanish.

Polish census data indicated nearly all Poles were white. On the other hand, an entry of "Polish" in the QD05ASIA other-specify mapped to "Other Asian." Geographic category codes also included ethnic groups where the racial identification was not immediately obvious. For example, a response of "Arab" would be automatically mapped to "white" if the response was a write-in response for QD05. However, as with the "Polish" entry, if the "Arab" response was a write-in response in QD05ASIA, the respondent was considered "Other Asian."

Indirectly Mapped Codes. Codes that were indirectly mapped also corresponded to countries where census data were used, but for indirect mapping the countries were racially heterogeneous. A racial category was chosen by generating a random number and allocating the race based on a comparison of the random number with the proportions of races in the country's census. For example, an entry of "Bolivian" would have a 55 percent chance of being allocated to the American Indian/Alaska Native category, because the latest Bolivian census indicated 55 percent of Bolivians were American Indian/Alaska Native. For countries where the census indicated a small proportion of some indistinct category such as "other" and the randomly generated number indicated an allocation to this proportion, the final race was left to imputation (appropriately constrained based upon the indistinct response). If two or three heterogeneous countries were entered in the other-specify response (e.g., Bolivian and Peruvian), the final race was allocated using the following procedure: (1) randomly assign races based on the proportions for each country mentioned; and (2) combine the results. Exceptions to these rules occurred with the categories Mexicans, Puerto Ricans, Cubans, Dominicans, Central or South Americans (no country listed), and Spanish, which were given codes described under the next heading, with a final value determined using the formal imputation procedures described in Section 4.3. Starting with the 2006 imputation process, the handling of indirectly mapped codes obtained from QD05ASIA has been simplified. In prior survey years, this type of write-in response was mapped to a race through country census information; since the 2006 NSDUH, all census-based write-in responses to the Asian race question were mapped directly to the "Other Asian" racial category.

Informative Codes for Formal Imputation Procedures. Some other-specify responses did not lead to definitive information about the respondent's race. However, the responses were used to limit the final imputation, which is described in Section 4.3. For example, a response of "mixed" resulted in an imputation among donors with two or more races, and a response of "brown" resulted in an imputation among donors who were not single-race white.

Noninformative Codes. Finally, a noninformative response (e.g., American) that was not accompanied by a response to one of the given (non-other-specify) categories resulted in an unrestricted imputation.

4.2.6.3.2 Subsequent Editing of Race Other-Specify Codes

Subsequent to the initial mapping of the other-specify codes, edits were sometimes implemented that revised or clarified the initial mapping before final races were allocated. These edits were necessary if multiple sources of information, including other-specify responses, provided conflicting or confusing information. These edits were implemented when (1) the final mapping depended upon the source question; (2) the responses were given to both the other-specify and non-other-specify categories of QD05 or QD05ASIA; or (3) the different other-specify responses were present in at least two of QD04, QD05, and QD05ASIA. In some cases, it was necessary to individually examine the responses to determine the appropriate mapping.

Final Mapping Depends upon the Source Question. In some cases, the final mapped value depended upon whether the other-specify code was in QD04, QD05, or QD05ASIA. An example from directly mapped codes is "Indian." This response would be mapped to "American Indian/Alaska Native" if the other-specify response was in QD05, but it would be mapped to "Asian Indian" if the other-specify response was in QD05ASIA. Indirectly mapped codes also could depend upon the source question. The census data from many countries included Asian categories. If the other-specify response was in QD05ASIA, the random imputation to a census category was limited to the Asian categories. Other-specify responses that were not specifically Asian sometimes occurred in the other-specify category of QD05ASIA. These were carefully examined, but the "Asian" part of the response was always preserved.

Responses Given to Both Other-Specify and Non-Other-Specify Categories. If other-specify responses to QD05 or QD05ASIA accompanied responses to the given (non-other-specify) categories of QD05 and QD05ASIA, it was necessary to reconcile these responses. In some cases, the combination of responses mapped to one of the multiple racial categories. For example, if a respondent selected "black/African American" in QD05 and wrote in "black and American Indian," then the respondent would be assigned both racial categories "black/African American" and "American Indian/Alaska Native." However, there were instances when the other-specify response was ignored because of responses to the non-other-specify categories. In particular, the other-specify response was always ignored if a non-other-specify category was selected, and the other-specify response was a geographic category code.[22] For example, if the interviewer selected the category for "black/African American" for the respondent and also wrote in "Polish," it was assumed that the respondent was a black Pole and, for racial identification purposes, was considered single-race black/African American. This was true even though the Polish census did not identify significant numbers of nonwhite persons in the Polish population.

Different Other-Specify Responses Present in at Least Two of QD04, QD05, and QD05ASIA. In some instances, it was necessary to reconcile the other-specify responses to QD04, QD05, and QD05ASIA. In these cases, the responses were examined on an individual basis, and sometimes a new code was assigned that more accurately reflected the situation.

4.2.6.3.3 Mapping of Hispanic/Latino Other-Specify Codes

Certain Hispanic/Latino codes were considered "Definitely Hispanic." If any of these appeared in QD05 or QD05ASIA, the respondent was considered Hispanic/Latino regardless of the response to QD03. Examples included "Hispanic" and "Dominicano" (Spanish for "Dominican"). There was also a code to handle respondents who were definitely not Hispanic/Latino. If this code appeared in QD04, QD05, or QD05ASIA, the respondent was considered non-Hispanic/Latino regardless of the response to QD03. All other Hispanic/Latino codes either mapped directly to one or more of the seven Hispanic/Latino group categories or provided no new information (e.g., Hispanic).

[22] Actually, this "edit" was not "subsequent" to the initial mapping. Instead, the initial mapping was ignored under the circumstances described.

4.2.6.4 Edited Variables, Race

4.2.6.4.1 Individual Racial Categories (EDQD051-EDQD0513)

Edited variables were created that correspond to the 13 racial categories described in Section 4.2.6.2.1. These variables were called EDQD05*xx*, where *xx* represented a number between 1 and 13, corresponding to each of the 13 categories.

EDQD05*xx* =

- 1, if the level *xx* was selected by the respondent in QD05 or QD05ASIA; else
- 2, if the level *xx* was indicated by a directly mapped code in QD05 or D05ASIA; else
- 3, if no EDQD05*xx* variables had values of 1 or 2, and the level *xx* was indicated by a directly mapped code in QD04; else
- 4, if (a) no EDQD05*xx* variables had values of 1, 2, or 3, and (b) the level *xx* was indicated by an indirectly mapped code in QD04, QD05, and/or QD05ASIA; else
- missing.

EDQD0513 (Asian nonspecific) was a little different from the others. In particular, there was no specific level of QD05 or QD05ASIA that corresponded to it. It was used mainly to preserve a response of "Asian" to QD05, even if the respondent selected nothing in QD05ASIA. The value of EDQD0513 was set to 1 if the respondent selected "Asian" in QD05 but mentioned nothing that mapped to a specific Asian category in QD05ASIA. It also could have values of 2, 3, or 4, depending on the other-specify codes.[23]

4.2.6.4.2 Broad Categories of Race (EDRACE)

The EDRACE variable indicates which of four broad racial categories (white, black/African American, American Indian/Alaska Native, Asian/Other Pacific Islander) were identified in QD04, QD05, and QD05ASIA, and it also has levels to indicate how the imputation should be restricted based on the race of the donor. The first three broad racial categories corresponded to EDQD051, EDQD052, and EDQD053, respectively. "Asian/Other Pacific Islander" was considered to have been identified if any of EDQD054 through EDQD0513 was nonmissing. EDRACE was created using the following rules, under five possible scenarios:

Scenario 1: If only one broad racial category was identified in QD04, QD05, and/or QD05ASIA, EDRACE =

- 1 (white only), if EDQD051 was nonmissing; else
- 2 (black/African American only), if EDQD052 was nonmissing; else
- 3 (American Indian/Alaska Native only), if EDQD053 was nonmissing; else

[23] A value of 2 indicated that the respondent wrote "Asian" in the QD05 other-specify blank. A value of 3 indicated that the response was obtained from the other-specify part of the Hispanic/Latino group question (QD04). Finally, a value of 4 indicated that the respondent gave a country of origin as a response to QD05, and the census for that country had "Asian" as one of its categories.

- 4 (Asian/Other Pacific Islander only), if any of EDQD054 through EDQD0513 were nonmissing.

Scenario 2: If two broad racial categories were identified in QD04, QD05, and/or QD05ASIA, EDRACE =

- 5 (white and black/African American only), if both EDQD051 and EDQD052 were nonmissing; else
- 6 (white and American Indian/Alaska Native only), if both EDQD051 and EDQD053 were nonmissing; else
- 7 (white and Asian/Other Pacific Islander only), if EDQD051 was nonmissing and at least one of EDQD054 through EDQD0513 were nonmissing; else
- 8 (black/African American and American Indian/Alaska Native only), if both EDQD052 and EDQD053 were nonmissing; else
- 9 (black/African American and Asian/Other Pacific Islander only), if EDQD052 was nonmissing and at least one of EDQD054 through EDQD0513 were nonmissing; else
- 10 (American Indian/Alaska Native and Asian/Other Pacific Islander only), if EDQD053 was nonmissing and at least one of EDQD054 through EDQD0513 were nonmissing.

Scenario 3: If three broad racial categories were identified in QD04, QD05, and/or QD05ASIA, EDRACE =

- 11 (white, black/African American, and American Indian/Alaska Native only), if all of EDQD051 through EDQD053 were nonmissing; else
- 12 (white, black/African American, and Asian/Other Pacific Islander only), if both EDQD051 and EDQD052 were nonmissing and at least one of EDQD054 through EDQD0513 were nonmissing; else
- 13 (white, American Indian/Alaska Native, and Asian/Other Pacific Islander only), if both EDQD051 and EDQD053 were nonmissing and at least one of EDQD054 through EDQD0513 were nonmissing; else
- 14 (black/African American, American Indian/Alaska Native, and Asian/Other Pacific Islander only), if both EDQD052 and EDQD053 were nonmissing and at least one of EDQD054 through EDQD0513 were nonmissing.

Scenario 4: If all four broad racial categories were identified in QD04, QD05, and/or QD05ASIA, EDRACE = 15.

Scenario 5: If none of the broad racial categories were identified in QD04, QD05, and/or QD05ASIA, EDRACE =

- 16 (multiple race, no other information), if an other-specify answer such as "biracial" or "mixed" appeared in QD04, QD05, or QD05ASIA; else

- 17 (nonwhite, no other information), if an other-specify answer such as "brown," "tan," or similar answers in Spanish appeared in QD04, QD05, or QD05ASIA; else
- 18 (white, or both white and American Indian/Alaska Native), if the random assignment of a census data code resulted in imputation restricted to donors who were either white, or both white and American Indian/Alaska Native; else
- 19 (not American Indian/Alaska Native, in part or in full), if the random assignment of a census data code resulted in imputation restricted to donors who were not American Indian/Alaska Native, in part or in full; else
- 20 (non-Hispanic Mexican), if "Mexican" was mentioned in the QD05 and/or QD05ASIA other-specify responses, but QD03 = 2; else
- 21 (non-Hispanic Cuban), if "Cuban" was mentioned in the QD05 and/or QD05ASIA other-specify responses, but QD03 = 2; else
- 22 (non-Hispanic Central or South American), if "Central or South American" was mentioned in the QD05 and/or QD05ASIA other-specify responses, but QD03 = 2; else
- 23 (non-Hispanic Dominican), if "Dominican" was mentioned in the QD05 and/or QD05ASIA other-specify responses, but QD03 = 2; else
- 24 (non-Hispanic Spanish), if "Spanish" was mentioned in the QD05 and/or QD05ASIA other-specify responses, but QD03 = 2; else
- missing.

4.2.6.4.3 Broad Categories of Race, No Multiple Race (EDRACEFORMODEL)

Because of the paucity and heterogeneity of multiple-race respondents, imputation models for race did not include a category for more than one race. Instead, predicted means were determined in multinomial logistic models with the following four categories:

1. American Indian/Alaska Native
2. Asian/Other Pacific Islander
3. Black/African American
4. White

In previous survey years, multiple-race respondents were assigned a single race based on the response to QD06, the multiple-race respondent's "main race." Multiple-race respondents who did not answer QD06 were allocated a "main race" based on an arbitrary priority rule (black/African American, Asian/Other Pacific Islander, American Indian/Alaska Native, white). Imputation donors were chosen with predicted means for these four categories close to those of the recipient with missing race. A respondent was imputed as being more than one race if the selected donor also identified more than one race.

As in past survey years, an edited variable that did not include a category for more than one race was necessary in the 2007 survey because (1) it was needed to build the imputation models; and (2) it was necessary as a base variable for the final imputation-revised variable that did not include a category for more than one race. However, with the absence of QD06 since the

2003 survey, the respondent did not have an opportunity to indicate a "main race," so a main race had to be assigned probabilistically using models. This edited variable (EDRACEFORMODEL) included the four broad categories given above. Using data pooled across the 2000-2002 survey years, a single race was imputed for multiple-race respondents using a series of logistic models. The modeling process is described in Appendix E. Eleven predictive mean models were fit, one for each multiple racial category (EDRACE between the values of 5 and 15 inclusive). The parameter estimates from the models were used to impute a "main" or "best" race by the following procedure:

Step 1: Estimate the probability that each respondent would mention each of the broad racial categories indicated as their "main" race, using the coefficients from the appropriate predictive mean model.

Step 2: Randomly select one of the broad racial categories based on these probabilities.

For example, consider a respondent in the 2007 survey with EDRACE = 5 (white and black/African American only). The covariates included in the model, as described in Appendix E, for respondents with EDRACE = 5 were age, region, race of householder, percentage owner-occupied households, percentage Asian population, percentage American Indian/Alaska Native population, and percentage black/African-American population. Using the values for these covariates for the 2007 respondent and the parameter estimates from the model, the probability that the respondent would select white as his main race could be estimated. If this probability was estimated at 50 percent, a random imputation was done such that the respondent was assigned white as his main race with probability 50 percent and black/African American as his main race with probability 50 percent.

The assignment of values for EDRACEFORMODEL is listed below:

EDRACEFORMODEL =

- EDRACE, if $1 \leq \text{EDRACE} \leq 4$; else
- randomly imputed main race, if $5 \leq \text{EDRACE} \leq 15$; else
- missing.

4.2.6.4.4 Finer Categories of Race (EDNWRACE)

EDNWRACE was a 15-level edited variable used as a base variable for the imputation-revised finer racial category variable IRNWRACE. It also had a 16th level to indicate when the imputation should be restricted to Asian-specific categories. It was created using the following rules, under three possible scenarios:

Scenario 1: If only one of EDQD051 through EDQD0513 was nonmissing,

EDNWRACE =

- 16 (Asian nonspecific only), if EDQD0513 was the nonmissing variable; else

- *xx* (one known racial category only), where EDQD05*xx* was the nonmissing variable out of EDQD051 through EDQD0512.

Scenario 2: If more than one of EDQD051 through EDQD0513 was nonmissing,

EDNWRACE =

- 13 (Native Hawaiian and Other Pacific Islander only), if both EDQD054 and EDQD055 were nonmissing, and all other EDQD05*xx* variables were missing; else
- 14 (Asian multiple category), if all of EDQD051 through EDQD055 were missing (i.e., at least two of the ordinary Asian categories were selected); else
- 15 (more than one race).

Scenario 3: If all of EDQD051 through EDQD0513 were missing,

EDNWRACE =

- 15 (more than one race), if EDRACE = 16; else
- missing.

4.2.6.5 Edited Variables, Hispanicity

4.2.6.5.1 Hispanic/Latino Indicator (EDHOIND)

An imputation-revised Hispanic/Latino indicator, EDHOIND, was created using responses to QD03 and, in rare cases, the other-specify responses to QD04, QD05, and/or QD05ASIA. This indicator variable was created as follows:

EDHOIND =

- 1 (Hispanic/Latino), if QD03 = 1 and no other-specify response stated that the respondent was definitely not Hispanic/Latino, or if the other-specify response to QD05 or QD05ASIA indicated that the respondent was definitely Hispanic/Latino; else
- 2 (not Hispanic/Latino), if QD03 = 2 and no other-specify response stated that the respondent was definitely Hispanic/Latino, or if the other-specify response to QD04, QD05, and/or QD05ASIA indicated that the respondent was definitely not Hispanic/Latino; else
- missing.

The race other-specify responses, which were considered "definitely Hispanic/Latino," and the single Hispanic/Latino other-specify response, which was considered "definitely not Hispanic/Latino," are listed in Appendix D.

4.2.6.5.2 Individual Hispanic/Latino Group Categories (EDQD041-EDQD047)

The edited variables EDQD041 through EDQD047 were created to match the seven Hispanic/Latino group categories described in Section 4.2.6.2.2: Mexican, Puerto Rican, Central or South American, Cuban, Dominican, Spanish, and Other Hispanic/Latino.

EDQD04*xx* =

- 1, if the level *xx* was selected by the respondent in QD04; else
- 2, if the other-specify response from QD04 mapped directly to level *xx*; else
- 3, if no EDQD04*xx* variables had values of 1 or 2, and the other-specify response from QD05 or QD05ASIA mapped directly to level *xx*; else
- missing.

4.2.6.5.3 Edited Hispanic/Latino Group (EDHOGRP)

The edited variable EDHOGRP was the base variable for creating an imputation-revised Hispanic/Latino group variable. It had seven levels to match the seven Hispanic/Latino group categories described in Section 4.2.6.2.2, plus several other more general Hispanic/Latino levels that could be used in a restricted imputation. Those respondents with EDHOIND = 2 were assigned EDHOGRP = 99. It was created using the following rules, under four possible scenarios:

Scenario 1: If EDHOIND = 2,

EDHOGRP = 99.

Scenario 2: If EDHOIND = 1 or missing and only one of EDQD041 through EDQD047 was nonmissing,

EDHOGRP = *xx*, where EDQD04*xx* was the nonmissing one.

Scenario 3: If EDHOIND = 1 or missing and more than one of EDQD041 through EDQD047 was nonmissing,

EDHOGRP =

- 1 (Mexican), if EDQD041 was nonmissing; else
- 2 (Puerto Rican), if EDQD042 was nonmissing; else
- 3 (Central or South American), if EDQD043 was nonmissing; else
- 4 (Cuban), if EDQD044 was nonmissing; else
- 5 (Dominican), if EDQD045 was nonmissing; else
- 6 (Spanish), if EDQD046 was nonmissing.

For the multiple-Hispanic/Latino group respondents, an arbitrary priority rule similar to the one used in the surveys prior to 2004 was applied in determining a single Hispanic/Latino group. The only difference is the addition of two more Hispanic/Latino group categories since the 2004 survey, resulting in the following order: Mexican, Cuban, Puerto Rican, Central or South American, Dominican, Spanish, and Other Hispanic/Latino.

Scenario 4: If EDHOIND = 1 or missing and all of EDQD041 through EDQD047 were missing,

EDHOGRP =

- EDRACE + 7 (imputation restricted by race), if $1 \leq$ EDRACE ≤ 14; else
- missing.

4.2.7 Highest Grade Completed (EDUC and EDEDUC)

EDUC and EDEDUC were created using the responses to the core education question QD11, which asked about the highest grade in school completed by the respondent. No editing was performed on other questionnaire information, and although EDUC contained codes describing the type of nonresponse, EDEDUC was set to missing if no response was given to QD11.

In the 2007 questionnaire, a single core question (QD11) asked about the respondent's education level, in terms of the highest grade that the respondent had completed:

QD11: What is the highest grade or year of school you have **completed**?

0	NEVER ATTENDED SCHOOL
1	1ST GRADE COMPLETED
2	2ND GRADE COMPLETED
3	3RD GRADE COMPLETED
4	4TH GRADE COMPLETED
5	5TH GRADE COMPLETED
6	6TH GRADE COMPLETED
7	7TH GRADE COMPLETED
8	8TH GRADE COMPLETED
9	9TH GRADE COMPLETED
10	10TH GRADE COMPLETED
11	11TH GRADE COMPLETED
12	12TH GRADE COMPLETED
13	COLLEGE OR UNIVERSITY / 1ST YEAR COMPLETED
14	COLLEGE OR UNIVERSITY / 2ND YEAR COMPLETED
15	COLLEGE OR UNIVERSITY / 3RD YEAR COMPLETED
16	COLLEGE OR UNIVERSITY / 4TH YEAR COMPLETED
17	COLLEGE OR UNIVERSITY / 5TH OR HIGHER YEAR COMPLETED

The creation of the edited variable derived from QD11, EDUC, is described in Kroutil and Chien (2008). The base variable for creating an imputation-revised version of education was called EDEDUC and was equivalent to EDUC, except that missing values were set to the SAS missing code (.) so that they were properly handled by the modeling programs.

4.3 Demographics Requiring Imputation

Missing values for the demographic variables of completed cases were imputed separately from all eligible (screener) rostered individuals. Moreover, almost no screener information was used in the imputation of questionnaire demographics for the completed cases. The exception involved an important covariate in the race imputation model, which is explained in Section 4.3.2. The descriptions that follow discuss the creation of imputation-revised demographic variables. Detailed descriptions of the screener-derived and segment-level[24] covariates used in the imputation models are given in Appendix F.

4.3.1 Marital Status

4.3.1.1 Imputation-Revised Marital Status (IRMARIT)

The variable of interest for marital status was a four-level nominal variable. The four substantive levels of the imputation-revised marital status variable, IRMARIT, were the same as the four answer categories in QD07 (married, widowed, divorced or separated, or never married) and its edited counterparts, MARITAL and EDMARIT, which are described in Section 4.2.5. Respondents younger than 15 years were automatically assigned an IRMARIT value of 99, a "legitimate skip" code. The PMN method, as applied to the marital status variable, is explained in the next four sections: setup for model building, computation of predicted means, assignment of imputed values, and constraints on multivariate predictive mean neighborhoods (MPMNs).

4.3.1.1.1 Setup for Model Building

Imputations at the hot-deck stage were conducted separately within each of three age groups: 12 to 17, 18 to 25, and 26 or older, though only a single model was fit across all age groups. All respondents with AGE younger than 15 years were assigned IRMARIT = 99. Only interview respondents with AGE of 15 years or older were considered donors.

An interview respondent was considered an item nonrespondent for marital status if his or her value for EDMARIT was missing. The weights of the item nonrespondents aged 15 years or older were reallocated to the item respondents aged 15 years or older, using an item response propensity model. The item response propensity model is a special case of the generalized exponential model (GEM),[25] which is described in Appendix B. The covariates in the item response propensity model were census region, gender, population density, age categories, percentage Hispanic/Latino population, percentage black/African-American population,

[24] Segments were the second-stage sample units in the multistage 2007 NSDUH sample. Each segment consisted of a set of U.S. Census Bureau blocks. Segment-level covariates were defined across the segment in which the respondent's household was located.

[25] The GEM macro, which was written in SAS/IML® software, was developed at RTI International (a trade name of Research Triangle Institute) for weighting procedures.

percentage American Indian/Alaska Native population, percentage Asian population, percentage owner-occupied households, and the interaction of age categories and gender.

4.3.1.1.2 Computation of Predicted Means

Using the adjusted weights, the probability of selecting each marital status category (married, widowed, divorced or separated, and never married) was modeled for all age groups together[26] using polytomous logistic regression.[27] The predictors included in the predictive mean model were census region, gender, population density, centered age, percentage Hispanic/Latino population, percentage black/African-American population, percentage American Indian/Alaska Native population, percentage Asian population, percentage owner-occupied households, and the interaction of centered age and gender. These variables were included in both the response propensity and the predictive mean models, unless a convergence problem occurred. If this happened, the model was reduced. A summary of the final set of covariates used in the model can be found in Appendix F.

4.3.1.1.3 Assignment of Imputed Values

Separate assignments were performed within each of the three age groups: 15 to 17, 18 to 25, and 26 or older. Respondents aged 12 to 14 were assigned the legitimate skip code 99. The constraints used to select donors are described in the next section.

4.3.1.1.4 Constraints on Multivariate Predictive Mean Neighborhoods

No logical constraints were used in defining neighborhoods for the marital status variable; only likeness constraints were utilized. In the first attempt to find a neighborhood for each item nonrespondent, two likeness constraints were used. The first constraint required each of the donor's three predicted means, as described in Section 4.3.1.1.2, to be within 5 percent of each of the recipient's three predicted means. The second constraint required donors and recipients to have an age difference of 3 years or less. If no item respondents met the above conditions for a particular item nonrespondent, the likeness constraints on the predicted means were removed. See Appendix G for the numbers of respondents meeting each set of likeness constraints on sets of eligible donors.

4.3.1.1.5 Imputation and Editing Summary for Marital Status

See Table 4.4 for a summary of item nonresponse for marital status (recorded in the variable IIMARIT).

[26] All age groups were modeled together because the distributions of the answers for the youngest two age groups were unbalanced, which made it difficult to find convergent models.

[27] SAS®-callable SUDAAN® was used to fit all polytomous logistic regression models. Details about the polytomous logistic regression model and additional references can be found in the *SUDAAN Language Manual Addendum, Release 9.0.3* (RTI International, 2007). SAS software is a registered trademark of SAS Institute, Inc. SUDAAN is a registered trademark of Research Triangle Institute.

Table 4.4 Marital Status Editing and Imputation Summary

Value of IIMARIT	Assignment of Marital Status	Frequency	Percent
1	From questionnaire	56,830	83.73
3	Statistically imputed	18	0.03
9	Legitimate skip (≤ 14 years old)	11,022	16.24

4.3.2 Race, Hispanic/Latino Origin Indicator, Hispanic/Latino Group

4.3.2.1 Introduction

As discussed in Section 4.2.6, race and Hispanicity were closely related in the 2007 survey. Moreover, race was used in the imputation of Hispanic/Latino origin, and Hispanicity was used in the imputation of race. The imputation of missing values in the race and Hispanicity variables will be discussed together in this section.

4.3.2.2 Imputation-Revised Race Variables

Sections 4.2.6.4.1 through 4.2.6.4.4 outline the edited variables describing race. Nearly all of these edited variables had imputation-revised counterparts, as shown in Table 4.5. Some of the individual racial category variables were collapsed at the imputation stage.

All of these variables could be imputed simultaneously, though the imputations of IRDETAILEDRACE, IRRACE2, and IRNWRACE occurred first, and then the imputations of the individual racial category variables (IRRACEWH, IRRACEBK, IRRACENA, IRRACENH, IRRACEPI, and IRRACEAS) were imputed. This was accomplished by assigning values for the individual racial category variables using the same donors as in the earlier imputation of IRDETAILEDRACE, IRRACE2, and IRNWRACE.

Table 4.5 Edited Race Variables and Their Imputation-Revised Counterparts

Edited Race Variable	Imputation-Revised Race Variable
EDQD051	IRRACEWH
EDQD052	IRRACEBK
EDQD053	IRRACENA
EDQD054	IRRACENH
EDQD055	IRRACEPI
EDQD056-EDQD0513 (collapsed)	IRRACEAS
EDRACE	IRDETAILEDRACE
EDRACEFORMODEL	IRRACE2
EDNWRACE	IRNWRACE

Whereas their edited counterparts had different codes depending upon the source of the information, the IRRACE*xx* variables were simply binary indicator variables, which were set to 1 if the respondent indicated the given race and were set to 0 otherwise. The extra information that was contained in the EDQD05*xx* variables was stored in the concomitant IIRACE*xx* variables. The variable IRDETAILEDRACE, which was the only one of these variables not released to the public use and analytic files, gives the same information as the IRRACE*xx* variables, all within a single variable. The final race variable IRRACE2 was a four-level nominal variable: American Indian/Alaska Native, Asian or Other Pacific Islander, black/African American, and white.[28] This variable has the same levels as IRRACE from previous surveys. The two variables differed in the way they were edited and in the handling of multiple race respondents. Because of the differences, the variable's name was changed. IRNWRACE was a 15-level nominal variable whose levels were the same as the first 15 levels of EDNWRACE.

The imputation-revised race variables were created using an MPMN method for imputation of missing values. The MPMN method, as applied to the race variables, is explained in the next four sections. It should be noted that the models used in PMN did not have a separate category for multiple-race respondents, because of the small number of these respondents as well as their disparate nature. Instead, a model with four broad categories was used, which were the same categories found in IRRACE2. Multiple-race respondents in the model were assigned a single race based on the models discussed in Appendix E. They were included in the model-building process as belonging to one of the four broad racial categories. Respondents requiring imputation were considered to be of more than one race if their donor in the hot-deck step of PMN was a multiple-race respondent.

4.3.2.2.1 Setup for Model Building

As with all other variables imputed using PMN methods, the race imputations were conducted separately within age groups. For race and other demographic variables, there were three age groups: 12 to 17, 18 to 25, and 26 or older. The separate age groups were used for ease of processing and consistency with other variables and not because of any strong correlation between age and race. Because all interview respondents were asked the race questions, no subsetting of the data was necessary.

Before predictive mean modeling was implemented, weights were adjusted for item nonresponse to the race questions. (Because the final weight adjustments for the 2007 survey were not completed at the time of the demographic imputations, the person-level sample design weights were adjusted to account for nonresponse at the household level using a simple ratio adjustment.[29]) An interview respondent was considered an item nonrespondent for race if either EDRACEFORMODEL was missing, EDNWRACE was missing or had a value of 16 (multiple races), or both. (If the respondent had missing data for either EDRACEFORMODEL or EDNWRACE, then he or she also had missing data for the other edited variables in Table 4.5 [EDQD051-EDQD0513 and EDRACE].) The weights of the item nonrespondents were redistributed among the item respondents using an item response propensity model, one for each

[28] To collapse the racial categories into these four levels, the following categories from QD05 were included in the category "Asian or Pacific Islander": Native Hawaiian, Other Pacific Islander, Chinese, Filipino, Japanese, Asian Indian, Korean, Vietnamese, and Other Asian.

[29] In subsequent text, the use of the word "weights" refers to these ratio-adjusted design weights.

of the three age groups. The covariates in these models included census region, household type (from the screener), age categories (for the respondents aged 26 or older), percentage Hispanic/Latino population, percentage owner-occupied households, percentage black/African-American population, percentage American Indian/Alaska Native population, and percentage Asian population.

4.3.2.2.2 Computation of Predicted Means

Using the adjusted weights, the probability of selecting each racial category was modeled within each age group using polytomous logistic regression. The predictors included in the models were the same as those used in the item response propensity model for race.

The PMN method for race was multivariate, as opposed to univariate, because the predictive mean vector contained more than one element. The three elements in the vector were the predicted probability of being identified with each of the first three racial categories (white, black/African American, American Indian/Alaska Native). The probability of being classified as Asian/Other Pacific Islander was not included, because it was completely defined by the first three elements in the predictive mean vector, being calculated as one minus their sum. A predictive mean vector of three predicted means was created from the polytomous logistic regression model. The covariates in these models included: census region, household type (from the screener), centered age, centered age squared, centered age cubed (oldest age group only), percentage Hispanic/Latino population, percentage owner-occupied households, percentage black/African-American population, percentage American Indian/Alaska Native population, and percentage Asian population. The number of covariates was reduced if convergence or stability problems occurred in the model-fitting process. A summary of the final set of covariates used in the model can be found in Appendix F.

Conditional probabilities were calculated for the few item nonrespondents with EDRACE values of 18 or 19. For details on the computation of these conditional probabilities, see Appendix G.

4.3.2.2.3 Assignment of Imputed Values

For the race questions, the PMN method required the selection of an item respondent who was similar to each item nonrespondent. Specifically, the item respondent "donated" his or her value for the relevant edited variables in Table 4.5 to the item nonrespondent. Most often, the selected item respondent, called the "donor," was randomly chosen from a "neighborhood" of potential donors. The item respondents in this neighborhood were the ones deemed to be most similar to the given item nonrespondent, that is, the "recipient." Item respondents who were deemed dissimilar to the recipient were discarded from the neighborhood by means of constraints. The predicted means calculated in the previous step were usually considered in these constraints. Because multiple variables were considered in the distance measure, "similarity" was defined in terms of the smallest Mahalanobis distance.[30] The PMN methodology is described in more detail in Appendix C. The constraints used for the race variables are described in the next section.

[30] See Appendix C for a definition of Mahalanobis distance. A definition also can be found in Manly (1986).

Separate assignments were performed within each of the three age groups. This type of age group-specific assignment was executed for almost all imputation-revised variables in the 2007 survey. If the recipient had missing values for EDRACEFORMODEL and EDNWRACE (as well as the other edited variables in Table 4.5), the donor gave values for all relevant variables to the recipient. In most cases, this ensured consistency between each of the imputation-revised variables. An exception occurred when a respondent listed only one specific category of race but indicated that he or she was more than one race in the other-specify entry. In these rare cases, the respondent was "more than one race" in IRNWRACE, but only one race was given in the IRRACE*xx* and IRDETAILEDRACE variables.

4.3.2.2.4 Constraints on Multivariate Predictive Mean Neighborhoods

For the MPMN method, there were two types of constraints: logical constraints and likeness constraints. Logical constraints were not loosened during the search for a donor. Likeness constraints were either loosened or removed if a donor was not found with the given constraints in effect. The logical constraints on the donors for EDRACEFORMODEL and EDNWRACE are listed below:

1. If the recipient was known to be Asian (i.e., EDNWRACE = 16), then the donor must also have been Asian.
2. If the recipient had EDRACE = 16 (multiple race, no other information), then the donor must have had EDNWRACE = 15.
3. If the recipient had EDRACE = 17 (nonwhite, no other information), then the donor must not have had EDNWRACE = 1.
4. If the recipient had EDRACE = 18 (white, or both white and American Indian/Alaska Native), then the donor must have had EDRACE = 1 or 6.
5. If the recipient had EDRACE = 19 (not American Indian/Alaska Native, in part or in full), then the donor must not have had EDRACE = 3, 6, 8, 10, 11, 13, 14, or 15.

In the first attempt to find a neighborhood for each item nonrespondent, a set of likeness constraints was used. The first likeness constraint stated that the donor must live in the same segment as the recipient. The second likeness constraint stated that each of the donor's three predicted means (two when the recipients had EDRACE = 19, one when EDRACE = 18), as described in Section 4.3.2.2.2, must be within 5 percent (within "delta") of each of the recipient's three predicted means. If no potential donors met both of the above conditions for a particular item nonrespondent, the constraint on the segment of the potential donor was removed first. If no potential donors met the "delta constraint," the delta constraint also was removed. In addition to these two constraints, a set of likeness constraints concerning the donor's Hispanicity were used when the recipient met one of the following conditions. These likeness constraints were never loosened or removed.

1. If the recipient was Hispanic/Latino nonspecific (EDHOIND = 1, and all of EDQD041 through EDQD046 were missing), then the donor must have been of Hispanic/Latino origin.

2. If the recipient selected one or more Hispanic/Latino categories: Mexican, Puerto Rican, Central or South American, Cuban, Dominican, Spanish (EDHOIND = 1, and one or more EDQD041 through EDQD046 were nonmissing), then the donor must have had an EDHOGRP value equal to one of the Hispanic/Latino groups mentioned by the recipient. For example, if the recipient chose Mexican and Central or South American, then the donor must have had EDHOGRP = 1 or 3.
3. If the recipient had EDRACE = 20 (non-Hispanic Mexican), then the donor must have been Mexican (but the donor could have been Hispanic/Latino or non-Hispanic/Latino).
4. If the recipient had EDRACE = 21 (non-Hispanic Cuban), then the donor must have been Cuban (but the donor could have been Hispanic/Latino or non-Hispanic/Latino).
5. If the recipient had EDRACE = 22 (non-Hispanic Central or South American), then the donor must have been Central or South American (but the donor could have been Hispanic/Latino or non-Hispanic/Latino).
6. If the recipient had EDRACE = 23 (non-Hispanic Dominican), then the donor must have been Dominican (but the donor could have been Hispanic/Latino or non-Hispanic/Latino).
7. If the recipient had EDRACE = 24 (non-Hispanic Spanish), then the donor must have been Spanish (but the donor could have been Hispanic/Latino or non-Hispanic/Latino).

The likeness and logical constraints for the race variables, along with the number of nonrespondents imputed using each set of constraints, are listed in Appendix G.

4.3.2.2.5 Imputation and Editing Summary for Race

To differentiate the final imputed values from nonmissing values, a concomitant indicator variable, IIRACE2, indicated how the levels of IRRACE2 were derived. Table 4.6 presents the levels for the indicators of the individual racial category variables (IIRACE*xx*). The levels for IRRACE2 are provided in Table 4.7. The 15-level race variable, IRNWRACE, also had a concomitant indicator variable. Table 4.8 shows the levels of IINWRACE, the concomitant indicator variable for IRNWRACE. No indicator variable was created for IRDETAILEDRACE.

Table 4.6 IRRACE*xx* Editing and Imputation Summary

Value of IIRACE*xx*	Assignment of IRRACE*xx*	*xx* = WH (white)		*xx* = BK (black/African American)		*xx* = NA (American Indian/Alaska Native)	
		Freq.	Pct.	Freq.	Pct.	Freq.	Pct.
1	Directly selected/not selected	65,730	96.85	65,944	97.16	65,863	97.04
2	From other-specify	203	0.30	42	0.06	79	0.12
3	From census data	77	0.11	24	0.04	68	0.10
4	Statistically imputed	1,860	2.74	1,860	2.74	1,860	2.74
Value of IIRACE*xx*	**Assignment of IRRACE*xx***	***xx* = NH (Native Hawaiian)**		***xx* = PI (Other Pacific Islander)**		***xx* = AS (Asian)**	
		Freq.	**Pct.**	**Freq.**	**Pct.**	**Freq.**	**Pct.**
1	Directly selected/not selected	66,010	97.26	66,008	97.26	65,732	96.85
2	From other-specify	0	0.00	2	0.00	275	0.41
3	From census data	0	0.00	0	0.00	3	0.00
4	Statistically imputed	1,860	2.74	1,860	2.74	1,860	2.74

Table 4.7 IRRACE2 Editing and Imputation Summary

Value of IIRACE2	Assignment of IRRACE2	Frequency	Percent
1	From single QD05 response	63,194	93.11
2	Logically assigned from alpha-specify response	419	0.62
3	Single race imputed from multiple responses	2285	3.37
4	Single race assigned with census data from country of origin	53	0.08
5	Multiple races assigned with census data, single race imputed	59	0.09
6	Statistically imputed (unrestricted)	27	0.04
7	Statistically imputed (restricted)	1,833	2.70

Table 4.8 IRNWRACE Editing and Imputation Summary

Value of IINWRACE	Assignment of IRNWRACE	Frequency	Percent
1	From QD05 response(s)	65,368	96.31
2	Logically assigned from alpha-specify response(s)	566	0.83
3	Assigned with census data from country of origin	109	0.16
4	Statistical imputation of "Asian" into finer categories	8	0.01
5	Statistically imputed (unrestricted)	27	0.04
6	Statistically imputed (restricted)	1,792	2.64

4.3.2.3 Imputation-Revised Hispanic/Latino Indicator (IRHOIND)

As with the imputation-revised race variables, a PMN method was used for the Hispanic/Latino indicator. However, because there was only one element in the predictive mean vector in this case, a univariate predictive mean neighborhood (UPMN) method was used. The PMN method, as applied to the Hispanic/Latino indicator, is explained in the next four sections.

4.3.2.3.1 Setup for Model Building

As with imputations for the race variables, the imputations for the Hispanic/Latino indicator were conducted separately within the three age groups: 12 to 17, 18 to 25, and 26 or older. The separate age groups were used more for ease of processing and consistency with other variables rather than because of any strong correlation between age and Hispanic/Latino origin. Because all interview respondents were asked the question about Hispanic/Latino origin, no subsetting of the data was necessary.

As for the race variables, weights were adjusted for item nonresponse to the Hispanic/Latino origin question, QD03, using item response propensity models, one for each age group. Weights were defined in a similar manner to the way weights were determined for other demographic variables. Details on how the weights were defined can be found in Section 4.3.2.2.1. The potential covariates in the item response propensity model were census region, imputation-revised race, age categories (for the 26-or-older age group), percentage Hispanic/Latino population, percentage owner-occupied households, percentage black/African-American population, percentage American Indian/Alaska Native population, and percentage Asian population.

4.3.2.3.2 Computation of the Predicted Means

Using the adjusted weights, the probability of an affirmative response to the Hispanic/Latino origin question was modeled within each age group using logistic regression. The predictors included in the models were census region, imputation-revised race, household type, centered age, centered age squared, centered age cubed, imputation-revised marital status, percentage Hispanic/Latino population, percentage owner-occupied households, percentage black/African-American population, percentage American Indian/Alaska Native population, and percentage Asian population. The number of covariates was reduced if convergence or stability problems occurred in the model-fitting process. A summary of the final set of covariates used in the model can be found in Appendix F.

4.3.2.3.3 Assignment of Imputed Values

Separate assignments were performed within each of the three age groups: 12 to 17, 18 to 25, and 26 or older. The constraints used to select donors are described in the next section.

4.3.2.3.4 Constraints on Univariate Predictive Mean Neighborhoods

No logical constraints were used in defining neighborhoods; only likeness constraints were utilized. In the first attempt to find a neighborhood for each item nonrespondent, two likeness constraints were used. The first likeness constraint stated that the donor must live in the same segment as the recipient. The second likeness constraint stated that the donor's predicted

mean, as described in Section 4.3.2.3.2, must be within 5 percent of the recipient's predicted mean. If no item respondents met the above conditions for a particular item nonrespondent, the constraint on the segment of the potential donor was removed. A donor was found for every item nonrespondent using this method. Therefore, no further loosening of constraints was necessary. See Appendix G for the numbers of respondents who met each set of likeness constraints on sets of eligible donors.

4.3.2.3.5 Imputation and Editing Summary for Hispanic/Latino Origin

Less imputation was required for the Hispanic/Latino indicator than for the race variables. Table 4.9 presents item nonresponse for the Hispanic/Latino indicator. This information was recorded in the variable IIHOIND.

Table 4.9 Hispanic/Latino Indicator Editing and Imputation Summary

Value of IIHOIND	Assignment of IRHOIND	Frequency	Percent
1	From questionnaire	67,761	99.84
2	From alpha-specify responses	2	0.00
3	Statistically imputed	107	0.16

4.3.2.4 Race and Hispanicity Recodes Used in Subsequent Processing

The imputation-revised race (IRRACE2) and imputation-revised Hispanic/Latino indicator (IRHOIND) variables were used to create several additional race/ethnicity variables. One of these was used in the subsequent processing of imputation-revised variables. Since the 2003 survey, this variable (RACE2) has had four levels: non-Hispanic/Latino white, non-Hispanic/Latino black/African American, Hispanic/Latino, and non-Hispanic/Latino other. However, it was similar to the variable RACE created in previous survey years from IRRACE and IRHOIND. RACE had the same levels as RACE2. Other variables were created from IRNWRACE and IRHOIND that were used extensively in the production of tables (NEWRACE1 and NEWRACE2).

4.3.2.5 Imputation-Revised Hispanic/Latino Group (IRHOGRP4)

4.3.2.5.1 Introduction

Because two additional Hispanic/Latino group categories (Dominican and Spanish) were added to QD04 since the 2004 survey, a final imputation-revised Hispanic/Latino group variable, IRHOGRP4, was created to differentiate from IRHOGRP3, which was created prior to 2004. With the added Hispanic/Latino group category "Dominican," there were very few respondents who classified themselves as non-Dominican Caribbean. Therefore, the Hispanic/Latino Caribbean level that was present in IRHOGRP3 was eliminated from IRHOGRP4 and was collapsed into the "Other" Hispanic/Latino group category. The edited variable EDHOGRP, described in Section 4.2.6.5.3, categorized Hispanic/Latino respondents into Hispanic/Latino groups. These categories were directly mapped to the same categories in the imputation-revised variable, IRHOGRP4, which had eight possible values: Puerto Rican, Mexican, Cuban, Central

or South American, Dominican, Spanish, Other Hispanic/Latino, and not Hispanic/Latino. It was created using an MPMN method similar to the method for IRMARIT. The predictive mean vector had only three elements associated with the first three levels of EDHOGRP: the predicted probabilities of the interview respondent being Mexican, Puerto Rican, or Cuban. Using only three predicted means made the computation of both predicted means and Mahalanobis distances more feasible.[31]

The PMN method, as applied to the Hispanic/Latino group variable, is explained in the next four sections.

4.3.2.5.2 Setup for Model Building

All respondents with IRHOIND = 2 were automatically assigned IRHOGRP4 = 99 and were excluded from the item response propensity model, the predictive mean model, and the set of potential donors. In contrast to the other demographic variables, imputations were not conducted separately within age groups. This was done for two reasons. First, with combined age groups, the models were likely to be better because none of the response categories were sparsely populated. Second, only respondents with IRHOIND = 1 were eligible to be donors, so it was necessary to keep all age groups in the same dataset to ensure sufficiently large donor pools.

An interview respondent was considered an item nonrespondent for the Hispanic/Latino group if his or her value for EDHOGRP was missing. The weights of the item nonrespondents were then redistributed among the item respondents using an item response propensity model. Covariates included census region, imputation-revised race, gender, age categories, percentage Hispanic/Latino population, percentage black/African-American population, percentage American Indian/Alaska Native population, percentage Asian population, percentage owner-occupied households, and the interaction of age categories and gender.

Starting in the 2004 survey, respondents who indicated multiple Hispanic/Latino groups also were excluded from the model-building process. In past survey years, if a respondent indicated multiple Hispanic/Latino groups, the single Hispanic/Latino group that was determined depended upon a priority rule: Mexican, Cuban, Puerto Rican, Central or South American, Caribbean Islander, and Other Hispanic/Latino. Even though this priority rule was arbitrary, respondents who were assigned a Hispanic/Latino group based on this priority rule have been used in the Hispanic/Latino group models since the 1999 survey. Because the Hispanic/Latino group model did not include a separate level for multiple Hispanic/Latino groups, respondents with multiple Hispanic/Latino groups were not considered item respondents in both the item response propensity model and the predictive mean model.

4.3.2.5.3 Computation of Predicted Means

Using the adjusted weights, the probability of selecting each of the first three Hispanic/Latino group categories (according to EDHOGRP) was modeled for all age groups together, using polytomous logistic regression. The predictors included in the predictive mean model were the same as the predictors used in the response propensity model, except for age-related covariates where the continuous version of age was used in the model. The number of

[31] The ordering of the levels of IRHOGRP4 differed from the questionnaire and from EDHOGRP. The levels were rearranged after all the imputation programs were complete.

covariates was reduced if convergence or stability problems occurred in the model-fitting process. A summary of the final set of covariates used in the model can be found in Appendix F.

4.3.2.5.4 Assignment of Imputed Values

All age groups were aggregated in this step, for the reasons given in Section 4.3.2.5.2. The constraints used to select donors are described in the next section.

4.3.2.5.5 Constraints on Multivariate Predictive Mean Neighborhoods

No logical constraints were used in defining neighborhoods; only likeness constraints were utilized. In the first attempt to find a neighborhood for each item nonrespondent, three likeness constraints were used. The first likeness constraint stated that the donor must live in the same segment as the recipient. The second likeness constraint stated that if the recipient had $8 \leq$ EDHOGRP ≤ 21, then the donor's IRDETAILEDRACE value had to indicate a subset of the racial categories mentioned by the recipient. For example, if the recipient had EDHOGRP = 13 (Hispanic/Latino group missing, and the only races mentioned were white and American Indian/Alaska Native), then the donor must have had IRDETAILEDRACE = 1 (white only), 3 (American Indian/Alaska Native only), or 6 (white and American Indian/Alaska Native only). The third likeness constraint stated that each of the donor's three predicted means, as described in Section 4.3.2.5.1, must be within 5 percent of each of the recipient's three predicted means. If no item respondents met the above conditions for a particular item nonrespondent, the constraint on the segment of the potential donor was removed. If still no donor was found, then the constraints on the predicted means also were removed. The constraint involving race was never loosened or removed. See Appendix G for the numbers of respondents who met each set of likeness constraints on sets of eligible donors.

4.3.2.5.6 Imputation and Editing Summary for Hispanic/Latino Group

To differentiate the final imputed values from nonmissing values, a concomitant indicator variable, II2HOGR4, gave the source of information for IRHOGRP4. The levels of II2HOGR4 are provided in Table 4.10. A variable that gave somewhat less information, IIHOGRP4, also was created to give the source of information for IRHOGRP4. The levels of IIHOGRP4 and II2HOGR4 were identical to the Hispanic/Latino group indicator variables created prior to the 2004 survey. Table 4.10 shows how the levels of II2HOGR4 mapped to those of IIHOGRP4. As previously stated in Section 4.2.6.5.3, a priority rule[32] was used to determine which group a respondent belonged to if he or she gave more than one response. The variable II2HOGR4 recorded these cases, whereas IIHOGRP4 merely considered these cases a "response from questionnaire."

[32] Amended slightly from previous surveys, the priority rule since the 2004 survey was the following: Mexican, Cuban, Puerto Rican, Central or South American, Dominicans, Spanish, and Other Hispanic/Latino.

Table 4.10 Hispanic/Latino Group Editing and Imputation Summary

Value of II2HOGR4	Assignment of IRHOGRP4	Frequency	Percent	Level of IIHOGRP4
1	Single Hispanic/Latino group from questionnaire	9,868	14.54	1
2	Single Hispanic/Latino group from alpha-specify response(s)	111	0.16	2
3	Single Hispanic/Latino group determined from multiple responses	261	0.38	1
4	Statistically imputed (unrestricted), or IRHOIND imputed to 2	118	0.17	3
5	Statistically imputed (restricted by race)	41	0.06	4
9	Legitimate skip (respondent was not Hispanic/Latino—nonimputed)	57,471	84.68	9

4.3.2.6 Imputation-Revised Multiple Hispanic/Latino Group (IRHOGRPM)

4.3.2.6.1 Introduction

As in past survey years, respondents were asked to choose the Hispanic/Latino group(s) that best described them in QD04. They were allowed to select more than one Hispanic/Latino group and also could write in an answer in the QD04 other-specify category. For the respondents from more than one Hispanic/Latino group, a priority rule (see Section 4.2.6.5.3) was used to determine the final Hispanic/Latino group for the respondent. In surveys prior to the 2004 NSDUH, there was no single variable that could report this information. Since the 2004 survey, a multiple Hispanic/Latino groups variable (IRHOGRPM) has been used to capture the information when a respondent identified multiple Hispanic/Latino groups.

4.3.2.6.2 Imputation and Editing Summary for Multiple Hispanic/Latino Group

The imputed variable IRHOGRPM was created based on the imputation-revised variables IRHOIND and IRHOGRP4 and the edited variables EDQD041 through EDQD047, as described in Sections 4.3.2.3, 4.3.2.5, and 4.2.6.5.2, respectively. For the Hispanic/Latino group nonrespondents, the values of EDQD041 through EDQD047 from the donors in IRHOGRP4 were used in place of the missing Hispanic/Latino group categories for the recipients. The first seven levels of IRHOGRM that represented the single Hispanic/Latino group respondents were Puerto Rican only, Mexican only, Cuban only, Other Hispanic/Latino only, Central or South American only, Dominican only, and Spanish only. Level eight represented the respondents from multiple Hispanic/Latino groups. A legitimate skip code of 99 was assigned to the non-Hispanic/Latino group respondents:

IRHOGRPM =

- 99, if IRHOIND = 2; else

- 1 to 7, if IRHOIND = 1 and only one of EDQD041 through EDQD047 was selected; else
- 8, if IRHOIND = 1 and more than one of EDQD041 through EDQD047 were selected.

The source information for IRHOGRPM was recorded in its indicator variable, IIHOGRPM, which appears in Table 4.11.

Table 4.11 Multiple Hispanic/Latino Group Editing and Imputation Summary

Value of IIHOGRPM	Assignment of IRHOGRPM	Frequency	Percent
1	Single or multiple Hispanic/Latino groups from questionnaire	10,120	14.91
2	Single or multiple Hispanic/Latino groups from alpha-specify response(s)	116	0.17
3	Multiple Hispanic/Latino groups from questionnaire and alpha-specify responses	4	0.01
4	Statistically imputed (unrestricted), or IRHOIND imputed to 2	118	0.17
5	Statistically imputed (restricted by race)	41	0.06
9	Legitimate skip (respondent was not Hispanic/Latino—nonimputed)	57,471	84.68

4.3.2.7 Hispanic/Latino Group Recodes Used in Subsequent Processing

Among the recoded variables that were created from IRHOGRP4, one was used in subsequent processing. The variable HISPGRP2 was created by collapsing the levels of IRHOGRP4 into four levels: Puerto Rican, Mexican, Other Hispanic/Latino (includes Cuban, Central or South American, Dominican, Spanish, and Other Hispanic/Latino), and not Hispanic/Latino.

4.3.3 Core Education

4.3.3.1 Imputation-Revised Highest Grade Completed (IREDUC)

As with the marital status, race, and Hispanic/Latino group variables, the predictive mean modeling for the highest grade completed variable was done using polytomous logistic regression. The base edited variable EDEDUC has 17 substantive levels (the same as in QD11), but these were collapsed into fewer levels for ease of modeling. For respondents aged 12 to 17, the predictive mean vector had four elements. For the other two age groups (18 to 25 and 26 or older), the predictive mean vector had three elements. The PMN method, as applied to the highest grade completed variable, is explained in the next four sections.

4.3.3.1.1 Setup for Model Building

The imputations for the highest grade completed variable in the hot-deck stage were conducted separately within the three age groups: 12 to 17, 18 to 25, and 26 or older. Because all interview respondents were asked this question, no subsetting of the data was necessary. Two of these age groups were aggregated for the modeling stage: 18 to 25 and 26 or older.

Weights were adjusted for item nonresponse to the highest grade completed question (QD11). The covariates in the item response propensity model were census region, imputation-revised race, gender, age categories (except in the 12- to 17-year-old age group), the interaction of age categories and gender (except in the 12- to 17-year-old age group), percentage Hispanic/Latino population, percentage black/African-American population, percentage American Indian/Alaska Native population, percentage Asian population, percentage owner-occupied households, and education levels.

4.3.3.1.2 Computation of Predicted Means

For ease of modeling, the 17 substantive levels of EDEDUC were collapsed into fewer levels. For respondents aged 12 to 17, the response variable in the predictive mean model had five levels: less than elementary school (EDEDUC = 1, 2, 3, 4, or 5), elementary school (EDEDUC = 6 or 7), middle school (EDEDUC = 8 or 9), some high school (EDEDUC = 10 or 11), and high school (EDEDUC = 12 or higher). For respondents aged 18 or older, the response variable had four levels: less than high school (EDEDUC < 12), high school (EDEDUC = 12), some college (EDEDUC = 13, 14, or 15), and college or higher (EDEDUC = 16 or 17).

Using the adjusted weights, the probability of the respondent having each level of the response variable was modeled using polytomous logistic regression. The respondents aged 12 to 17 were modeled separately from the two older age groups. For the youngest age group, the predictors included in the model were census region, imputation-revised race, gender, centered age, centered age squared, centered age cubed, the interaction of centered age and gender, the interaction of centered age squared and gender, percentage Hispanic/Latino population, percentage black/African-American population, percentage American Indian/Alaska Native population, percentage Asian population, and percentage owner-occupied households. For the other two age groups, the predictors included in the model were census region, imputation-revised race, gender, centered age, centered age squared, centered age cubed, the interaction of centered age and gender, the interaction of centered age squared and gender, percentage Hispanic/Latino population, percentage black/African-American population, percentage American Indian/Alaska Native population, percentage Asian population, percentage owner-occupied households, and imputation-revised marital status. The number of covariates was reduced if convergence or stability problems occurred in the model-fitting process. A summary of the final set of covariates used in the model can be found in Appendix F.

4.3.3.1.3 Assignment of Imputed Values

Separate assignments were performed within each of the three age groups: 12 to 17, 18 to 25, and 26 or older. The constraints used to select donors are described in the next section.

4.3.3.1.4 Constraints on Multivariate Predictive Mean Neighborhoods

No logical constraints were used in defining neighborhoods for the education level variable; only likeness constraints were utilized. For the two youngest age groups, three likeness constraints were used in the first attempt to find a neighborhood for each item nonrespondent. The first required the donor to be the same age as the recipient. The second stated that the donor must live in the same segment as the recipient. The third likeness constraint stated that the donor's predicted means, as described in Section 4.3.3.1.2, must be within 5 percent of the recipient's predicted means. If no item respondents met the above conditions for a particular item nonrespondent, the constraint on the segment of the potential donor was removed. If potential donors still were not found, then the delta constraints were removed. For the oldest age group, the constraints were the same except that the constraint on the donor's age was not applied. See Appendix G for the numbers of respondents meeting each set of likeness constraints on sets of eligible donors.

4.3.3.1.5 Imputation and Editing Summary for Highest Grade Completed

Table 4.12 shows item nonresponse for the highest grade completed variable. This information was recorded in the variable IIEDUC.

Table 4.12 Highest Grade Completed Editing and Imputation Summary

Value of IIEDUC	Assignment of IREDUC	Frequency	Percent
1	From questionnaire	67,860	99.99
3	Statistically imputed	10	0.01

4.3.3.2 Education Records

EDUCCAT2, a recoded education variable, was created using the imputation-revised highest grade completed variable (IREDUC). EDUCCAT2 had five levels (less than high school and aged 18 or older, high school graduate and aged 18 or older, some college and aged 18 or older, college graduate and aged 18 or older, or 12 to 17 years old).

5. Noncore Demographics

5.1 Introduction

For the 2007 National Survey on Drug Use and Health (NSDUH),[33] missing values were imputed in two sets of variables in the noncore demographics module: the immigrant status and employment status variables. The core demographics that were imputed in the 2007 survey are discussed in Chapter 4.

For immigrant status, two imputation-revised variables (IRBORNUS and IRENTAG2) were created using the edited variables BORNINUS and ENTRYAG2 as base variables.[34] Respectively, these variables recorded whether a respondent was born in the United States, and if not, the variables recorded the age of entry into the United States. The name of the imputation-revised age-of-entry variable (IRENTAG2) was changed in the 2004 survey because of changes in the questionnaire that are described in Section 5.2. IRENTAG2 is analogous to the variable IRENTAGE, which was used for previous surveys. The final goal was to create a data file containing variables that would indicate whether respondents could be included in incidence analyses based on their immigrant status.

The variables describing current employment status were determined from multiple questions in the noncore demographics module. Instead of a single question asking the respondent to describe his or her "current" employment status, several questions were asked regarding the respondent's employment situation during the week preceding the interview and whether that week was atypical. The employment status questions were asked of only respondents aged 15 or older. A single imputation-revised variable, EMPSTATY, was created from the series of employment status questions. Unlike other imputation-revised variables, for historical reasons this variable was not preceded by an "IR" prefix. However, it was accompanied by imputation indicators that did have the requisite "II" prefix: II2EMSTY and IIEMPSTY.

Respondents who either worked during the week preceding the interview or said they had a job were asked to write in the industry for which they worked, their occupation, and their main duties at work. Edited versions of the responses to some of these questions are discussed in a separate document (Kroutil & Chien, 2008). Even though responses were edited, missing values were not imputed.

[33] This report presents information from the 2007 National Survey on Drug Use and Health (NSDUH), an annual survey of the civilian, noninstitutionalized population of the United States aged 12 or older. Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA).

[34] Although these variables are called "immigrant status" variables for convenience purposes, the immigrant questions also included information from eligible respondents who lived in the United States and were not born in the United States, but had no intention of staying permanently in the United States (e.g., foreign students are not immigrants). For this reason, respondents who indicated that they were not born in the United States are called non-U.S.-born respondents in this chapter.

5.2 Immigrant Status

The edited immigrant status variables used to create IRBORNUS and IRENTAG2 are described in Section 5.2.1. The edited variable BORNINUS, the base variable used for creating IRBORNUS, was derived from questions QD14 and QD15 and is described in Section 5.2.1.1. Whereas the indicator of whether the respondent was born in the United States did not change from previous surveys, the determination of the length of time non-U.S.-born respondents had lived in the United States did change in the 2004 survey. In surveys prior to 2004, the length-of-time information was obtained from a single question (QD16) in categorical increments. However, since the 2004 survey, this information was obtained from three questions (QD16a, QD16b, and QD16c), from which a continuous amount of time lived in the United States was obtained. The edited variables LIVUS1YR, LIVUSYRS, and LIVUSMOS were derived from these questionnaire questions and were consolidated into a single variable: LNGTHLIV. The edited age-of-entry variable (ENTRYAG2) was derived from LNGTHLIV and was used as the base variable for IRENTAG2. The variables LIVUS1YR, LIVUSYRS, and LIVUSMOS are discussed in Section 5.2.1.2; LNGTHLIV and CONTAGE are discussed in Section 5.2.1.3; and ENTRYAG2 is discussed in Section 5.2.1.4.

Imputation-revised immigrant status variables were imputed using the weighted sequential hot-deck (WSHD) method for the 2002 and 2003 surveys. However, partly because of the changes in the questionnaire, inconsistency problems were not easily resolved using the WSHD method when creating IRENTAG2. To alleviate this problem and to promote consistency with how imputations were conducted with other variables in NSDUH, imputations on the immigrant status variables IRBORNUS and IRENTAG2 have been conducted using the predictive mean neighborhood (PMN) method since the 2004 survey, which is discussed in Section 5.2.2. The variables IRBORNUS and IRENTAG2 were subsequently used to create recoded variables for the purposes of analysis. The recoded Hispanic/Latino group variable HISPGRP2, which was used specifically for the WSHD imputation of missing values of the immigrant status variables in the 2002 and 2003 surveys, was no longer required in the PMN method. Nevertheless, to maintain consistency with previous surveys, HISPGRP2 is still created as described in Section 5.2.3.

5.2.1 Edited Immigrant Status Variables

5.2.1.1 Born-in-U.S. Indicator (BORNINUS)

All respondents were asked in QD14 whether they were born in the United States (excluding U.S. territories). Responses were limited to "yes" or "no," and if the response was "no," the respondent was asked to name the country of origin in QD15. The edited variable BORNINUS was created using the responses to QD14. As part of the standard editing procedures, if the interviewer entered a U.S. State in QD15, the "no" in QD14 was overwritten with a logically assigned "yes." Other levels of BORNINUS were standard NSDUH missing data codes corresponding to "don't know," "refused," or "blank." More details about editing procedures are provided in a separate document (Kroutil & Chien, 2008).

5.2.1.2 Length of Time Lived in the United States (LIVUS1YR, LIVUSYRS, and LIVUSMOS)

As previously stated, the 2007 NSDUH questions recording the length of time that a non-U.S.-born respondent had lived in the United States changed from surveys prior to 2004. In those surveys, the length of time that non-U.S.-born respondents had lived in the United States was obtained from a single question (QD16). However, since the 2004 survey, respondents have been given the choice to write in the amount of time they had lived in the United States in years (in QD16b) or in months (in QD16c), depending upon their answer to QD16a (asking if they had lived in the United States for at least 1 year). The edited variables associated with QD16a, QD16b, and QD16b were called LIVUS1YR, LIVUSMOS, and LIVUSYRS, respectively. A legitimate skip code was assigned to LIVUSMOS if the respondent had lived in the United States for 1 year or more (LIVUS1YR = 1). Similarly, a legitimate skip code was assigned to LIVUSYRS, if the respondent had lived in the United States for less than 1 year (LIVUS1YR = 2). Codes for "don't know," "refused," "blank," and "bad data" also were applied to these variables at the editing stage. More editing details on these three variables are described by Kroutil and Chien (2008).

5.2.1.3 Continuous Age (CONTAGE) and Continuous Length of Time Lived in the United States (LNGTHLIV) for Non-U.S.-Born Respondents

In order to compute the age at which a non-U.S.-born respondent entered the United States, the continuous form of the respondent's age and length of time living in the United States was produced for all non-U.S.-born respondents. Because QD16b and QD16c were designed to be mutually exclusive, the edited variables LIVUSMOS and LIVUSYRS were combined to create a continuous estimate of how many years a non-U.S.-born respondent had lived in the United States: LNGTHLIV. In most cases, LNGTHLIV had the same value as LIVUSYRS. However, if the respondent had lived in the United States for less than 1 year, his or her LNGTHLIV values were obtained from LIVUSMOS by converting the number of months into fractions of 1 year. The variable was set to missing when LIVUSYRS and LIVUSMOS had missing data codes. CONTAGE, the continuous age variable, was defined as CONTAGE = (interview date – birth date + 1)/365.25. Because interview date and birth date, as described in Chapter 4, had no missing values, CONTAGE also had no missing values. A legitimate skip code of 999 was assigned to the respondents who were born in the United States for both LNGTHLIV and CONTAGE.

5.2.1.4 Age of Entry (ENTRYAG2)

The variable ENTRYAG2 is the base variable for creating the imputation-revised variable IRENTAG2 and represents the (continuous) age at which an immigrant entered the United States. ENTRYAG2 was defined as ENTRYAG2 = CONTAGE – LNGTHLIV and was set to missing if LNGTHLIV was missing. It also had a legitimate skip code (999) for respondents who were born in the United States.

5.2.2 Imputation-Revised Immigrant Status Variables

5.2.2.1 Imputation-Revised Born-in-U.S. Indicator (IRBORNUS)

As with all other demographic variables requiring imputation, except birth date, the PMN method was used to impute missing values in the born-in-U.S. indicator variable. Because the born-in-U.S. indicator was a single dichotomous discrete variable, the assignment of imputed values was performed using the univariate predictive mean neighborhood (UPMN) method, which is described in Appendix C. The UPMN method, as applied to IRBORNUS, is explained in the next four sections: setup for model building, computation of predicted means, assignment of imputed values, and constraints on UPMNs.

5.2.2.1.1 Setup for Model Building

Imputation of missing values in the born-in-U.S. indicator was conducted within three age categories: 12 to 17, 18 to 25, and 26 or older. The separate age groups were used more for ease of processing and consistency with other variables rather than because of any strong correlation between whether a respondent was born in the United States and age. Because all interview respondents were asked the question whether they were born in the United States, no subsetting of the data was necessary.

If a valid response ("yes" or "no") was provided for the born-in-U.S. measure, the person was defined as an item respondent. The weights were adjusted for item nonresponse using item response propensity models, one for each age group. The item response propensity model is a special case of the generalized exponential model (GEM),[35] which is described in Appendix B. The covariates in these models included gender, age categories, gender by age category interactions, imputation-revised race/ethnicity, imputation-revised education level, imputation-revised employment status, imputation-revised marital status, percentage owner-occupied households, core-based statistical area, and census region.

5.2.2.1.2 Computation of Predicted Means

After the weight adjustment, the probability of an affirmative response to the question for whether a respondent was born in the United States was modeled within each age group using logistic regression. The predictors included gender, centered age, centered age squared, centered age cubed, gender by centered age interaction, gender by centered age squared interaction, gender by centered age cubed interaction, imputation-revised race/ethnicity, imputation-revised education level, imputation-revised employment status, imputation-revised marital status, percentage owner-occupied households, core-based statistical area, and census region. The number of covariates was reduced if convergence or stability problems occurred in the model-fitting process. A summary of the final set of covariates used in the model can be found in Appendix F.

[35] The GEM macro, which was written in SAS/IML® software, was developed at RTI International (a trade name of Research Triangle Institute) for weighting procedures.

5.2.2.1.3 Assignment of Imputed Values

Separate assignments were performed within each of the three age groups: 12 to 17, 18 to 25, and 26 or older. The constraints used to select donors are described in the next section.

5.2.2.1.4 Constraints on UPMNs

No logical constraints were used in defining neighborhoods; only likeness constraints were utilized. In the first attempt to find a neighborhood for each item nonrespondent, two likeness constraints were used. The first likeness constraint stated that the donor must live in the same segment as the recipient. The second likeness constraint stated that the donor's predicted mean, as described in Section 5.2.2.1.2, must be within 5 percent of the recipient's predicted mean. If no item respondents met the above conditions for a particular item nonrespondent, the constraint on the segment of the potential donor was removed. If a potential donor could still not be found, the delta constraints were removed. In the 2007 survey, a donor was found for every item nonrespondent using this method. Therefore, no further loosening of constraints was necessary. See Appendix G for the number of respondents that met each set of likeness constraints on sets of eligible donors.

5.2.2.1.5 Imputation and Editing Summary for Born in the United States

Table 5.1 summarizes item nonresponse for the born-in-U.S. variable. The source information was recorded in the indicator variable IIBORNUS.

Table 5.1 IRBORNUS Editing and Imputation Summary

Value of IIBORNUS	Assignment of IRBORNUS	Frequency	Percent
1	From questionnaire	67,835	99.95
2	Logically assigned	4	0.01
3	Statistically imputed	31	0.05

5.2.2.2 Imputation-Revised Immigrant Age of Entry (IRENTAG2)

The PMN method was utilized for imputing missing values in the variable recording the age of entry into the United States of non-U.S.-born respondents. It followed the same general procedures as the imputation of other demographic variables. A linear regression model was fitted using a logit transformation of the respondent's age of entry as the response variable. Because the immigrant's age of entry was a single continuous variable, the UPMN method was used in the imputation-revised age-of-entry assignment. The UPMN method, as applied to IRENTAG2, is explained in the next four sections.

5.2.2.2.1 Setup for Model Building

All respondents who were born in the United States (IRBORNUS = 1) were assigned a legitimate skip code (IRENTAG2 = 999) and were excluded from the item response propensity model, the predictive mean model, and the set of potential donors. Imputations of missing values in the age-of-entry variable were not conducted separately within age groups, because the

number of non-U.S.-born respondents was too small to support quality models and sufficient donor pools in three separate age groups.

An interview respondent was considered an item nonrespondent for the age-of-entry variable if the edited variable ENTRYAG2 had missing data. The weights were adjusted for item nonresponse using item response propensity models to match the entire non-U.S.-born population. The covariates in these models included gender, age categories, gender by age category interactions, imputation-revised race/ethnicity, imputation-revised education level, imputation-revised employment status, imputation-revised marital status, percentage owner-occupied households, core-based statistical area, and census region.

5.2.2.2.2 Computation of Predicted Means

The predicted means for an immigrant's age of entry was estimated using a linear regression model. To control the upper and lower bounds of predicted means for age of entry, it was necessary to perform a logit transformation to the response variable. The response variable in the model was the immigrant age at entry as a proportion of the continuous version of current age CONTAGE, as described in Section 5.2.1.3. The expression of the proportion is $P_i = Y_i/N_i$, where $Y_i = Age\ at\ Entry_i$ and $N_i = Continuous\ Age_i$ (CONTAGE). After the weight adjustment, the following empirical logit transformation was used as the response variable in a weighted linear univariate regression:

$$\log\left[\left(Y_i + 0.5\right)/\left(N_i - Y_i + 0.5\right)\right].$$

This transformation was nearly equivalent to the standard logit transformation:

$$Y_i^* = \log\left[P_i/\left(1 - P_i\right)\right],$$

which was not used because it might be unstable for respondents who entered the country at their current age. Variables included in the regression model were gender, centered age, centered age squared, centered age cubed, gender by centered age interaction, gender by centered age squared interaction, gender by centered age cubed interaction, imputation-revised race/ethnicity, imputation-revised education level, imputation-revised employment status, imputation-revised marital status, percentage owner-occupied households, core-based statistical area, and census region. The number of covariates was reduced if convergence or stability problems occurred in the model-fitting process. A summary of the final set of covariates used in the model can be found in Appendix F.

5.2.2.2.3 Assignment of Imputed Values

The assignment was performed on the full sample and was not separated into age categories, for reasons given in Section 5.2.2.2.1. The constraints used to select donors are described in the next section.

5.2.2.2.4 Constraints on UPMNs

Two logical constraints and two likeness constraints were utilized in the definition of neighborhoods for IRENTAG2. Both logical constraints involved the respondent's age. One required that the donor's age of entry be less than the recipient's current age. The other logical constraint required that the difference between the recipient's current age and the donor's age of entry be less than 1 year if the recipient lived in the United States for less than 1 year (as indicated by QD16a) or greater than 1 year if the recipient lived in the United States for more than 1 year.

In the first attempt to find a neighborhood for each item nonrespondent, two likeness constraints were used. The first likeness constraint stated that the donor must live in the same segment as the recipient. The second likeness constraint stated that the donor's predicted mean, as described in Section 5.2.2.2.2, must be within 5 percent of the recipient's predicted mean. If no item respondents met the above conditions for a particular item nonrespondent, the constraint on the segment of the potential donor was removed. If a potential donor could still not be found, the delta constraints were removed. See Appendix G for the number of respondents that met each set of likeness constraints on sets of eligible donors.

5.2.2.2.5 Imputation and Editing Summary for Immigrant Age of Entry

The associated indicator variable for the imputation-revised immigrant age of entry was IIENTAG2. Table 5.2 summarizes item nonresponse for the age-of-entry variable.

Table 5.2 IRENTAG2 Editing and Imputation Summary

Value of IIENTAG2	Assignment of IRENTAG2	Frequency	Percent
1	From questionnaire	7,343	10.82
2	Logically assigned	4	0.01
3	Statistically imputed (including those imputed to IRBORNUS = 2)	59	0.09
9	Legitimate skip (BORNINUS = 2)	60,464	89.09

5.2.3 Recoded Hispanic/Latino Group Variable (HISPGRP2)

When the weighted sequential hot-deck method was used for the 2002 and 2003 surveys, two variables—HISPGRP2 and AGEADULT—were created specifically to aid in the imputation of missing values in the immigrant status variables. These variables were no longer needed when the PMN method was used starting with the 2004 survey. The variable AGEADULT, which has not been produced since the 2004 survey, would be equivalent to CATAG3, which is described in Chapter 4. The variable HISPGRP2 was created in the 2007 survey in the same way as in previous surveys; that is, it was derived from the variable IRHOGRP4. Some of the levels in IRHOGRP4 were collapsed to generate a more condensed version of the Hispanic/Latino group variable. As a result, HISGRP2 had four levels: 1 = Puerto Rican (IRHOGRP4 = 1), 2 = Mexican (IRHOGRP4 = 2), 3 = Other Hispanic/Latino (IRHOGRP4 = 3, 4, 5, 6, or 7), and 4 = Non-Hispanic/Latino (IRHOGRP4 = 99).

5.3 Current Employment Status

The edited employment status variables used to create EMPSTATY are described below. Sections 5.3.2 and 5.3.3 discuss the imputation procedure for EMPSTATY, and Section 5.3.4 discusses the creation of EMPSTAT4, a recoded version of EMPSTATY.

5.3.1 Edited Employment Status Variables

5.3.1.1 JBSTATR and WRKHRSUS

The main edited variable used to summarize the respondent's current work situation was JBSTATR, which was subsequently used to create EMPSTATY. This edited variable combined information from QD26, QD29, QD30, QD31, QD32, and QD33. The categories for JBSTATR are shown in Table 5.3. WRKHRSUS was an edited variable created from QD29, which asks, "Do you **usually** work 35 hours or more per week at **all** jobs or businesses?" WRKHRSUS was used in some cases to determine whether employed respondents were employed full time or part time. Both variables are described in more detail in Kroutil and Chien (2008).

Table 5.3 Categories of JBSTATR

Code	Employment Situation	Code	Employment Situation
1	Worked at full-time job, past week	**12**	No job: in school/training
2	Worked at part-time job, past week	**13**	No job: retired
3	Has job but out: vacation/sick/temp absence	**14**	No job: disabled for work
4	Has job but out: layoff, looking for work	**15**	No job: didn't want a job
5	Has job but out: layoff, not looking for work	**190**	Has full-time job, reason for not working unknown
6	Has job but out: waiting to report to new job	**191**	Has part-time job, reason for not working unknown
7	Has job but out: self-employed, no business past week	**199**	Has job, no further information
8	Has job but out: in school/training	**290**	No job, no further information
9	No job: looking for work	**299**	Other, not in labor force
10	No job: layoff, not looking for work	Remaining codes in the 900 series have their standard meanings in NSDUH: Don't know (994), Refused (997), Blank (998), Legitimate skip (999)	
11	No job: keeping house full time		

5.3.1.2 EDEMPY

The base variable EDEMPY, which was used to create the imputation-revised employment status variable EMPSTATY, was derived from JBSTATR and the edited variable WRKHRSUS in the following manner:

EDEMPY =

- 99, if the respondent is 12 to 14 years old; else
- 1 (full time), if JBSTATR = 1 or 190, or if JBSTATR = 3, 6, 7, 8, or 199 and WRKHRSUS = 1; else
- 2 (part time), if JBSTATR = 2 or 191, or if JBSTATR = 3, 6, 7, 8, or 199 and WRKHRSUS = 2; else
- 3 (unemployed), if JBSTATR = 4, 5, 9, or 10; else
- 4 (other), if JBSTATR = 11-15, 290, or 299; else
- 5 (part time or full time), if JBSTATR = 3, 6, 7, 8, or 199 and WRKHRSUS was missing (i.e., greater than 2); else
- missing.

5.3.2 Imputation-Revised Employment Status (EMPSTATY)

Missing values in the edited employment status variable EDEMPY were replaced with imputed values using a multivariate predictive mean neighborhood (MPMN) method. This method is described in Appendix C. The MPMN method was applied to employment status variables for the first time in the 2001 survey. It was enhanced in the 2002 survey to account for partial knowledge of employment status. The imputation procedure for employment status in the 2007 survey was similar to the procedures that have been used since the 2002 survey. The MPMN method, as applied to the employment status variable, is explained in the next four sections.

5.3.2.1 Setup for Model Building

Similar to the imputations that were performed on other demographic variables, imputations for employment status variables in the hot-deck stage of the PMN method were conducted separately within the same three age groups: 12 to 17, 18 to 25, and 26 or older. All respondents with AGE < 15 were assigned EMPSTATY = 99. Only interview respondents with AGE ≥ 15 were used in the models or were considered donors. At the modeling stage of PMN, two of these age groups were aggregated: 15 to 17 and 18 to 25.

An interview respondent was considered an item nonrespondent for employment status if his or her value for EDEMPY = 5 (employed, part time versus full time unclear) or missing. The weights of the item nonrespondents aged 15 or older were reallocated to the item respondents aged 15 or older. In the 2007 survey, the final analysis weights were used if they were available. However, because the final weight adjustments were not completed at the time of the

demographic imputations, the person-level sample design weights were adjusted to account for nonresponse at the household level using a simple ratio adjustment.[36] Respondents aged 15 to 25 were modeled separately from respondents aged 26 or older.[37] The initial set of covariates in the two models were the same: census region, imputation-revised race, gender, age categories, the interaction of age categories and gender, percentage Hispanic/Latino population, percentage black/African-American population, percentage American Indian/Alaska Native population, percentage Asian population, and percentage owner-occupied households.

5.3.2.2 Computation of Predicted Means

Using the adjusted weights, the probability of selecting each employment status category (employed full time, employed part time, unemployed, and other) was modeled using polytomous logistic regression.[38] The predictors included in the model for the respondents aged 15 to 25 were census region, imputation-revised race, gender, centered age, centered age squared, the interaction of centered age and gender, the interaction of centered age squared and gender, percentage Hispanic/Latino population, percentage black/African-American population, percentage American Indian/Alaska Native population, percentage Asian population, and percentage owner-occupied households. The predictors included in the model for the respondents aged 26 or older were census region, imputation-revised race, gender, centered age, centered age squared, centered age cubed, the interaction of centered age and gender, the interaction of centered age squared and gender, percentage Hispanic/Latino population, percentage black/African-American population, percentage American Indian/Alaska Native population, percentage Asian population, percentage owner-occupied households, and imputation-revised marital status. The number of covariates was reduced if convergence or stability problems occurred in the model-fitting process. A summary of the final set of covariates used in the model can be found in Appendix F. The predictive mean vector used in the imputation procedure had three elements (three predicted probabilities) corresponding to the first three levels of EDEMPY.

5.3.2.3 Assignment of Imputed Values

The imputations were performed separately within each of three age groups: 15 to 17, 18 to 25, and 26 or older. All constraints used to select donors are described in the next section.

5.3.2.4 Constraints on MPMNs

One logical constraint was used in defining neighborhoods for the employment status variable: if the recipient had EDEMPY = 5, the donor must have been employed either part time or full time (EDEMPY = 1 or 2).

---

[36] In subsequent text, the use of the word "weights" will refer to the ratio-adjusted design weights.

[37] The 15- to 17-year-old respondents were separated from the 18- to 25-year-old respondents in the stage where final imputed values were assigned. This separating of age groups was done because these two age groups had very different work patterns. However, in both the response propensity models and the predictive mean models, these two age groups were combined because of the insufficient number of 15- to 17-year-old working respondents.

[38] SAS®-callable SUDAAN® was used to fit all polytomous logistic regression models. Details about the polytomous logistic regression model and additional references can be found in the *SUDAAN Language Manual Addendum, Release 9.0.3* (RTI International, 2007). SAS software is a registered trademark of SAS Institute, Inc. SUDAAN is a registered trademark of Research Triangle Institute.

Conditional probabilities were used to take advantage of the partial information that was available. Recipients with EDEMPY = 5 were known to be employed. Instead of the usual three predicted means using the model's predicted probabilities directly, a single predicted mean was derived using a conditional probability, which was the probability that the recipient was employed full time, given that the respondent was employed. See Appendix G for more details on missingness patterns for employment status.

In addition to the logical constraint, three likeness constraints were used. In the first attempt to find a neighborhood for each item nonrespondent, the donor's age was required to be within 4 years of the recipient's age; the donor was required to live in the same segment as the recipient; and each of the donor's three predicted means (one predicted mean for recipients with EDEMPY = 5), as described in Section 5.3.2.2, were required to be within 5 percent of each of the recipient's three predicted means. If no item respondents met the above conditions for a particular item nonrespondent, the constraint on the donor's segment was removed. If still no donors were found, the delta constraints were removed. See Appendix G for the numbers of respondents meeting each set of likeness constraints on the sets of eligible donors.

5.3.3 Imputation and Editing Summary for Employment Status

See Table 5.4 for a summary of item nonresponse for employment status. The table shows the values of both the detailed imputation indicator II2EMSTY and the simpler indicator IIEMPSTY.

Table 5.4 EMPSTATY Editing and Imputation Summary

Assignment of EMPSTATY	Frequency	Percent	Value of IIEMPSTY	Value of II2EMSTY
From Questionnaire	56,807	83.70	1	1
Statistically Imputed (Unrestricted)	26	0.04	3	3
Statistically Imputed (Restricted to Full Time or Part Time)	15	0.02	3	4
Legitimate Skip (Respondent Was 12 to 14 Years Old)	11,022	16.24	9	9

5.3.4 Imputation-Revised Employment Status Recode (EMPSTAT4) and Indicators (II2EMST4 and IIEMPST4)

EMPSTAT4 was a direct recode of EMPSTATY and AGE. For respondents who were younger than 15 or older than 17, EMPSTAT4 and EMPSTATY were equivalent. For 15- to 17-year-olds, responses for EMPSTATY were overwritten with a code indicating that the respondent was too young to have his or her employment status recorded for the variable. This was the same code that was used for 12- to 14-year-olds for EMPSTATY (and EMPSTAT4).

The same relationship was held between both II2EMSTY and II2EMST4 and IIEMPSTY and IIEMPST4. II2EMSTY was equivalent to II2EMST4, and IIEMPSTY was equivalent to IIEMPST4 for respondents younger than 15 or older than 17. For respondents aged 15 to 17, II2EMST4 = IIEMPST4 = 9.

6. Drugs

6.1 Introduction

Major changes were introduced in the imputation procedures for the drug use variables in the computer-assisted interviewing (CAI) sample of the 1999 National Household Survey on Drug Abuse (NHSDA), which was renamed the National Survey on Drug Use and Health (NSDUH) in 2002.[39] In particular, for the CAI sample of the 1999 survey, a new imputation methodology (i.e., predictive mean neighborhood [PMN]) was developed specifically for NSDUH. This methodology is a combination of weighted regression and nearest neighbor hot-deck imputation, where the hot deck is random whenever possible.[40] Its application to the drug use variables for the 2007 survey was similar to that of previous survey years, as is explained in the following sections.

This chapter describes how the PMN method was applied to the drug use variables. In some cases, imputations were required because the respondent did not answer a given question. However, other responses were altered in the editing process because of inconsistencies. In these cases, the original response was set to missing, or, in the case of recency of use, a specific recency was edited to a more general recency that was consistent with other responses, and determination of the specific recency was left to imputation. For example, a recency-of-use response might be edited to past year usage, where past month use versus past year but not past month use could be determined by imputation. These editing processes are summarized in the following two reports (Kroutil & Handley, 2008; Kroutil, Handley, Felts, Bradshaw, & Chien, 2008).

The models for these imputations, which are described in the following sections, were either weighted logistic regression models (binomial or multinomial) or weighted multiple linear regression models with the response variable appropriately transformed. Using the PMN method, the predicted means from these models were used to determine neighborhoods from which donors were randomly selected for the final assignment of imputed values. If no donors were available within a very small distance of the recipient's predicted mean, the donor with the closest predicted mean was chosen. The neighborhoods were created based on a single predicted mean (a univariate predictive mean neighborhood [UPMN]) or using several predicted means at once (a multivariate predictive mean neighborhood [MPMN]). Even if the neighborhood was constructed from a univariate predicted mean, the assignment of imputed values could be either univariate or multivariate. The members of the neighborhood were restricted to satisfy two types of constraints: logical constraints and likeness constraints. Constraints that made the imputed values consistent with preexisting values of other variables were called logical constraints and were required for the candidate donor to be a member of the neighborhood. Likeness constraints were implemented to make donors and recipients as much alike as possible. Although logical

[39] This report presents information from the 2007 National Survey on Drug Use and Health (NSDUH), an annual survey of the civilian, noninstitutionalized population of the United States aged 12 or older. Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA).

[40] The nearest neighbor hot deck is described in Appendix A.

constraints could not be loosened, likeness constraints could be loosened if they forced the donor pool to be too sparse. Details of these PMN imputation procedures are provided in Appendix C.

In the 2007 survey, because drug use was highly correlated with age, and to facilitate easier implementation of the imputation procedures, the model building and final assignment of imputed values for all drug use variables were performed separately within three distinct age groups: 12 to 17, 18 to 25, and 26 or older.[41]

Although statistical imputation of the drug use variables could not proceed separately within each State because of insufficient pools of donors, information about the State of residence of each respondent was incorporated in the modeling and hot-deck steps in the sample. States were classified into three drug usage categories within each age group: States with high usage of a given drug were placed into one category, States with medium usage into another, and the remainder into a third category. Respondents were then assigned values for a three-level "State rank" variable, depending on their State of residence. The indicator variables resulting from this categorical State-rank variable were used as covariates in the imputation models. In addition, for all of the drug use measures, eligible donors for each item nonrespondent were restricted, if possible, to be from States with the same level of usage (the same State rank) as the item nonrespondent. The definition of "level of usage" (i.e., which measure of usage was used to categorize the States) depended on the drug use measure being imputed.

As with the CAI instruments used in the 1999 through 2006 surveys, the 2007 survey had several different drugs and drug use measures than are found in pre-1999 surveys. Table 6.1 summarizes the drugs and drug use measures that were imputed and whether the imputations were univariate or multivariate. If no symbol is present in Table 6.1, then no information regarding that particular drug use measure was gathered in the questionnaire for the given drug.

6.2 Hierarchy of Drugs and Drug Use Measures

The first step in the imputation process was to determine the order in which drugs and drug use measures were to be modeled so that drugs and drug use measures earlier in the sequence could be used, if applicable, as covariates for models fitted later in the sequence. Because the lifetime indication of use questions (see Table 6.3) in the 2007 survey were the basis for all subsequent drug data, it was necessary that the imputation of missing values for lifetime drug use for all drugs preceded imputations of all other drug use measures. These lifetime use indicators were temporary in the sense that they were created within the drug recency and frequency-of-use variables, but they were not included in the public use file themselves. The hierarchy of models for drugs for the lifetime usage models is discussed in Section 6.3.1.

Once all the lifetime usage indicators had been determined, the imputations of the remaining measures proceeded. Where indicated in Table 6.1, a multivariate imputation was implemented within each drug for recency of use, 12-month frequency of use, 30-day frequency

[41] The modeling procedures were performed separately within each of the three age groups, regardless of the response variable.

of use, and binge drink 30-day frequency (alcohol only). For a given drug, recency of use[42] was included in the model for frequency of use, 12-month frequency of use was included in the model for 30-day frequency, and 30-day frequency of use of alcohol was included in the model for the binge drink frequency variable.

Table 6.1 Drugs and Drug Use Measures, Univariate versus Multivariate Imputation[1]

Drug	Drug Use Measure						
	Lifetime Usage	Recency of Use	12-Month Frequency of Use	30-Day Frequency of Use	Binge Drink Frequency	Age at First Use	Age at First Daily Use
Cigarettes	✓✓	×		×		✓	✓
Smokeless Tobacco[2]	✓✓	××		××		✓×	
Cigars	✓✓	×		×		✓	
Pipes	✓✓	✓					
Alcohol	✓✓	×	×	×	×	✓	
Inhalants	✓✓	×	×	×		✓	
Marijuana	✓✓	×	×	×		✓	
Hallucinogens[3]	✓✓	××	××	××		✓×	
Pain Relievers[4]	✓✓	××	××			✓×	
Tranquilizers	✓✓	×	×			✓	
Stimulants[5]	✓✓	××	××			✓×	
Sedatives	✓✓	×	×			✓	
Cocaine and Crack	✓✓	××	××	××		✓×	
Heroin	✓✓	×	×	×		✓	

✓ Univariate neighborhood; univariate assignment of imputed values.
✓✓ Multivariate neighborhood across all lifetime drug use variables; multivariate assignment of imputed values across all lifetime drug use variables.
× Multivariate neighborhood across recency of use, 12-month frequency of use where applicable, 30-day frequency of use where applicable, and the 30-day binge drink frequency variable (alcohol only); multivariate assignment of imputed values across measures.
×× Multivariate neighborhood across recency of use, 12-month frequency of use where applicable, and 30-day frequency of use where applicable; multivariate assignment of imputed values across measures and across certain drugs (see Section 6.5.5.1.3).
✓× Univariate neighborhood and multivariate assignment of imputed values (see Section 6.7.1.7).

[1] If no symbol is present, then no information regarding that particular drug use measure was available for the given drug.
[2] Includes chewing tobacco and snuff.
[3] Includes LSD, PCP, and Ecstasy.
[4] Includes OxyContin.
[5] Includes methamphetamine.

[42] Missing values were replaced by imputed values in these recency-of-use and frequency-of-use variables. The imputed values were provisional because the final values were not known until the multivariate imputation, after the completion of the modeling.

Finally, age at first use was required to be consistent (in a number of ways) with the other measures (see Section 6.7). Hence, age at first use was imputed after the imputation for the other measures was completed.[43] The following sections describe the imputation procedures for each drug use measure.

Some of the rows in Table 6.1 refer to both a general drug category and one or more subcategories. In the remainder of this chapter, to highlight the relationship between them, these drugs are described using the terms "parent drug" for the general drug category and "child drug" for the drug subcategory. For a drug to be considered a child drug, data must have been gathered on some combination of recency, frequency, and age at first use. Parent/child drug pairs sometimes occurred in modules that included "subgate" questions. However, they also could appear in separate modules. The parent/child drug combinations included smokeless tobacco (parent) and chewing tobacco and snuff (child); cocaine (parent) and crack (child); hallucinogens (parent) and LSD, PCP, and Ecstasy (child); pain relievers (parent) and OxyContin (child); and stimulants (parent) and methamphetamine (child). Smokeless tobacco differs from the other parent drugs in that data were not collected on this drug. Respondents were asked about only the two child drugs (chewing tobacco and snuff). Any measures of smokeless tobacco can be considered as recoded variables because they were not directly imputed. Table 6.2 illustrates all the drugs in parent/child relationships and the data that were gathered on them.

Table 6.2 Drugs in a Parent/Child Relationship

Parent Drug	Child Drug	Parent Data Collected	Child Data Collected	"Other" Lifetime Use Indicator[1]
Smokeless Tobacco	**Chewing Tobacco, Snuff**	None	Recency, 30-day frequency, age at first use	No
Hallucinogens	**LSD, PCP, Ecstasy**	Recency, 12-month frequency, 30-day frequency, age at first use	Recency, age at first use	Yes
Pain Relievers	**OxyContin**	Recency, 12-month frequency, age at first use	Recency, 12-month frequency, age at first use	Yes
Stimulants	**Methamphetamine**	Recency, 12-month frequency, age at first use	Recency, 12-month frequency, age at first use	Yes
Cocaine	**Crack**	Recency, 12-month frequency, 30-day frequency, age at first use	Recency, 12-month frequency, 30-day frequency, age at first use	No

[1] See Section 6.3.7.3.

[43] For cigarettes, both age at first use and age at first daily use had to be consistent with the other measures. Hence, age at first use was imputed after the other measures, followed by the imputation of age at first daily use.

6.3 Imputing Lifetime Drug Use Indicators

As with the 1999 through 2006 surveys, the 2007 survey implemented automatic routing of the respondent through the questionnaire. Using a series of gate questions, the instrument asked the respondent whether he or she had ever used a number of drugs in his or her lifetime. Based on the response to each gate question, the instrument either routed the respondent through the current drug module or skipped him or her to the next module. Thus, the respondent was not necessarily required to answer all questions in the questionnaire. The respondent could have skipped a module if he or she either indicated nonusage of the drug in the gate question or did not answer the gate question. Therefore, the gate question response was crucial to the range of responses available for subsequent questions in each module.

6.3.1 Hierarchy of Drugs

Because PMN was used for the lifetime usage imputations, a drug hierarchy was required, the use of which was motivated in general for PMN. As stated in Section 6.2, this hierarchy allowed drug use measures earlier in the sequence to be used as covariates for models fitted later in the sequence. Experience from past survey years has indicated a substantial correlation between lifetime drug use indicators. Although models were built using respondents with complete data across all the drugs, predicted means were calculated for both item respondents and nonrespondents for lifetime use. When calculating the predicted means for the lifetime usage of a given drug for respondents who did not answer all the lifetime usage questions, a predictor value could have been missing. Hence, it was sometimes necessary to use imputed lifetime usage values. These imputed values were provisional because the final imputed lifetime usage indicators were not known until the final multivariate imputation, after the completion of the modeling.

Therefore, the first step in the imputation of lifetime indicators was to determine the order in which the drugs would be modeled, where drugs later in the sequence would have more predictors in their models. The order in which the lifetime indicators of use were imputed is shown in Table 6.3.

6.3.2 Setup for Model Building and Hot-Deck Assignment

Once the hierarchy of drugs was established, the next step was to define respondents, nonrespondents, and the item response mechanism. As stated earlier, imputations for all drug use measures were conducted separately within the three age groups: 12 to 17, 18 to 25, and 26 or older. For an individual to be considered a lifetime-use item respondent, he or she must have complete data within each age group for all of the drug module gate questions: cigarettes; cigars; chewing tobacco; snuff; pipes; alcohol; marijuana; cocaine; crack; heroin; inhalants; LSD; PCP; Ecstasy; hallucinogens other than LSD, PCP, and Ecstasy; OxyContin; pain relievers other than OxyContin; tranquilizers; methamphetamine; stimulants other than methamphetamine; and sedatives. Response propensity adjustments were then computed for each age group to make the item respondent weights representative of the entire sample. (Because the modeling of the final weight adjustments was not completed at the time of the drug imputations, the person-level sample design weights were adjusted to account for nonresponse at the household level using a

simple ratio adjustment.)[44] An adjustment was calculated that reallocated weights from item nonrespondents to item respondents. Because item respondents were defined across all drugs, this adjustment was computed only once per age group and then used in the modeling of lifetime use for all drugs. The item response propensity model is a special case of the generalized exponential model (GEM),[45] which is described in Appendix B.

Table 6.3 Lifetime Indication of Use (Gate) Questions (in Order of Imputation)[1]

Drug	Questions
Cigarettes	CG01
Smokeless Tobacco[2]	CG17, CG25
Cigars	CG34
Pipes	CG42
Alcohol	AL01
Inhalants	IN01a, IN01b, IN01c, IN01d, IN01e, IN01f, IN01g, IN01h, IN01i, IN01j, IN01l
Marijuana	MJ01
Hallucinogens[3]	LS01a, LS01b, LS01c, LS01d, LS01e, LS01f, LS01h
Pain Relievers[4]	PR01, PR02, PR03, PR04, PR04a, PR05
Tranquilizers	TR01, TR02, TR03, TR04, TR04a, TR05
Stimulants[5]	ST01, ST02, ST03, ST04, ST04a, ST05
Sedatives	SV01, SV02, SV03, SV04, SV04a, SV05
Cocaine	CC01
Crack	CK01
Heroin	HE01

[1] Follow-up questions also were considered in the lifetime imputation.
[2] Includes chewing tobacco (CG17) and snuff (CG25).
[3] Includes LSD (LS01a), PCP (LS01b), and Ecstasy (LS01f).
[4] Includes OxyContin (option 12 in PR04a).
[5] Includes methamphetamine (ST01).

For certain categories of drugs, multiple gate questions within a drug module were used to assess lifetime use or nonuse of the overall group of drugs within that module (e.g., LSD, PCP, Ecstasy, and a number of other substances within the drug module for hallucinogens were used to assess usage of hallucinogens). For these drug groups, if any of the gate questions were answered "yes" (i.e., the respondent indicated using the drug once or more in his or her lifetime), then the lifetime use indicator for the overall drug group was set to "yes." For example, to assess lifetime use of the overall drug group "inhalants," the respondent was asked through 11 different questions if he or she had ever, even once, inhaled any of the following with the intention of getting high: (1) amyl nitrite, "poppers," locker room odorizers, or "rush"; (2) correction fluid, degreaser, or cleaning fluid; (3) gasoline or lighter fluid; (4) glue, shoe polish, or toluene; (5)

[44] In subsequent text, the use of the word "weights" will refer to the ratio-adjusted design weights.
[45] The GEM macro, which was written in SAS/IML® software, was developed at RTI International (a trade name of Research Triangle Institute) for weighting procedures.

halothane, ether, or other anesthetics; (6) lacquer thinner or other paint solvents; (7) lighter gases, such as butane or propane; (8) nitrous oxide or "whippets"; (9) spray paints; (10) some other aerosol spray; and (11) any other inhalant. If the response to any of these questions was "yes," the respondent was deemed a lifetime user of inhalants, even if some of the other responses to the gate questions in the inhalants module were unanswered. Similarly, composite lifetime indications of use were formed for hallucinogens, pain relievers, tranquilizers, stimulants, sedatives, and smokeless tobacco. To be considered a lifetime nonuser of a drug module with multiple gate questions, the respondent had to answer "no" to all of the gate questions. If none of the gate questions in a drug module was answered affirmatively, but some of the gate questions were unanswered, the individual was considered a nonrespondent for that module.

6.3.3 Sequential Model Building

Starting with cigarettes, the probability of lifetime use of each drug was modeled for item respondents within each age group, using the nonresponse-adjusted weights. Logistic regression[46] was used to determine the parameter estimates. The predictors in each model included lifetime use of drugs already imputed; centered age;[47] centered age squared; centered age cubed; gender; race/ethnicity; first-order interactions of centered age, race/ethnicity, and gender; a three-level State-rank variable (incorporating the proportion of lifetime users of the drug of interest in the respondent's State of residence); population density; and census region.[48] For the age groups 18 to 25 and 26 or older, the variables for marital status, education level, and employment status also were included. For a complete summary of the lifetime use imputation models, see Appendix F.

6.3.4 Computation of Predicted Means and Creation of Univariate Predictive Mean Neighborhoods

Using the parameter estimates from the probability of lifetime usage model for a given drug, predicted probabilities of use were computed for both item respondents and nonrespondents. These predicted values were then used to temporarily impute a value for each nonrespondent, using the UPMN imputation method described in Appendix C. Although models were built using respondents with complete data across all drugs, predicted probabilities were required for all respondents. In order to use lifetime usage of a given drug as a predictor for a drug later in the sequence, it was therefore necessary to utilize these temporary imputed values in cases where the original lifetime usage indicator was missing. If possible, provisional donors

[46] SAS®-callable SUDAAN® was used to fit all binomial and polytomous logistic regression models. Details about the logistic regression model and additional references can be found in RTI International (2007). SAS software is a registered trademark of SAS Institute, Inc. SUDAAN is a registered trademark of Research Triangle Institute.

[47] The covariate age was centered within each age group in order to reduce the effects of multicollinearity, particularly with the squared and cubed age terms. For more information on "centering" and "multicollinearity," refer to Draper and Smith (1981).

[48] These variables were included in every model unless convergence problems arose. If this occurred, the model was reduced.

were chosen with predicted means within the delta of the recipient,[49] where the value of delta varied depending on the value of the predicted means, which in this case were predicted probabilities of lifetime use. In particular, delta was defined as 5 percent of the predicted probability if the probability was less than 0.5, and it was defined as 5 percent of 1 minus the predicted probability if the probability was greater than 0.5. This allowed a looser delta for predicted probabilities close to 0.5, and it allowed a tighter delta for predicted probabilities close to 0 or 1. The range of values for delta across various predicted probabilities is shown in Table 6.4. If no donors were available with predicted means within delta of the recipient, the neighborhood was abandoned and the donor with the closest predicted mean was chosen.

Table 6.4 Values of Delta for Various Predicted Probabilities of Lifetime Use

Predicted Probability (p)	Delta
$p \leq 0.5$	$0.05p$
$p > 0.5$	$0.05(1 - p)$

6.3.5 Assignment of Provisional Imputed Values

Subject to the constraints described in the next section, separate assignments of provisional values were performed within each of the three age groups. The final lifetime imputations were multivariate across lifetime drug use variables and are further described in Section 6.3.8.

6.3.6 Constraints on Univariate Predictive Mean Neighborhoods

In a general UPMN imputation, the neighborhood is restricted by two types of constraints: (1) logical constraints (which cannot be loosened) to make imputed values consistent with a nonrespondent's preexisting nonmissing values of other variables, and (2) likeness constraints (which can be loosened) to make candidate donors in the neighborhood as similar to recipients as possible. The next paragraph discusses the likeness constraints and the order in which they were loosened or removed.

As with all other drug use measures, neighborhoods for lifetime use indicators were restricted so that candidate donors and recipients would be within the same age group (12 to 17, 18 to 25, and 26 or older). Models were built separately within these three groups, so this likeness constraint was never loosened. A small delta also could be considered a likeness constraint, which could be loosened by enlarging or removing delta. As previously stated, if no donors were found in the delta, the neighborhood was abandoned and the donor with the predicted mean closest to the recipient was chosen.[50] If possible, donors and recipients were required to be from States with the same level of usage of a given drug (State rank), where the

[49] "Delta" refers to the value that defined the neighborhood of donors that were "close" to the item nonrespondent. The difference between the predicted mean of the item nonrespondent and the predicted means of the item respondents in the neighborhood must be less than delta. See Appendix C for more details.

[50] Although using neighborhoods is important for the calculation of the variance due to imputation, methods to account for donor-predicted means differing greatly from recipient-predicted means had not been devised at the time these imputations were implemented.

level of usage was defined in terms of the weighted proportion of a given State's residents who were lifetime users of the drug.[51] An additional likeness constraint required the donor to match the recipient on any nonmissing lifetime use indicators for child drugs. For example, if the lifetime use indicator for overall hallucinogens was missing, but the recipient was known to be a lifetime nonuser of LSD, then the donor must also have been a lifetime nonuser of LSD. If insufficient donors were available within these constraints, they were loosened in the following order: (1) the neighborhood was abandoned and the donor with the closest predicted mean was chosen; and (2) both the State-rank and child lifetime drug indicator constraints were removed, and the delta constraint was reapplied.

No logical constraints were placed on the neighborhoods for any of the lifetime usage indicators. Occasionally, more than one substance was associated with a single predicted mean, leading to a multivariate assignment of imputed values. Even in those cases, however, the imputation was performed so that no logical constraints were necessary.

6.3.7 Multivariate Assignments

Although the methodology for determining the nearest neighbor neighborhood was univariate in terms of the predicted probability of lifetime use, peculiarities associated with drugs in parent/child pairs sometimes required the assignment step to be multivariate. These drugs are discussed in the following sections.

6.3.7.1 Smokeless Tobacco (Chewing Tobacco and Snuff)

Many respondents who indicated lifetime use of smokeless tobacco seemed to be confused regarding the difference between chewing tobacco ("chew") and snuff, as was demonstrated by their responses to questions regarding specific brands. For example, many respondents who indicated use of chewing tobacco entered a snuff brand, such as Copenhagen™, when asked about the specific brand of chew they used. As a result, one model for smokeless tobacco (a combination of the chew and snuff responses) was fitted, rather than individual models for chew and snuff. The nearest neighbor hot-deck neighborhood was then based on the overall smokeless tobacco predicted probability of lifetime use. Missing values for chew and/or snuff were replaced with the values from a donor within this neighborhood. For individuals missing the lifetime usage indicator for either chew or snuff but not both, only the missing value was replaced. However, for individuals missing both chew and snuff, both lifetime usage indicators were replaced by values from the same donor. No logical constraints were necessary in the assignment step. This was because chew and snuff were assigned values independently, and then combined at the end to form a final lifetime usage indicator for smokeless tobacco.

6.3.7.2 Cocaine and Crack

Because cocaine and crack were in distinct modules in the 2007 NSDUH questionnaire, separate models were fitted for the two substances. However, crack is a type of cocaine, so donors for the two substances were obtained using a single neighborhood. This neighborhood was defined in terms of the deltas shown in Table 6.4, which were based on the predicted probabilities of lifetime use for both cocaine and crack. An item respondent was eligible to be a

[51] Those with a missing lifetime use indicator for the drug were treated as lifetime nonusers.

donor for a given item nonrespondent if his or her predicted probability of lifetime cocaine use was within delta of the item nonrespondent's cocaine predicted probability and his or her predicted probability of lifetime crack use was within delta of the item nonrespondent's crack-predicted probability. This was true regardless of whether the item nonrespondent was missing only crack, only cocaine, or both crack and cocaine.[52] Once the neighborhood was defined, missing values for crack and/or cocaine were replaced with the values from a donor within this neighborhood. For individuals missing a lifetime usage indicator for only crack, or only cocaine, but not both crack and cocaine, only the missing value was replaced. However, for individuals missing both crack and cocaine, both lifetime usage indicators were replaced by values from the same donor.

6.3.7.3 Hallucinogens (LSD, PCP, Ecstasy, and "Other" Hallucinogens), Pain Relievers (OxyContin and "Other" Pain Relievers), and Stimulants (Methamphetamine and "Other" Stimulants)

The hallucinogens, pain relievers, and stimulants modules all included multiple gate questions (called "subgate questions"), and some of the substances referenced in the subgate questions were child drugs. For hallucinogens, there were three child drugs: LSD, PCP, and Ecstasy. For pain relievers, there was one child drug: OxyContin. For stimulants, there was also one child drug: methamphetamine.

Predicted probabilities were calculated for the parent drugs, and these probabilities were used to determine neighborhoods for each group of drugs. An "other" category was created by combining all the other subgate questions with the exception of the ones referring to the child drugs. In the final assignment step, lifetime usage indicators were assigned for LSD, PCP, Ecstasy, and "other" hallucinogens; OxyContin and "other" pain relievers; and methamphetamine and "other" stimulants. The final lifetime usage indicators for the parent drugs were created by combining the constituent parts, including the "other" group of substances.

6.3.7.3.1 Hallucinogens

The lifetime usage indicator for "other hallucinogens" was created using the lifetime usage information from all the hallucinogens' subgate questions except LSD, PCP, and Ecstasy. It is important to note that if a respondent was a user of at least one of the other hallucinogens, he or she was considered a user of other hallucinogens, even if some of the other hallucinogens' subgate questions were unanswered. A missing value for other hallucinogens arose if at least one of the other hallucinogens' subgate questions was unanswered and all the other hallucinogens' subgate questions that were answered had a negative response. Using the neighborhood created from the hallucinogens' predicted probability of lifetime use, missing values for LSD and/or PCP and/or Ecstasy and/or other hallucinogens were replaced with the values from a donor within this neighborhood. For individuals missing a lifetime usage indicator for LSD and/or PCP and/or Ecstasy and/or other hallucinogens, only the missing value(s) was replaced. For individuals missing two or more of these lifetime usage indicators, the missing values were replaced by

[52] It would seem impossible for an individual to be missing the lifetime usage indicator for cocaine, but not for crack, because respondents who fail to respond to the cocaine lifetime use question never proceed to the crack module. However, because of editing rules, two cases in the 2007 survey were actually considered lifetime nonusers of crack but missing for cocaine.

values from the same donor. As with smokeless tobacco, the subcategories for hallucinogens were assigned values separately, making logical constraints unnecessary. As a final step, a lifetime usage indicator for the parent drug was created by combining the lifetime usage indicators for the three subgroups.

6.3.7.3.2 Pain Relievers

The procedure for pain relievers was similar to the procedure used for hallucinogens. The major difference is that there was no subgate question focusing solely on the specific child drug OxyContin. Specifically, OxyContin was one of 18 types of pain relievers, which appeared both in question PR04a and on a card shown to the respondents by the interviewers when the respondents reached these questions:

PR04 Please look at the pain relievers shown **below** the red line on Card A.

Have you ever, even once, used any of these pain relievers when they were **not** prescribed for you or that you took only for the experience or feeling they caused?

1 Yes
2 No

PR04a [IF PR04 = 1] Which of the pain relievers shown **below** the red line on Card A have you used when they were **not** prescribed for you or that you took only for the experience or feeling they caused?

To select more than one drug from the list, press the space bar between each number you type. When you have finished, press [ENTER].

4 Codeine
5 Demerol
6 Dilaudid
7 Fioricet
8 Fiorinal
9 Hydrocodone
10 Methadone
11 Morphine
12 OxyContin
13 Phenaphen with Codeine
14 Propoxyphene
15 SK-65
16 Stadol
17 Talacen
18 Talwin
19 Talwin NX
20 Tramadol
21 Ultram

Respondents could have selected any number of drugs listed on the card. A lifetime usage indicator for "other pain relievers" was created using information from all the pain relievers' subgate questions, except the OxyContin item in PR04a. As with hallucinogens, a respondent's other pain relievers' lifetime usage indicator was missing only if the subgate questions, other than the item that dealt with OxyContin, were all unanswered or if these questions were a combination of unanswered questions and "no" responses. Using the neighborhood created from the pain relievers' predicted probability of lifetime use, the missing value(s) for OxyContin and/or other pain relievers was replaced with the value(s) from a donor within this neighborhood. For individuals missing a lifetime usage indicator for either OxyContin or other pain relievers but not both, only the missing value was replaced. For individuals missing both of these lifetime usage indicators for pain relievers, the missing values were replaced by values from the same donor. As with smokeless tobacco, the subcategories for pain relievers were assigned values separately, making logical constraints unnecessary. As a final step, a lifetime usage indicator for the parent drug was created by combining the lifetime usage indicators for the two subgroups.

6.3.7.3.3 Stimulants

The procedure for stimulants was almost identical to the procedure used for pain relievers. However, as for hallucinogens, there was a specific subgate question on the child drug methamphetamine. Three lifetime usage indicators were created: one for "other stimulants," one for methamphetamine, and one for all stimulants.

6.3.8 Multivariate Imputation for Lifetime Drug Use

Section 6.3.2 summarized how all of the respondents in the 2007 survey were separated into item respondents and item nonrespondents for the lifetime drug variables. Subsequent sections summarized model building, computation of predicted means and delta neighborhoods, and the assignment of imputed values for these measures using a univariate predicted mean. In most cases, however, these univariate assignments were only provisional. As indicated in Table 6.1, the final imputed values for these drug use measures were obtained by building neighborhoods upon a vector of predicted means using the MPMN method described in Appendix C. In a manner consistent with the univariate imputations, the multivariate assignments were done separately within three age groups: 12 to 17, 18 to 25, and 26 or older. As described in earlier sections, a respondent was eligible to be a donor for a given item nonrespondent if he or she had complete data across all the lifetime drug use variables and was within the same age group.

The values missing for a given respondent define the "pattern of missingness." Respondents with missing lifetime indicators were separated into two groups: respondents missing only one lifetime drug use measure and respondents missing more than one lifetime drug use measure. The respondents missing only one lifetime use indicator were imputed using UPMN. Respondents missing more than one lifetime use indicator were imputed using MPMN.

Only one logical constraint was utilized in the multivariate imputation of lifetime use. Those item nonrespondents who were known to have used pain relievers, but both their OxyContin and "other" pain reliever indicators were missing, were required to have a donor who was a lifetime user of pain relievers. This pattern of nonresponse occurs when respondents respond affirmatively to PR04 but fail to select any drugs from the card in PR04a.

In addition, if possible, donors and recipients were required (as likeness constraints) to come from States with similar drug usage patterns for illicit drugs, and donors were required to have each element of the multivariate predictive mean vector "close to" (i.e., within the delta distance of) the recipient's elements of the predictive mean vector. Because the imputation was multivariate, the set of deltas was also multivariate, where a different delta corresponded to each element of the predictive mean vector. The elements of the predictive mean vector corresponded to the predicted values of the recipient's missing lifetime use indicators. Initially, donors and recipients were required to have, if possible, the same values for all nonmissing lifetime use indicators. If this initial constraint did not produce a big enough donor pool, donors and recipients were required to have the same values for only lifetime indicators within the same or related drug modules. The number of respondents for whom donors were found within various likeness constraints is summarized in Appendix G. In general, the likeness constraints were loosened in the following order: (1) remove the requirement that donors and recipients have the same values for all nonmissing lifetime usage indicators; (2) remove the requirement that donors and recipients have the same values for all nonmissing lifetime usage indicators only within a common or related drug module; (3) abandon the neighborhood and choose the donor with the closest predicted mean; and (4) remove the requirement that donors and recipients be from States with similar illicit drug usage levels.

The full predictive mean vector contained elements for each lifetime drug use measure. However, only a portion of the full predictive mean vector was used. Specifically, only those elements corresponding to the recipient's missing lifetime drug use were used. If the missing lifetime usage indicators corresponded to only one predicted mean, a UPMN imputation similar to the provisional UPMN was utilized. Otherwise, an MPMN imputation was employed. The Mahalanobis distance[53] was then calculated using only the portion of the predictive mean vector associated with the given missingness pattern. If no donors were available who had predicted means within a multivariate delta of the recipient's vector of predicted means, the neighborhood was abandoned and the respondent with the closest Mahalanobis distance was selected as the donor. The procedure is described in Appendix C.

No final imputation-revised variables indicating lifetime usage alone were created, because this information was recorded in the final imputation-revised recency-of-use variables. Imputation indicators also were not created, though temporary variables indicating that lifetime usage was imputed were maintained to inform the creation of the recency-of-use imputation indicators.

6.4 Editing of Drug Recency of Use, 30-Day Frequency of Use, and Age at First Use

Most of the editing procedures that were applied to the raw data on recency of use, frequency of use, and age at first use are discussed in the 2007 NSDUH editing and coding reports (Kroutil & Handley, 2008; Kroutil et al., 2008). However, a few edits were implemented just before imputation and are discussed below. In general, these edits affected only a few records. They were implemented mostly to resolve inconsistencies, which prevented the

[53] See Appendix C for a definition of Mahalanobis distance. A definition also can be found in Manly (1986).

determination of a valid interval for the assignment of date of first use (see Section 6.7.1.8). There are other edits that could have been implemented, but were not implemented for one of the following reasons:

1. The pattern of inconsistency was not discovered until after processing began.
2. It was decided that the effort required to implement the edit exceeded the benefit derived from this edit.
3. No decision had been made on whether to implement the edit by the time processing began.

6.4.1 Edits Involving "Other" Hallucinogens, "Other" Pain Relievers, and/or "Other" Stimulants

For respondents who were known to have never used "other" hallucinogens, "other" pain relievers, and/or "other" stimulants, certain deductions could be made regarding the relationship between the parent drug data and the child drug data. Note that these edits also could have been applied to respondents who were imputed to lifetime nonuse of the "other" variable.

1. If the respondent was known never to have used "other" hallucinogens, the overall hallucinogens recency was missing, and none of the recencies for the child drugs were missing, then the overall hallucinogens recency was assigned to the most recent of the child recencies. (This also was applied for pain relievers and stimulants.)
2. If the respondent was known never to have used "other" hallucinogens, the overall hallucinogens recency was past month, one of the child recencies was past year (where past month vs. not past month use could not be determined), and no other child recency was past month, then the child recency that was past year (where past month vs. not past month use could not be determined) was edited to past month.
3. If the respondent was known never to have used "other" hallucinogens (or pain relievers or stimulants), the parent age at first use was nonmissing, only one child age at first use was missing, and the minimum of the nonmissing child ages at first use was greater than the parent age at first use, then the missing child age at first use was edited to the parent age at first use.

6.4.2 Edits Applied to Respondents Imputed to Lifetime Use of Child Drug(s)

Once the imputation of the lifetime use indicators was complete, certain edits that were applied to the raw recency and frequency data had to be reapplied. The list of these edits is below:

1. If the parent drug recency of use was known to be lifetime but not past year, and the respondent was imputed to lifetime use of the child drug(s), then the child drug recency was set to lifetime but not past year. This can be deduced because the respondent could not have used the child drug more recently than the parent drug.
2. This edit only applied to OxyContin, methamphetamine, and crack, which are the only child drugs with frequencies. If the respondent used the parent drug on exactly 1

day in the past 12 months, and the respondent was imputed to lifetime use of the child drug, then the child drug recency of use was set equal to the parent drug recency of use, and the child drug 12-month frequency of use was set to 1 day. This can be deduced because the respondent could not have used the child drug on any days when the parent drug was not used, so the recencies and frequencies cannot differ.

3. If the parent drug incidence data indicated a date of first use in the past year, the parent drug recency of use was past year but not past month, and the respondent was imputed to lifetime use of the child drug(s), then the recency of use for the child drug(s) was set to past year but not past month. This can be deduced because the respondent could not have used the child drug more recently than the parent drug (eliminating the possibility of past month recency), and the respondent also could not have started using the child drug before the parent drug (eliminating the possibility of lifetime but not past year recency).
4. Similarly, if the parent drug incidence data indicated a date of first use in the past year, the parent drug recency of use was past month, and the respondent was imputed to lifetime use of the child drug(s), then the recency of use for the child drug(s) was set to past year (where past month vs. not past month use could not be determined). This can be deduced because the respondent could not have started using the child drug before the parent drug (eliminating the possibility of lifetime but not past year recency).

6.4.3 Other Age-at-First-Use Edits

1. This edit applied to all parent age-at-first-use variables: cigarettes, overall hallucinogens, overall pain relievers, overall stimulants, and cocaine. If the parent age-at-first-use value was missing and the minimum of the child age-at-first-use values was 3 years, then the parent age-at-first-use value was edited to 3 years. This could be deduced because respondents with age-at-first-use values of less than 3 years were ineligible to be donors (see Section 6.7.1.6).
2. This edit applied to all child age-at-first-use variables: daily cigarettes, LSD, PCP, Ecstasy, OxyContin, methamphetamine, and crack. If the parent age at first use was equal to the age, all missing child age-at-first-use values were edited to the age.
3. This edit also applied to all child age-at-first-use variables. If the parent age at first use was equal to 1 less than the age, the child recency[54] was lifetime but not past year (or, for cigarettes, past 3 years but not past year), and the child age-at-first-use value was missing, then the child age-at-first-use value was assigned to 1 less than the age. In particular, the child age at first use must be either less than AGE – 1, greater than AGE – 1, or equal to AGE – 1. It cannot be less than AGE – 1, because the parent age at first use is AGE – 1, and the respondent could not have begun using a child drug before using the parent drug. It also cannot be greater than AGE – 1, because the recency implies that the respondent did not use the drug while at his or her current age (because he or she did not use the drug at all in the past year). If the respondent

[54] Because there was no recency question associated with daily cigarettes, the overall cigarette recency was used instead.

did not use the drug at all in the past year, then he or she could not have begun using the drug in the past year. Because the child age at first use cannot be less than AGE – 1 or greater than AGE – 1, it must be equal to AGE – 1.

4. If the age at first cigarette use was equal to AGE – 3, cigarette recency was lifetime but not past 3 years, and age at first daily cigarette use was missing, then age at first daily cigarette use was assigned to AGE – 3. The logic is similar to the above edit: age at first daily cigarette use must be either less than AGE – 3, greater than AGE – 3, or equal to AGE – 3. The age at first cigarette use precludes the possibility that the age at first daily cigarette use was less than AGE – 3, and the cigarette recency precludes the possibility that the age at first daily cigarette use was greater than AGE – 3.

6.5 Imputation-Revised Drug Recency of Use, 12-Month Frequency of Use, 30-Day Frequency of Use, and 30-Day Binge Drinking Frequency

In the 2007 survey, recency of use, frequency of use in the past 12 months, frequency of use in the past 30 days, and (for alcohol) 30-day binge drinking frequency[55] were modeled separately for each drug. These measures of drug usage constituted a multivariate set within each drug. Provisional values replaced missing values for use in subsequent models, where necessary, using the UPMN method. After having modeled all of the drug use measures for a given drug, the MPMN method was employed to determine final imputed values using the predicted values from these models. Separate multivariate imputations were conducted for each drug.

The implementation of the PMN method required the identification of a modeling hierarchy. However, for the multivariate imputations described in this section, two separate modeling hierarchies were employed. Within a multivariate set, recency of use was modeled first, followed by the 12-month frequency of use (where applicable), 30-day frequency of use (where applicable), and (for alcohol) 30-day binge drinking frequency. Once the multivariate imputation for a given drug was completed, the recency of use for the next drug in the sequence was modeled.

6.5.1 Recency of Use

6.5.1.1 Hierarchy of Drugs

A complete drug hierarchy, as described in Appendix C, was not required for recency of use, because only cigarettes, alcohol, and marijuana recencies were used as covariates in models for subsequent drugs. This was because of difficulties that would have arisen if too many covariates were included in the polytomous logistic models. (Lifetime usage indicators of other drugs were included instead of recency-of-use indicators.) However, for the sake of convenience, the recency-of-use imputations did follow the same hierarchy as described in Section 6.2.

[55] "Binge drinking" was defined as having five or more drinks on the same occasion on a given day. The 30-day binge drinking frequency was defined as the number of days out of the past 30 days in which the respondent had five or more drinks on the same occasion.

6.5.1.2 Setup for Model Building and Hot-Deck Assignment

As with all the drug use measures, the recency-of-use imputations were conducted separately for respondents aged 12 to 17, 18 to 25, and 26 or older. To impute missing recency-of-use values for each drug, it was first necessary to define the eligible population within each of these age groups. Using the imputation-revised lifetime indication of use, the file was reduced to lifetime users. Among these lifetime users, item respondents and nonrespondents for each drug were identified across recency of use and (where applicable) the 12-month, 30-day, and (for alcohol only) 30-day binge drinking frequency-of-use measures. If a valid response was provided for each drug use measure, the person was deemed an item respondent for the drug. Otherwise, he or she was an item nonrespondent.

Before modeling, the respondents' weights were adjusted so that they represented all lifetime users. Because item respondents were defined at the drug level, these adjustments were made separately for each drug (and within the three age groups). The covariates in the item response propensity model included imputation-revised cigarette, alcohol, and marijuana recencies (where applicable); lifetime indicators of usage of drugs other than cigarettes, alcohol, and marijuana; gender; age;[56] race/ethnicity; first-order interaction of gender and race/ethnicity; marital status; education level; employment status;[57] census region; and a CBSA[58] indicator.[59] In addition, a three-level State-rank variable was defined by clustering States according to the prevalence of past month use of the drug of interest and was included as a covariate in the models.[60]

6.5.1.3 Sequential Model Building

Using the adjusted weights, the probability of selecting each recency-of-use category was modeled within each age group using, where possible, polytomous logistic regression. The predictors included in the models were imputation-revised cigarette, alcohol, and marijuana recencies (where applicable); lifetime indicators of usage of drugs other than cigarettes, alcohol, and marijuana; centered age; centered age squared; centered age cubed; gender; race/ethnicity; first-order interactions of centered age, gender, and race/ethnicity; marital status; education level;

---

[56] The covariate "categorical age" was divided into five categories to match the categories used in sample selection (12 to 17, 18 to 25, 26 to 34, 35 to 49, and 50 or older). For the 12-to-17 and 18-to-25 age groups, categorical age was not included as a covariate in the item response propensity models.

[57] Marital status, education level, and employment status were included as covariates for the 18-to-25 and 26-or-older age groups only.

[58] CBSAs, developed in response to standards put forth by the Office of Management and Budget (OMB), are metropolitan and micropolitan areas that were designated using data from the 2000 census. More information about CBSAs can be retrieved from http://www.census.gov/hhes/www/housing/resseg/cbsa.html.

[59] These variables were included in every model unless convergence problems arose. If this occurred, the model was reduced.

[60] For drug/age group combinations where the proportion of past month users was low, age groups were aggregated within the given drug. Sometimes all three age groups were aggregated, and other times two neighboring age groups were aggregated. This was done mainly to keep the State rank from being easily influenced by only one or two users in the sample. Also, all States with no users were placed in the lowest State-rank category, so for especially rare drugs such as heroin and sedatives, substantially more than one third of the States received the lowest State rank. Those individuals whose past month use status was unknown were treated as if they were not past month users.

employment status; census region; a CBSA indicator; and State rank.[61] For a summary of the variables included in each drug model, see Appendix F.

For certain drugs, the proportion of users who were past year users was quite small when compared with the total number of lifetime users. The lopsided distributions[62] for these drugs caused convergence problems when fitting multinomial logistic models. This problem occurred with the following set of drugs that were either rare overall or were rare within one or more age groups: inhalants, hallucinogens, sedatives, stimulants, tranquilizers, and heroin. To alleviate this problem, the single multinomial logistic model was replaced with two binary logistic models[63] that were fitted in a hierarchical manner.

As with the multinomial logistic model, the first binary logistic model was fitted among lifetime users, but the past month and past year but not past month categories in the response variable were collapsed into a single level. In a similar manner to other recency-of-use models, respondents' weights were adjusted so that they represented all lifetime users. The predicted probability of past year use given lifetime use was obtained from this model.

The second model was limited to past year users, where the response variable had two levels: past month and past year but not past month users. For the second model, respondents' weights were adjusted so that they represented all past year users. In order to do this, it was necessary to completely define the domain of past year users. Missing values were provisionally imputed to past year or not past year use by randomly allocating the response utilizing the predicted means from the first model.

From the two binary logistic models, both the probability of past month use and the probability of past year but not past month use were obtained and utilized in the provisional hot-deck program for recency, which is discussed in subsequent sections. Once the predicted means were determined from the two models, a single vector of predicted means conditional on lifetime usage, as with the multinomial logistic models, was determined in the following manner:

$$P(\text{past month use} \mid \text{lifetime use}) = P(\text{past month use} \mid \text{past year use}) * P(\text{past year use} \mid \text{lifetime use}), \text{ and}$$

$$P(\text{past year, not past month use} \mid \text{lifetime use}) = P(\text{past year, not past month use} \mid \text{past year use}) * P(\text{past year use} \mid \text{lifetime use}).$$

6.5.1.4 Computation of Predicted Means and Univariate Predictive Mean Neighborhoods

Because recency-of-use and frequency-of-use variables for a given drug were considered part of a multivariate set, the calculation of predicted means for the frequency-of-use variables

[61] These variables were included in every model unless convergence problems arose. If this occurred, the model was reduced.

[62] A "lopsided distribution" in the context of recency of use is where, among the categories past month use, past year but not past month use, and lifetime not past year use, only a small minority of respondents gave a response of "past month use."

[63] The set of covariates used for these binary logistic models were the same as those for logistic modeling given earlier in this section.

required the item nonrespondents to be identified as provisional past month and/or past year users. Within a given drug and within each age group, predicted probabilities for each of the recency categories were computed for both item respondents and item nonrespondents using the parameters from the appropriate logistic model(s). The predicted probabilities from the recency models were used to assign provisional values using the UPMN imputation method. A vector of predicted probabilities for each respondent was created by the logistic regression model(s). Because only a single predicted mean was used to determine the neighborhood when determining provisional values, not all of the predicted probabilities from the model were used.[64] Also, because past month use was the most critical measure of recency of drug use, the neighborhoods were defined based on the probability of past month use. If possible, provisional donors were chosen with predicted means within the delta of the recipient, where the value of delta varied depending on the value of the predicted means, which in this case were predicted probabilities of past month use.[65] In particular, delta was defined as 5 percent of the predicted probability if the probability was less than 0.5, and it was defined as 5 percent of 1 minus the predicted probability if the probability was greater than 0.5. This allowed a looser delta for predicted probabilities close to 0.5, and it allowed a tighter delta for predicted probabilities close to 0 or 1. If no donors were available with predicted means within delta of the recipient, the neighborhood was abandoned and the donor with the closest predicted mean was chosen.

6.5.1.5 Assignment of Provisional Imputed Values

Subject to the constraints described in the next section, separate assignments of provisional values were performed within each of the three age groups. The final recency-of-use imputations were multivariate across drug measures and are further described in Section 6.5.5.

6.5.1.6 Constraints on Univariate Predictive Mean Neighborhoods

As stated in the lifetime usage section, a UPMN can be restricted by logical constraints (which cannot be loosened) and by likeness constraints (which can be loosened) to make candidate donors in the neighborhood as similar to recipients as possible. The likeness constraints and logical constraints that were applied are described below.

As with all other drug use measures, neighborhoods for recency of use were restricted so that candidate donors and recipients would be within the same age group (12 to 17, 18 to 25, and 26 or older). Models were built separately within these three groups, so this likeness constraint was never loosened. A small delta also could be considered a likeness constraint, which could be loosened by enlarging or removing delta. As previously stated, if no donors were found in the delta, the neighborhood was then abandoned and the donor with the predicted mean closest to the recipient was chosen. If possible, donors and recipients were required to be from States with the same level of usage of a given drug (State rank), where the level of usage was defined in terms of

[64] A multivariate procedure could have been used to determine the provisional values that would be used for all of the predicted probabilities in the predictive mean vector. However, the amount of effort and computation time associated with multivariate imputation is considerably greater with multivariate procedures than with univariate procedures. Because the imputation was only provisional, a univariate imputation was used.

[65] The probability of past month use was used to define univariate neighborhoods for recency of use, even when it was known that the respondent was not a past month user. This could occur if the edited recency of use was, for example, lifetime not past month use.

the weighted proportion of a given State's residents who had used a given drug in the past month.[66] If insufficient donors were available within these constraints, they were loosened in the following order: (1) the neighborhood was abandoned and the donor with the closest predicted mean was chosen; and (2) donors and recipients were no longer required to be from States with similar usage levels.

The only logical constraints placed on the neighborhoods involved cases where a general recency category was available for a respondent and imputation was required to determine the specific recency categories. The general recency categories that appeared are shown in Table 6.5. Logical constraints ensured that only donors with allowable specific recency categories were included in the neighborhood. Other logical constraints involving a very small number of respondents were not applied to the provisional imputations. The complete list of constraints used in the multivariate imputation of recency and frequency of use is provided in Section 6.5.5.

Table 6.5 General Incomplete Recency Categories for Tobacco and Nontobacco

General Incomplete Recency Category	Allowable Specific Recency Categories (Tobacco)	Allowable Specific Recency Categories (Nontobacco)
Lifetime	1. Lifetime but not past 3 years 2. Past 3 years but not past year 3. Past year but not past month 4. Past month	1. Lifetime but not past year 2. Past year but not past month 3. Past month
Past Year	1. Past year but not past month 2. Past month	1. Past year but not past month 2. Past month
Lifetime, Not Past Year	1. Lifetime but not past 3 years 2. Past 3 years but not past year	N/A (for nontobacco, this was a specific recency category)
Lifetime, Not Past Month	1. Lifetime but not past 3 years 2. Past 3 years but not past year 3. Past year but not past month	N/A
Lifetime, Not Past Month but within Past 3 Years	1. Past 3 years but not past year 2. Past year but not past month	N/A
Past 3 Years	1. Past 3 years but not past year 2. Past year but not past month 3. Past month	N/A

N/A = not applicable.

6.5.1.7 Multivariate Assignments

Occasionally, more than one substance was associated with a single predicted mean, leading to a multivariate assignment of imputed values. However, for the provisional imputed

[66] Those individuals whose past month use status was unknown were treated as if they were not past month users.

values, a multivariate assignment was necessary only if the substances associated with a single predicted mean were of equal standing. This occurred with smokeless tobacco, which consists of chewing tobacco and snuff. No provisional imputed values were determined for substances that were a subset of the substance associated with the predicted mean ("parent/child" drugs). Examples of such situations included cocaine (parent) and crack (child); pain relievers (parent) and OxyContin (child); stimulants (parent) and methamphetamine (child); and hallucinogens (parent) and LSD, PCP, and Ecstasy (children). The multivariate assignment of imputed values for chew and snuff is discussed below.

For reasons discussed in Section 6.3.7.1, one model for smokeless tobacco (a combination of the chew and snuff responses) was fitted, rather than individual models for chew and for snuff. The nearest neighbor hot-deck neighborhood was then based on the predicted probability of past month use of smokeless tobacco. Missing recency-of-use values for chew and/or snuff were replaced with the (provisional) values from a donor within this neighborhood. At this stage in the process, lifetime use or nonuse of either chew or snuff was considered known (employing information from the lifetime usage imputation). For lifetime users of chew or snuff who were missing some or all of their recency-of-use information for either chew or snuff, but not both, only the missing specific recency-of-use values were replaced.[67] However, for individuals missing recency-of-use information for both chew and snuff (given that the respondent was known or was imputed to be a chew user and a snuff user), values for both were obtained from the same donor. The provisional recency of use for smokeless tobacco was obtained by combining the recency-of-use information from chew and snuff.

6.5.2 12-Month Frequency of Use

6.5.2.1 Hierarchy of Drugs

The modeling of 12-month frequency sequentially followed that of recency of use for each drug. Across drugs, the sequence was exactly the same as the one used for recency of use. Data on 12-month frequency of use were not collected for all of the drugs. Thus, these imputations were conducted for a subset of the drugs (see Table 6.1).

6.5.2.2 Setup for Model Building and Hot-Deck Assignment

As with all the drug use measures, the 12-month frequency-of-use imputations were conducted separately for respondents aged 12 to 17, 18 to 25, and 26 or older. The eligible population for the imputation of 12-month frequency of use was past year users of the drug in question (as defined by the provisional recency of use). Among the past year users of each drug, the item response indicator and the response propensity adjustment were defined. Item respondents were defined using the same criterion as was used in the recency-of-use imputations. Namely, the respondent had to have a valid response to all of the applicable measures for the drug of interest. The item response propensity adjustment was then computed so that the respondents' weights accurately represented all past year users of the drug. The predictors in the

[67] For respondents missing all of their recency information, the only known information was that they were lifetime users (either from their survey response or from imputation). For respondents missing some of their recency information, they might be assigned a general recency category (outlined in Table 6.3), and if so, then specific recency values were imputed.

response propensity adjustment modeling included the provisional indicator of past month use for the drug of interest; (where available) recencies of use for cigarettes, smokeless tobacco, cigars, pipes, alcohol, marijuana, cocaine, crack, heroin, hallucinogens, inhalants, pain relievers, tranquilizers, stimulants, and sedatives;[68] categorical age; race/ethnicity; gender; census region; and a CBSA indicator.[69]

6.5.2.3 Model Building

As indicated in the previous section, only past year users of the drug of interest were used to build the 12-month frequency-of-use model. The response variable of interest in the 12-month frequency-of-use models for most respondents, prior to a normalizing transformation, was the proportion of the days in a full year (365.25) on which a respondent used a particular drug. For example, if a respondent entered a 12-month frequency of 100, the (untransformed) response variable of interest would be 100/365.25. Some respondents, however, started using the drug within the past year. If they responded to the month-at-first-use question, the difference between the month of first use and the date of the interview indicated the total time period during which they could have been using drugs.[70] If the date of the interview was July 10, for example, and the month of first use was March of the same year, the maximum period during which the respondent could have used is the number of days between March 1 and July 10 (inclusive), or 101. Thus, if a respondent entered a 12-month frequency of 100, the (untransformed) response variable of interest would be 100/101 instead of 100/365.25. The range of values for the proportion was from (greater than) 0 to 1. Hence, in order to model 12-month frequency of use, the following empirical logit transformation was computed for all respondents:

$$\log\left[\left(Y_i + 0.5\right)/\left(N_i - Y_i + 0.5\right)\right],$$

where Y_i is the observed 12-month frequency for respondent i and N_i is the total number of days in the year that respondent i could have used the substance. This transformation is nearly equivalent to the standard logit transformation:

$$Y_i^* = \log\left[P_i/\left(1 - P_i\right)\right],$$

where P_i is defined as the proportion of days in the past year in which respondent i used the drug. The standard logit transformation was not used because it was not defined for daily users. Using the adjusted weights, a linear univariate regression model using SUDAAN® software was then fitted for the log-transformed variable Y_i within each age group.

[68] If the recency of use for a particular drug was not yet defined, the lifetime indication of use was used instead. The recency of use of the drug being modeled (past month use vs. past year but not past month use) was always defined.

[69] These variables were included in every model unless convergence problems arose. If this occurred, the model was reduced.

[70] If a respondent initiated use in the past year (according to his or her age-at-first-use response), but did not answer the month-at-first-use question, the maximum period the respondent could have been using drugs was assumed to be 365.25 because no other information was available.

Because the 12-month frequency models were limited to past year users, only two recency categories could have resulted: past month use and past year but not past month use.[71] Hence, recency of use for the drug being modeled was represented as a covariate in the 12-month frequency-of-use model by a single indicator variable representing these two categories. Imputation-revised recency of use for other drugs was used if available. If the missing values for a given drug's recency of use had not yet been imputed, a single covariate was used that indicated lifetime usage of that drug. To control for State variations in drug use, the State-rank groups defined for the recency-of-use imputations were included as covariates in the 12-month frequency-of-use models.[72] Thus, the models included a provisionally imputed indicator of past month use of the given drug; (where available) the imputation-revised recencies of use for cigarettes,[73] smokeless tobacco, cigars, pipes, alcohol, marijuana, cocaine, crack, heroin, hallucinogens, inhalants, pain relievers, tranquilizers, stimulants, and sedatives; centered age; centered age squared; centered age cubed; gender; race/ethnicity; first-order interactions of centered age, gender, and race/ethnicity; marital status; education level; employment status; census region; a CBSA indicator; and State rank (based on past month prevalence of the drug).[74] Predicted 12-month frequencies of use were defined by back-transforming the resulting predicted values. For a complete summary of the 12-month frequency-of-use models, see Appendix F.

The predicted mean that resulted from the 12-month frequency-of-use model was a logit of the proportion of the year used. This logit was back-transformed into a proportion for use as the variable from which the neighborhoods were created. This proportion could be treated as a probability, which, in turn, could be multiplied by the probability of past year use to make the predicted mean conditional on lifetime use of the drug in question. When calculating predicted means for some item nonrespondents, sometimes it was not known whether they were past year users. Hence, to make the predicted means conditional on the same recency of use, the variables were transformed to make them conditional on what was known.

6.5.2.4 Computation of Predicted Means and Univariate Predictive Mean Neighborhoods

Within a given drug, predicted means from the 12-month frequency-of-use models were computed for both item respondents and item nonrespondents using the parameters from the regression model. The logits were converted back to proportions, which were in turn multiplied by the probability of past year use to make the predicted mean conditional on lifetime use.[75] Using the UPMN method, neighborhoods were defined based on these predicted means. If

[71] For item nonrespondents, where parameter estimates were used to determine predicted means, past year use was defined based on a provisional imputation.

[72] As with the recency-of-use models, for a few cases, the State-rank variable could not be included in the model. Usually, but not always, the age group/drug combination that had problems was the same for recency of use and 12-month frequency of use.

[73] The covariates based on recency-of-use variables that corresponded to drugs other than the one being modeled (if the recency of use was available) were defined by a series of dummy variables reflecting the different recency categories.

[74] These variables were included in every model unless small sample sizes precluded the use of such a large pool of covariates. If this occurred, the model was reduced.

[75] The dependent variable in the model used was the empirical logit, as described in Section 6.5.2.3. The value that was back-transformed was obtained by solving for Y/N, where Y is the number of days of use (in a year) and N is the number of potential days of use in the year.

possible, provisional donors were chosen with predicted means within delta of the recipient, where the value of delta varied depending on the value of the predicted means, which in this case were predicted proportions of the year used. In particular, delta was defined as 5 percent of the predicted proportion if the proportion was less than 0.5, and it was defined as 5 percent of 1 minus the predicted proportion if it was greater than 0.5. This allowed a looser delta for predicted proportions close to 0.5, and it allowed a tighter delta for predicted proportions close to 0 or 1. As with recency of use, if no donors were available with predicted means within delta of the recipient, the neighborhood was abandoned and the donor with the closest predicted mean was chosen.

6.5.2.5 Assignment of Provisional Imputed Values

For all drug use measures except 12-month frequency, the observed value of interest was donated directly to the recipient. However, because donors and recipients could potentially have had a different maximum possible number of days in the year that they could have used a substance, the observed proportion of the total period was donated, rather than the observed 12-month frequency. In the assignment step, the donor's proportion of the total period was multiplied by the recipient's maximum possible number of days in the year on which he or she could have used the substance in order to arrive at a 12-month frequency-of-use value for the recipient. Separate assignments were performed within each of the three age groups, subject to the constraints described in the next section. For the 12-month frequency of use, "level of usage" for the State-rank groups was defined in terms of the proportion of a given State's residents who had used a given drug in the past month. Assignments were not required for tobacco because the tobacco module did not have 12-month frequency-of-use questions. Also, assignments were not needed for "pills"[76] because pills did not have a 30-day frequency-of-use question, making it unnecessary to obtain provisionally imputed 12-month frequencies. The final 12-month frequency-of-use imputations were multivariate across drug measures and are further described in Section 6.5.5.

6.5.2.6 Constraints on Univariate Predictive Mean Neighborhoods

An obvious logical constraint for 12-month frequency of use was that all donors were past year users. Other logical constraints involved the interview date, month of first use, birthday, recency of use, and 30-day frequency of use. See Section 6.5.5 for a discussion of the multivariate imputation of recency and frequency of use.

Two likeness constraints used in the assignment of values for 12-month frequency of use were identical to those of recency of use: the three age groups and the State-rank groups based on level of past month usage. As with the recency-of-use models, delta was set so that the predicted means of all potential donors were within 5 percent of the item nonrespondent's predicted mean, where the predicted mean was defined to be the proportion of the year (or maximum period within a year) during which a respondent used a drug. Finally, recipients and donors were required to have the same recency of use (past month vs. past year but not past

[76] "Pills" were defined as pain relievers, tranquilizers, stimulants, and sedatives.

month), whether that recency of use was reported or imputed.[77] If no donors were available within these constraints, they were loosened in the following order: (1) the neighborhood was abandoned and the donor with the closest predicted mean was chosen; (2) donors and recipients were no longer required to be from States with similar usage levels; and (3) donors and recipients were no longer required to have the same recency of use.

Occasionally, more than one substance was associated with a single predicted mean. However, for the provisional imputed values, only the "parent" drug was of interest (e.g., only the provisionally imputed cocaine 12-month frequency was needed, not the crack 12-month frequency). Therefore, multivariate assignments were not needed for the provisional UPMNs, but they did occur in the final multivariate imputation of recency and frequency.

6.5.2.7 Multivariate Assignments

Although more than one substance was occasionally associated with a single predicted mean, the provisionally imputed 12-month frequencies were required only if they were needed for calculating predicted means using the coefficients from a subsequent model. A multivariate assignment was necessary only if the substances associated with a single predicted mean were of equal standing. This occurred with smokeless tobacco, which consists of chewing tobacco and snuff. However, no question about 12-month frequency was asked of smokeless tobacco users. Moreover, no provisionally imputed values were required for substances that were a subset of the substance associated with the predicted mean, which would be referred to as "parent/child" drugs (see Section 6.2). Hence, no multivariate assignments were required for the provisionally imputed 12-month frequency.

6.5.3 30-Day Frequency of Use

6.5.3.1 Hierarchy of Drugs

The modeling of 30-day frequency followed that of recency and 12-month frequency of use for each drug. Across drugs, the sequence was exactly the same as that for recency of use. Data on 30-day frequency of use were not collected for all of the drugs. Thus, these imputations were performed for only a subset of the drugs (see Table 6.1).

6.5.3.2 Setup for Model Building and (for Alcohol Only) Hot-Deck Assignment

The file was first reduced to the eligible population, which was past month users, as defined by the provisional recency variable. Next, item respondents and nonrespondents were defined according to the same criterion used for the recency and 12-month frequency imputations. To be an item respondent, the individual had to have provided valid responses to all applicable measures for the drug of interest. The item response propensity adjustment was then computed so that the respondents' weights accurately represented all past month users of the drug. Predictors for the response propensity models included the provisional 12-month frequency

[77] Because all respondents in the 12-month frequency of use imputation were past year users by definition, item nonrespondents who were past month users required donors who were past month users, and item nonrespondents who were past year but not past month users required donors who matched that specific recency category.

for the drug of interest (where applicable); (where available) recencies of use for cigarettes, smokeless tobacco, cigars, pipes, alcohol, marijuana, cocaine, crack, heroin, hallucinogens, inhalants, pain relievers, tranquilizers, stimulants, and sedatives;[78] categorical age; race/ethnicity; gender; census region; and a CBSA indicator.[79]

6.5.3.3 Model Building

As is apparent from the previous section, only past month users of the drug of interest were used to build the 30-day frequency-of-use model. The response variable of interest in the 30-day frequency-of-use models for most drugs, prior to a normalizing transformation, was the proportion of the days in a month (30) on which a respondent used a particular drug. The range of values for the proportion was from (greater than) 0 to 1. Hence, to model 30-day frequency of use, the following empirical logit transformation was computed for all respondents:

$$\log\left[\left(Y_i + 0.5\right)/\left(N - Y_i + 0.5\right)\right],$$

where Y_i was the observed 30-day frequency for respondent i and N was 30, the total number of days in the month that the respondent could have used the substance. This transformation was nearly equivalent to the standard logit transformation:

$$Y_i^* = \log\left[P_i/\left(1 - P_i\right)\right],$$

where P_i was defined as the proportion of days in the past year on which respondent i used the drug. The standard logit transformation was not used because it was not defined for daily users.[80] Using the adjusted weights, a linear univariate regression model was then fitted using SUDAAN software for the log-transformed variable Y_i within each age group.

Because the 30-day frequency models were limited to past month users, only one provisional recency category was relevant for the drug of interest.[81] Hence, provisional recency of use for the drug of interest could not be included in the 30-day frequency-of-use model. However, imputation-revised recency of use of other drugs could be included. For drugs where the recency of use was not yet modeled, the lifetime indication of use served as a surrogate for the recency-of-use indicators. Covariates representing the State-rank groups (defined by the level of past month use) were included to adjust for any State drug use differences. Other covariates included the provisional 12-month frequency of use for the drug of interest (where applicable); census region; centered age; centered age squared; centered age cubed; gender; race/ethnicity; the first-order interactions of centered age, gender, and race/ethnicity; marital status; education

[78] If the recency of use for a particular drug was not yet defined, the lifetime indication of use was used instead. The recency of use of the drug being modeled was not used, because all respondents in the model were past month users.

[79] These variables were included in every model unless convergence problems arose. If this occurred, the model was reduced.

[80] If the respondent was a daily user of the substance, then $\log[(Y + 0.5)/(N - Y + 0.5)] \approx \log[(N + 0.5)/0.5]$ with N = 30 so that it was defined for all respondents. (See Cox and Snell [1989] for a discussion of the empirical logit transformation.)

[81] For item nonrespondents, where parameter estimates were used to determine predicted means, past month use was determined based on a provisional imputation.

level; employment status; and a CBSA indicator.[82] The predicted 30-day frequencies of use were defined by back-transforming the predicted values from the models. For a complete summary of the 30-day frequency-of-use models, see Appendix F.

The predicted mean that came out of the 30-day frequency-of-use model was a logit of the proportion of the month used. This logit was back-transformed into a proportion for use as the variable from which the neighborhoods were created. This proportion was treated as a probability, which in turn was multiplied by the probability of past month use in order to have made the predicted means conditional on lifetime use of the drug in question.[83] When calculating predicted means for some item nonrespondents, sometimes it was not known whether they were past month users or not. Hence, to make the predicted means conditional on the same recency of use, the variables were transformed to make them conditional on what was known.

For cigarettes, chewing tobacco, and snuff, the empirical distribution for 30-day frequency of use was in fact a mixture distribution, with a positively skewed distribution from 1 to 29 and a spike at 30. These substances were modeled using two separate models. One was a logistic model for daily use versus nondaily use among past month users. For the nondaily past month users (i.e., those who had used between 1 and 29 days), a model much like the 30-day frequency-of-use models for other substances was used. In this case, the response variable in a linear regression model was a logit of the proportion of the period (30 days) during which a respondent used the substance. The same pool of covariates was used in the logistic model and the regression model with the logit as the response variable. It should be noted that, unlike recency of use, the 30-day frequencies for chewing tobacco and snuff were not combined into a single value for smokeless tobacco. Because it was not possible to determine if the x days using chewing tobacco overlapped with the y days using snuff, separate models were fitted for chewing tobacco and snuff.

6.5.3.4 Computation of Predicted Means and Univariate Predictive Mean Neighborhoods

Within a given drug, predicted means from the 30-day frequency-of-use models were computed for both item respondents and item nonrespondents using the parameters from the regression model. The 30-day frequency models were fitted after recency of use and 12-month frequency of use. The only drug for which provisional 30-day frequency values were required was alcohol because provisional 30-day frequencies were required to calculate 30-day binge drinking provisional values. Neighborhoods were created for each alcohol item nonrespondent using the UPMN method. The predicted means used to create the neighborhoods were given by the product of the predicted proportion of the month used (conditioned on past month use) and the probability of past month use given lifetime use (taken from the recency-of-use models).

[82] These variables were included in every model unless small sample sizes precluded the use of such a large pool of covariates. If this occurred, the model was reduced.

[83] The dependent variable in the model used was the empirical logit given in Section 6.5.3.3. The value that was back-transformed was obtained by solving for Y/N, where Y is the number of days of use (in a month) and N is the number of potential days of use in the month (30.4375).

6.5.3.5 Assignment of Provisional Imputed Values (Alcohol Only)

Separate assignments for the 30-day frequency of alcohol use were performed within each of the three age groups, subject to the constraints described in the next section. For the 30-day frequency of use, "level of usage" was defined in the same manner as the recency of use and 12-month frequency of use.

6.5.3.6 Constraints on Univariate Predictive Mean Neighborhoods (Alcohol Only)

An obvious logical constraint was that all donors had to be past month users. In addition, the donated 30-day frequency was required to be less than or equal to the respondent's preexisting 12-month frequency—whether that 12-month frequency was reported or imputed—and greater than or equal to the respondent's preexisting 30-day binge drinking frequency. Two likeness constraints used in the assignment of values for 30-day frequency of use were identical to those used for recency of use and 12-month frequency of use. The two likeness constraints were the three age groups and the State-rank groups based on level of past month usage. As with the recency-of-use models, delta was set so that the predicted means of all potential donors were within 5 percent of the item nonrespondent's predicted mean, where the predicted mean was defined to be the proportion of the month during which a respondent used a drug. If no donors were available within these constraints, they were loosened in the following order: (1) the neighborhood was abandoned and the donor with the closest predicted mean was chosen; and (2) donors and recipients were no longer required to be from States with similar usage levels.

6.5.3.7 Multivariate Assignments

Although more than one substance was occasionally associated with a single predicted mean, the provisionally imputed 30-day frequencies were required only if they were needed for calculating predicted means using the coefficients from a subsequent model. Of the substances within the multivariate set of recency of use and frequencies of use, only alcohol contained a measure (30-day binge drinking frequency) that was lower in the sequence than 30-day frequency of use. Because alcohol is not a "parent/child" drug (see Section 6.2 for a definition of "parent/child" drug), no multivariate assignments were required for provisionally imputed 30-day frequency.

6.5.4 30-Day Binge Drinking Frequency

For alcohol, an additional variable was defined that measured level of usage. In particular, the variable DR5DAY measured the binge drinking frequency or the number of days in the past month during which the respondent had five or more drinks. The imputation of the 30-day binge drinking frequency was similar to the imputation of 30-day frequency of alcohol use. However, the 30-day binge drinking frequency model included the provisional alcohol 30-day frequency of use[84] as a covariate. Moreover, the model was built using all past month users of alcohol, whether they were binge drinkers or not. Item respondents for alcohol were defined across recency, 12-month frequency, 30-day frequency, and the 30-day binge drinking frequency

[84] The provisional 30-day frequency of use was defined by randomly selecting donors from univariate neighborhoods, which were defined by using the respondent and nonrespondent predicted values.

measures. Therefore, the weight adjustment used in the modeling of the 30-day binge drinking frequency was the same as was used for the 30-day frequency model.

The response variable of interest in the 30-day binge drinking frequency model, prior to a normalizing transformation, was the proportion of the days in a month (30) on which a respondent drank five or more drinks. The range of values for the proportion was from 0 to 1. Hence, to model 30-day binge drinking frequency of use, the following empirical logit transformation was computed for all respondents:

$$\log\left[\left(Y_i + 0.5\right)/\left(N - Y_i + 0.5\right)\right],$$

where Y_i was the observed 30-day binge drinking frequency for respondent i and N was 30, the total number of days in the month that the respondent could have binge drunk. This transformation was nearly equivalent to the standard logit transformation:

$$Y_i^* = \log\left[P_i/\left(1 - P_i\right)\right],$$

where P_i was defined as the proportion of days in the past month during which respondent i had five or more drinks. The standard logit transformation was not used, because it was not defined for daily binge drinkers, nor was it defined for nonbinge drinkers among past month users.[85] Using the adjusted weights, a linear univariate regression model was then fitted for the log-transformed variable Y_i within each age group.

The predicted means from this model were used solely in the multivariate predictive mean vector used in the final MPMN imputation. No UPMN step was taken, and no provisional imputed values were determined.

6.5.5 Multivariate Imputation for Recency of Use, 12-Month Frequency of Use, 30-Day Frequency of Use, and 30-Day Binge Drinking Frequency

Sections 6.5.1 through 6.5.4 summarized how the set of lifetime drug users in the sample of the 2007 survey was separated into item respondents and item nonrespondents for the recency-of-use, 12-month frequency-of-use, 30-day frequency-of-use, and (for alcohol) 30-day binge drinking frequency drug use measures. These sections also summarized model building, computation of predicted means and delta neighborhoods, and the assignment of imputed values for these measures using a univariate predicted mean. In most cases, however, these univariate assignments were only provisional. As indicated in Table 6.1, the final imputed values for these drug use measures were obtained by building neighborhoods upon a vector of predicted means using the MPMN method. In a manner consistent with the univariate imputations, the multivariate assignments were done separately within three age groups: 12 to 17, 18 to 25, and 26 or older. As indicated in earlier sections, a respondent was eligible to be a donor for a given item nonrespondent if he or she had complete data across the drug use measures for the drug in

[85] If the respondent was a daily binge drinker of alcohol, then $\log[(Y + 0.5)/(N - Y + 0.5)] \approx \log[(N + 0.5)/0.5]$, where Y was the observed 30-day binge drinking frequency and N was the total number of days that the respondent could have used (usually 30). If the proportion was 0, then $\log[(Y + 0.5)/(N - Y + 0.5)] \approx \log[0.5/(N + 0.5)]$.

question and was within the same age group. As with the provisional imputations, the donated value for the 12-month frequency-of-use variable was determined by taking the product of the donated proportion of the year that the donor had used the substance of interest and the recipient's maximum number of possible days that he or she could have used the substance.

6.5.5.1 Constraints on Multivariate Predictive Mean Neighborhoods

6.5.5.1.1 Logical Constraints

The logical constraints required in the provisional univariate imputations discussed in Sections 6.5.1 through 6.5.3 also were required in the multivariate imputations. However, some constraints that potentially could have been applied in the provisional recency-of-use and provisional 12-month frequency imputations were not applied because of the very small number of respondents affected, and thus they are not listed in Table 6.5 or mentioned in Sections 6.5.1 or 6.5.2. However, these constraints were applied in the multivariate imputations. In particular, the possible recencies of use were limited based on the respondent's current age, the time between the interview date and the birthday, the time between the interview date and the month of first use, and any nonmissing frequency-of-use information. In addition, if the respondent was (or could have been) a past month user and was known to have used the drug at least once between 1 month before the interview date and 1 year before the interview date (because of the given month and/or year of first use), donors were required to have a 12-month frequency that reflected this. In general, the application of these constraints depended on what information was missing in the recency-of-use and frequency-of-use variables. The values missing for a given respondent define the "pattern of missingness." For example, one pattern of missingness for marijuana could be as follows: past year user of marijuana (recency partially missing), 12-month frequency not missing, and 30-day frequency missing. In this example, the logical constraints have to make the imputed 30-day frequency consistent with the preexisting 12-month frequency. In the case where the 12-month frequency-of-use variable was missing, an additional logical constraint involved the product of the donated proportion and the recipient's maximum possible number of days used in a year (called the "donated 12-month frequency product"). Because this product involved both the donor and the recipient, it had to be consistent with the 30-day frequency of use, regardless of whether the 30-day frequency was a preexisting nonmissing value or a donated value. It also had to be greater than 1 and/or greater than the 30-day frequency when it was known that the respondent was a past month user, but started using prior to the past month in the past year. The various patterns of missingness for each drug, the logical constraints imposed on the set of donors, and the frequency with which each missingness pattern occurred are provided in Appendix G.

6.5.5.1.2 Likeness Constraints

In addition, if possible, donors and recipients were required (as likeness constraints) to come from States with similar drug usage patterns for the drug in question, and donors were required to have each element of the multivariate predictive mean vector "close to" (i.e., within the delta distance) the recipient's elements of the predictive mean vector. Because the imputation was multivariate, the set of deltas was also multivariate, where a different delta corresponded to each element of the predictive mean vector. Finally, for drug modules with multiple substances (i.e., parent/child relationships), if the recency of use for one or more of the substances within the module was not missing, donors and recipients were required to have, if possible, the same

values for these recency-of-use indicators.[86] The number of respondents for whom donors were found within various likeness constraints, by missingness pattern and age group, is summarized in Appendix G. In general, the likeness constraints were loosened in the following order: (1) for drug modules with multiple substances, likeness constraints requiring donors and recipients to have the same recency-of-use values for nonmissing variables were removed, while any necessary logical constraints were maintained; (2) the neighborhood was abandoned and the donor with the closest predicted mean was chosen; and (3) donors and recipients were no longer required to be from States with similar usage levels.

6.5.5.1.3 More than One Substance for a Single Predictive Mean Vector

Occasionally, more than one substance was associated with a single predicted mean, whether it was for recency-of-use or frequency-of-use variables. This could be two substances of equal standing considered together when modeling (snuff and chewing tobacco) or could be drugs with a parent/child relationship (see Section 6.2 for a definition of parent/child relationship). The assignment of imputed values for these substances was unique for each situation.

Smokeless Tobacco. As noted in Sections 6.3.7.1 and 6.5.1.7, one model for smokeless tobacco recency of use (a combination of the chew and snuff responses) was fitted, rather than individual models for chew and snuff. The nearest neighbor hot-deck neighborhood was then based on the predicted probability of past month use of smokeless tobacco. The assignment of recency-of-use values for smokeless tobacco followed the same logical constraints in the multivariate imputation as the constraints in the univariate imputation discussed in Section 6.5.1.7.

Unlike recency of use, however, separate models for chew and snuff were built for 30-day frequency of use. The predicted means from these models were conditioned on past month use. In the 30-day frequency-of-use imputations, discussed in Section 6.5.3.3, the predicted means used to form the neighborhoods were conditioned on lifetime usage rather than past month usage. Because the 30-day frequency models gave predicted means conditioned on past month use, it was necessary to determine the probability of past month use given lifetime use, which could be obtained from the recency models. Because the 30-day frequencies for chew and snuff could not be combined, recency-of-use models were built for chewing tobacco and snuff separately.[87] (This was in addition to the regular recency-of-use model that was built for smokeless tobacco.) The covariates used in the models are provided in Appendix F.

---

[86] Donors also were required to match the recipient with respect to lifetime use of "other" drugs for hallucinogens. During the processing, this constraint was loosened later than the constraint involving the recency-of-use indicators for the three child drugs. No likeness constraints involving "other" drugs were applied to pain relievers or stimulants.

[87] To properly condition the respective 30-day frequency predicted means for chewing tobacco and snuff, it was not possible to use the predicted probabilities available for the recency of use of smokeless tobacco as a whole. Instead, separate recency-of-use models for chewing tobacco and snuff were used to obtain the predicted probabilities of both past month use and past year but not past month use of these substances. These were the values utilized in the construction of conditional probabilities for the 30-day frequencies of chewing tobacco and snuff. See Appendix G for details.

Cocaine and Crack. Even though cocaine and crack were in distinct modules, single models were fitted for recency-of-use and frequency-of-use variables using the information from the cocaine module. Crack is a type of cocaine, so donors for the two substances were obtained using a single neighborhood. As with smokeless tobacco, use or nonuse of crack was considered known (using information from the lifetime imputations). Hence, as a logical constraint, users of crack with incomplete recency (or frequency) information required donors who were also crack users. Moreover, if the cocaine recency was not missing, the donated crack recency could not be more recent than the preexisting cocaine recency. Similarly, if the crack recency was not missing, but the cocaine recency was missing, the donated cocaine recency could not be less recent that the preexisting crack recency.

If at least one of the frequency-of-use variables was missing, but the cocaine recency was not, the cocaine recency of use for donors and recipients had to match. In addition, donors and recipients were required to have the same crack recency of use if it was known that the recipient used crack in the past year. Both of these constraints were applied regardless of the pattern of missingness among the frequency-of-use variables.[88] Additional logical constraints involved the donated 12-month frequency products for both crack and cocaine. If both the crack and cocaine 12-month frequency-of-use values were missing, it was necessary to check the donated products against each other for consistency because this product depended upon both the donor and recipient, even though the donated proportions came from the same donor. Both also had to be checked for consistency against the 30-day frequency-of-use values (if the respondent was a past month user of crack and/or cocaine), regardless of whether those variables were preexisting nonmissing values or donated imputed values. If only one of the 12-month frequency-of-use variables were missing, the donated product was checked for consistency against the preexisting nonmissing 12-month frequency-of-use value and against the 30-day frequency-of-use variables, imputed or not.

Hallucinogens (LSD, PCP, Ecstasy, and "Other" Hallucinogens), Pain Relievers (OxyContin and "Other" Pain Relievers), and Stimulants (Methamphetamine and "Other" Stimulants). As stated in Section 6.3.7.3, the modules for hallucinogens, pain relievers, and stimulants included subgate questions referring to child drugs. Hallucinogens had three child drugs (LSD, PCP, and Ecstasy); pain relievers had one (OxyContin); and stimulants had one (methamphetamine). Recency-of-use information for the parent drugs was used in subsequent models, and recency-of-use information for the child drugs was not used. Hence, obtaining provisional values for the recency of use of the child drugs was not necessary. Predictive recency probabilities were calculated for the parent drugs, and these probabilities were used to determine neighborhoods for each group of drugs. As with smokeless tobacco, use or nonuse of the child drugs was considered known (including values that were imputed in the lifetime usage imputations).

Hallucinogens. Using the neighborhood created from the predictive mean vector, missing specific recency categories for LSD and/or PCP and/or Ecstasy and/or hallucinogens as a whole, were replaced with the specific recency categories from a single donor. Child drug

[88] In one case, a donor could not be found within the respondent's age group (26 or older) who met all the logical constraints after all the likeness constraints had been loosened. Therefore, the 12-month frequency for crack and the 30-day frequency for cocaine was randomly assigned within the appropriate range.

(LSD, PCP, and/or Ecstasy) users with incomplete recency information were constrained to have donors who were lifetime users of the specific child drug(s). Moreover, donors were constrained so that a preexisting child drug recency could not be more recent than a donated parent drug recency. Conversely, a preexisting parent drug recency value could not be less recent than any donated child recency value. In addition, donors were constrained for those respondents missing the parent recency who used no "other" type of hallucinogen so that the donated parent recency was equal to the minimum of the child recencies, whether donated or not. For individuals missing recency information for the parent drug or the child drugs, only the missing value(s) was replaced. For individuals missing recency information for two or more of these substances, the missing categories were replaced by values from the same donor.

No 12-month frequency-of-use variables were available for any of the three child drugs. However, the donated 12-month frequency product for all hallucinogens was required to be consistent with the 30-day frequency-of-use value for all hallucinogens, whether it was imputed or was a preexisting nonmissing value.

Pain Relievers. A similar procedure was followed for the pain relievers module. Using the neighborhood created from the predictive mean vector, missing specific recency-of-use categories for OxyContin and/or pain relievers as a whole were replaced with the specific recency categories from a single donor within this neighborhood. OxyContin users with incomplete recency information were constrained to have donors who were also OxyContin users. Moreover, donors were constrained so that a preexisting OxyContin recency-of-use value could not be more recent than a donated pain relievers recency-of-use value, and, conversely, a preexisting pain reliever recency-of-use value could not be less recent than the donated OxyContin recency of use. In addition, donors were required to have an overall pain reliever recency equal to their OxyContin recency for those respondents missing both overall pain reliever recency and OxyContin recency, who used no "other" type of pain reliever. For individuals missing recency information for OxyContin and/or pain relievers as a whole, only the missing categories were replaced. For individuals missing recency information on both of these substances, the missing categories were replaced by values from the same donor.

The major difference between hallucinogens and pain relievers was that a 12-month frequency-of-use variable was available for the child drug OxyContin. Even though separate 12-month frequency questions were asked for overall pain relievers and, more specifically, OxyContin, 12-month frequency was modeled for overall pain relievers only. As with cocaine and crack, additional logical constraints involved the product of the donated proportion and the recipient's maximum possible number of days used in a year for both OxyContin and pain relievers. If both the pain relievers and OxyContin 12-month frequency-of-use values were missing, it was necessary to check the donated products against each other for consistency because this product depended upon both the donor and recipient, even though the donated proportions came from the same donor. No additional check was necessary, because pain relievers did not have a 30-day frequency-of-use variable. If only one of the 12-month frequency-of-use variables was missing, the donated product was checked for consistency against the preexisting nonmissing 12-month frequency-of-use value.

Stimulants. The procedure used for the stimulants module was very similar to the one followed for the pain relievers module. As for the pain relievers, a 12-month frequency-of-use

variable was available for the child drug methamphetamine. The constraints that were applied, and the predicted means that were used, were the same as for pain relievers.

6.5.5.2 Final Multivariate Assignment

The full predictive mean vector contained several elements for recency-of-use (different probabilities associated with each of the recency categories) and frequency-of-use variables. Each element in the full vector of predicted means was adjusted so that all elements were conditioned on the same usage status whenever possible. The resulting elements in the predictive mean vector that could have potentially resulted are shown in Table 6.6. It is important to note that not all drugs contained all the elements given, as is apparent by looking at the rightmost column in Table 6.6. Table 6.6 assumes that only the lifetime usage is known. If other information about the recency of use is known (e.g., past year user), the predictive mean vector is adjusted accordingly. Table 6.7 shows the full predictive mean vector for each drug. The portion of the full predictive mean vector used to determine the neighborhood for a particular item nonrespondent was dependent on the pattern of missingness for that item nonrespondent. If partial information was available regarding recency of use, that information was used to adjust the recency-of-use probabilities. The portions of the full predictive mean vector used to create the MPMNs for each missingness pattern, with accompanying adjustments, are provided in Appendix G. The Mahalanobis distance was then calculated using only the portion of the predictive mean vector that was associated with the given missingness pattern, with elements appropriately adjusted. If no donors were available who had predicted means within a multivariate delta of the recipient's vector of predicted means, the neighborhood was abandoned and the respondent with the closest Mahalanobis distance was selected as the donor.

The construction of the predictive mean vectors for the drug families mentioned in Section 6.5.5.1.3 was often complex. The main reason for the complexity is that recency and frequency models were not fit for all child drugs. The predicted means from the models for the parent drug were often used as surrogates for the child drug predicted means. When constructing the predictive mean vectors, the following general principles were followed:

1. If both the parent drug recency and the child drug recency(ies) were missing, condition on the general recency category of the parent drug.
2. For smokeless tobacco, if both the chewing tobacco recency and the snuff recency were missing, condition on whichever was "more" missing. Specifically, condition the recency predictive mean vectors on the more general recency category. For example, if chewing tobacco recency was "not past month" and snuff recency was "not past year," condition on the chewing tobacco recency category because it is more general.
3. Condition all elements of the predictive mean vector on the same general recency level.

Table 6.6 Elements of Full Predictive Mean Vector

Drug Use Measure and Category of Interest	Predicted Mean	Substance
Recency of Use, Past Month[1]	*P*(past month user | lifetime user)	All substances
Recency of Use, Past Year but Not Past Month[1]	*P*(past year but not past month user | lifetime user)	All substances except pipes
Recency of Use, Past 3 Years but Not Past Year[1]	*P*(past 3 years but not past year user | lifetime user)	Tobacco products[2] only
12-Month Frequency of Use	*P*(use on a given day in the year | past year user)**P*(past year user | lifetime user)[3]	All substances except tobacco
30-Day Frequency of Use for Alcohol and Substances with Few Daily Users[4]	*P*(use on a given day in the month | past month user)**P*(past month user | lifetime user)[5]	All substances except cigarettes, chew,[6] snuff, pipes, and pills[7]
30-Day Frequency of Use for Substances with Many Daily Users (excluding Alcohol)	*P*(use on a given day in the month | past month user, not a daily user)**P*(not a daily user | lifetime user)**P*(past month user | lifetime user)[5]	Cigarettes, chewing tobacco, snuff
Daily Use	*P*(daily user | past month user)**P*(past month user | lifetime user)[5]	Cigarettes, chewing tobacco, snuff
30-Day Binge Drinking Frequency	*P*(drank 5 or more drinks on a given day in the past month | past month user)**P*(past month user | lifetime user)[5]	Alcohol only

[1] The final category for recency (lifetime but not past year or lifetime but not past 3 years) was not needed in the predictive mean vector because the multinomial probabilities added to 1, and this probability was determined by the other probabilities.

[2] "Tobacco products" includes cigarettes, cigars, chewing tobacco, and snuff.

[3] Interpreting the proportion of the year used as a probability of use on a given day in the year assumed that the probability of use on each day in the year was equal. However, this was not true. The violation of this assumption did not seriously affect the ability to find a reasonable variable to use for finding a neighborhood, and it did allow the predicted mean to be made conditional on what was known.

[4] Alcohol, with many daily users, was included in this group because the distribution did not show a severe drop-off from 30 days a month to 29 days a month, as was apparent with cigarettes, chewing tobacco, and snuff.

[5] Interpreting the proportion of the month used as a probability of use on a given day in the month assumed that the probability of use on each day in the month was equal, which was not true, in the same manner as the 12-month frequency of use (see note #3 within this table).

[6] "Chew" is short for "chewing tobacco."

[7] "Pills" includes pain relievers, tranquilizers, stimulants, and sedatives.

Table 6.7 Full Predictive Mean Vector for Sample Drugs

Drug Use Measure and Category of Interest	Drug			
	Tobacco Products[1]	Alcohol	Marijuana, Cocaine, Crack, Heroin, Inhalants, Hallucinogens	Pain Relievers, Stimulants, Sedatives, Tranquilizers
Recency of Use, Past Month Use	✓	✓	✓	✓
Recency of Use, Past Year but Not Past Month Use	✓	✓	✓	✓
Recency of Use, Past 3 Years but Not Past Year Use	✓			
12-Month Frequency of Use		✓	✓	✓
30-Day Frequency of Use	✓	✓	✓	
30-Day Binge Drinking Frequency		✓		

[1] "Tobacco products" includes cigarettes, cigars, and smokeless tobacco (chewing tobacco and snuff). The imputation of pipes was completed in the univariate step because only two recency categories (past month and not past month) and no frequency-of-use variables were available for pipes.

6.5.5.3 Final Recency-of-Use and Frequency-of-Use Variables

As with all other imputation-revised variables, the final imputation-revised recency-of-use and frequency-of-use variables were identified with the prefix IR, followed by a 5-letter identifier, where a 3-letter code identified the drug[89] and the final 2 letters identified the measure (RC = recency; FY = frequency of use in past 12 months; FM = frequency of use in past 30 days). Each IR variable was accompanied by two imputation indicators, one with an II prefix and the other with an II2 prefix. The levels for the II indicator were the standard levels used for all imputation-revised variables: 1 = questionnaire data; 2 = logically assigned; 3 = statistically imputed; and 9 = legitimate skip (where applicable). The II2 indicators contained more details, including information from the lifetime usage imputations indicating whether lifetime usage was imputed. The imputation indicator levels are provided in Table 6.8.

[89] The exception to this rule occurred with marijuana, which for historical reasons contained only a two-letter code (MJ). Marijuana variables therefore ended with a four-letter identifier, rather than a five-letter identifier.

Table 6.8 Detailed Imputation Indicators for Recency and Frequency of Use

Level	Measure	
	Recency of Use	Frequency of Use
1	Questionnaire data	Questionnaire data
2	Logically assigned	Logically assigned[1]
3	Lifetime usage imputed	Lifetime usage imputed
4	Edited recency = 9 (lifetime user)	Lifetime usage not imputed
5	Edited recency = 8 (past year user)	N/A
6	Edited recency = 19 (lifetime not past month user)	N/A
7	Edited recency = 14 (lifetime not past year user)	N/A
9	N/A	Legitimate skip

N/A = not applicable.

[1] The logically assigned cases for 12-month frequency of use were not all included in Level 2; some were included in Level 1. This occurred if the 12-month frequency of use was trimmed because of (1) 30-day frequency; (2) estimated 30-day frequency; or (3) month and year of first use.

6.6 Special Section: Core-Plus-Noncore Methamphetamine and Stimulants Lifetime Use and Recency of Use

New questions were added to the noncore special drugs module in the 2005 NSDUH to capture information from respondents who may have used methamphetamine but did not recognize it as a prescription drug and therefore did not report use in the core stimulants module. Additional follow-up items were included in the 2007 NSDUH to resolve inconsistencies between responses in the core stimulants module and responses in the noncore special drugs module. Questions were added for 12-month frequency, age at first use, and date of first use.

Findings from the methamphetamine analysis report (Ruppenkamp et al., 2007) showed that it would be important to use responses from the noncore special drugs module in order to determine the best estimate of the prevalence of methamphetamine use in NSDUH. Therefore, after the normal imputation processing of the drug variables was complete, new imputation-revised versions of lifetime use and recency-of-use variables for both methamphetamine and stimulants were created, which incorporated responses from the noncore special drugs module as well as the core module. These versions of the methamphetamine variables were presented in a special section in the 2007 detailed tables but not in the main tables showing the standard list of drugs.[90] For more information on the reporting of methamphetamine prevalence in the 2007 NSDUH, see Appendix B.4.6 of the national findings report (Office of Applied Studies, 2008). New imputation-revised variables were created using the new questions in the noncore section of the questionnaire on 12-month frequency, age at first use, and date of first use.

A detailed description of the creation of these imputation-revised variables follows. In general, the approach was the same as for normal processing, except that a different set of edited variables was used as the base for imputation.

[90] Available at http://www.oas.samhsa.gov/WebOnly.htm#NHSDAtabs.

6.6.1 Final Creation of Base Variables for Imputation

The edited recency-of-use variables MTHREC06 and STMREC06, created by the editing team, were used as a starting point for the final creation of the base variables for imputation. These variables are described in Kroutil et al. (2008). They are similar to METHREC and STIMREC, the edited recency-of-use variables used in normal processing, except that they incorporate responses from the noncore special drugs module and the core module.

The final base variable for imputation of lifetime use of methamphetamine was called EDMTHLIFE. It was created as follows:

EDMTHLIFE =

- 1 (lifetime user), if MTHREC06 was 1, 2, 3, 8, 9, 11, 12, or 13; else
- 2 (lifetime nonuser), if MTHREC06 was 81 or 91; else
- missing.

The final base variable for imputation of lifetime use of stimulants, EDSTMLIFE, was created in exactly the same manner.

The final base variable for imputation of recency of use of methamphetamine was called EDMTHREC. It was created as follows:

EDMTHREC =

- 1 (past month user), if MTHREC06 was 1, or if MTHREC06 was 11 and METHREC was not equal to 11; else
- 2 (past year but not past month user), if MTHREC06 was 2, or if MTHREC06 was 12 and METHREC was not equal to 12; else
- 3 (lifetime but not past year user), if MTHREC06 was 3 or 13; else
- MTHREC06.

This was done based on SAMHSA's directive to treat noncore responses as known rather than as logically assigned. Respondents with known responses were treated as item respondents, while respondents with logically assigned responses were treated as item nonrespondents for the purpose of determining which cases were eligible donors.

The final base variable for imputation of recency of use of stimulants, called EDSTMREC, was created in exactly the same manner.

The edited 12-month frequency-of-use variables MTHTOT07 and STMTOT06 were created in a similar manner to the variables MTHYRTOT and STMYRTOT. These variables were used in subsequent steps instead of MTHYRTOT and STMYRTOT.

6.6.2 Reimputation of Lifetime Use Indicators

Using EDMTHLIFE and EDSTMLIFE, the processing of the lifetime use indicators proceeded, as described in Section 6.3. The set of item respondents did not change between the original imputation of the lifetime indicators and the reimputation of the lifetime indicators; therefore, it was not necessary to readjust the weights for item. As shown in Table 6.3, the stimulants lifetime drug use indicator was modeled toward the end of the hierarchy. So lifetime models were refit for stimulants, sedatives, cocaine, crack, and heroin, and provisional imputations were completed. After these models were refit, all the lifetime use indicators were reimputed. However, the only imputation-revised lifetime use variables used in further processing were the ones for stimulants and methamphetamine.

6.6.3 Reimputation of Recency of Use

Using EDMTHREC, EDSTMREC, MTHTOT07, and STMTOT06 instead of METHREC, STIMREC, MTHYRTOT, and STMYRTOT, the processing of the recency and frequency data proceeded, as described in Section 6.5. Final recency-of-use and frequency-of-use variables for methamphetamine and stimulants were created. However, only the new recency-of-use variables were used in subsequent steps. No imputation indicators were created.

6.7 Age at First Use and Related Variables

Unlike the recency and 12-month frequency-of-use variables, age at first drug use was not statistically imputed in the surveys prior to 1999. Instead, missing values were excluded from subsequent analyses. However, as with the 30-day frequency, missing age-at-first-use values were imputed since the 1999 survey. Also, recent drug initiates (i.e., those whose current age was equal to or 1 year greater than the reported age at first use) were asked the year and month of their first use. To have this information for all users, both missing year and missing month of first use for less recent initiates (and recent initiates who did not report year and month of first use) were replaced by assigning values consistent with the respondent's current age, interview date, imputation-revised age at first use, and imputation-revised recency and frequency variables. To have complete date-of-first-use information, day of first use was randomly assigned for all users. The combined data gave the respondent's age at first use along with the date of first use. It is important to note that in addition to age at first use for cigarettes, those respondents classified as lifetime daily cigarette users also were asked their age at first daily cigarette use.

6.7.1 Age at First Use

The age-at-first-drug-use imputations followed the same general procedures as the imputation of other drug use measures. A linear regression model utilizing SUDAAN software was fitted using a logit transformation of the respondent's age at first drug use as a proportion of his or her current age as the response variable. UPMNs were formed using the predicted mean from the regression model. Each item nonrespondent's neighborhood was restricted by logical constraints and likeness constraints. From these neighborhoods, a final imputation-revised age at first use was created. In addition, a randomly assigned date (i.e., year, month, and day) of first use was constructed that remained consistent with the imputed age at first drug use and other drug use measures.

6.7.1.1 Hierarchy of Drugs

The first step in the imputation of age at first use was to determine the order in which drugs would be modeled. As with the other drug use measures, it was expected that age at first use of other drugs would be strong predictors of age at first use of each drug of interest. Therefore, a hierarchy was chosen in order to get the greatest benefit from using the previously imputed age-at-first-use values as predictors for the drug of interest. The hierarchy for age at first use was identical to the lifetime and recency/frequency-of-use hierarchy shown in Table 6.3.

6.7.1.2 Setup for Model Building and Hot-Deck Assignment

As with the imputation of other drug use measures, the file was broken into three age categories for the imputation of age at first use (12 to 17, 18 to 25, and 26 or older), and all subsequent procedures were performed separately within each of these age groups. To impute missing age at first use for each drug, it was necessary to define the eligible population. Using the imputed recency of use, the files were reduced to lifetime users for each drug. If a valid response was provided for the age-at-first-use measure,[91] the person was deemed an item respondent. Before modeling, the respondent weights were adjusted, using a response propensity model, to match the entire population of lifetime users. The following categorical covariates were included in the models: imputed recency of use for cigarettes, smokeless tobacco, cigars, pipes, alcohol, marijuana, cocaine, crack, heroin, hallucinogens, inhalants, pain relievers, tranquilizers, stimulants, and sedatives (where available, otherwise lifetime indicators were used); categorical age; race/ethnicity; gender; census region; and a CBSA indicator.[92]

6.7.1.3 Sequential Model Building

The response variable in the model for age at first use, before a normalizing transformation, was the age at first use as a proportion of the current age. The numerator in this proportion was an integer representing age at first use. However, because this integer was in fact a truncated version of the real age at first use, the value was made continuous by adding a random component between 0 and 1. Hence, expressing the proportion as $P_i = Y_i/N_i$, the numerator was given as

$$Y_i = \textit{Age at First Use}_i + \textit{Uniform(0,1) random number.}^{93}$$

The denominator in the proportion was the total age. The true age was known, based on the interview date and birth date. Expressing it in years rather than days required dividing by the number of days in the year:

$$N_i = (\text{Interview Date} - \text{Birth Date} + 1)/365.25.$$

[91] Respondents who reported age at first use of 1 or 2 were not included in the model.

[92] These variables were included in every model unless convergence problems arose. If this occurred, the model was reduced.

[93] In the event that the age at first use was equal to the age, Y_i was constrained so that it was equally likely to be anywhere on the interval [$\textit{Age at First Use}_i$, N_i]. Thus, Y_i was prevented from being greater than N_i.

After a weight adjustment, the empirical logit transformation was used as the response variable in a weighted linear univariate regression:

$$\log\left[\left(Y_i + 0.5\right)/\left(N_i - Y_i + 0.5\right)\right].$$

This transformation was nearly equivalent to the standard logit transformation:

$$Y_i^* = \log\left[P_i/\left(1 - P_i\right)\right],$$

which was not used because it might be unstable for respondents who started using at their current age. Variables included in the regression equation were modified 12-month and 30-day frequencies for the drug in question; modified versions of the imputed age at first drug use for previously imputed drugs; imputed recency of use for cigarettes, smokeless tobacco, cigars, pipes, alcohol, marijuana, cocaine, crack, heroin, hallucinogens, inhalants, pain relievers, tranquilizers, stimulants, and sedatives (where available, otherwise lifetime indicators were used); centered age; centered age squared; centered age cubed; gender; race/ethnicity; State rank (based on the recency variable, see Section 6.5.1); first-order interactions of centered age, centered age squared, gender, and race/ethnicity; marital status; education level; employment status; census region; and a CBSA indicator.[94] The modified variables for 12-month frequency of use (where applicable), 30-day frequency of use (where applicable), and age at first use (AFU) were defined as follows:

new12$_i$	=	0	if respondent did **not** use the i^{th} drug in the past 12 months
	=	12-month frequency	if respondent used the i^{th} drug in the past 12 months
new30$_i$	=	0	if respondent did **not** use the i^{th} drug in the past month
	=	30-day frequency	if respondent used the i^{th} drug in the past month
AFU$_i$	=	0	if respondent is **not** a lifetime drug user of the i^{th} drug
	=	age at first use	if respondent is a lifetime drug user of the i^{th} drug

Naturally, the full model for age at first use did not include the lifetime indicator for the drug in question because the model was built on users of this substance. A summary of the final models can be found in Appendix F.

6.7.1.4 Computation of Predicted Means and Univariate Predictive Mean Neighborhoods

From the final model, a predicted value (based on the *Y* variable) was computed for each user of the drug of interest, which was then back-transformed to produce a predicted age at first use. The imputation-revised age-at-first-use assignment was conducted using the UPMN imputation, where the predicted mean was the predicted age at first use. Again, this procedure defined a neighborhood of respondents by requiring that the respondents' predicted age-at-first-use values be within a certain relative distance (delta) of the nonrespondent's value. The value of

[94] These variables were included in every model unless small sample sizes precluded the use of such a large pool of covariates. If this occurred, the model was reduced.

delta was set so that donors were required to have a predicted age at first use within 5 percent of that of the item nonrespondent. If no donors were available with predicted means within 5 percent of the recipient's predicted mean, the neighborhood was abandoned and the respondent with the closest predicted age at first use was chosen as the donor.

6.7.1.5 Assignment of Imputed Values

Subject to the constraints described in the next section, separate assignments of provisional values were performed within each of the three age groups. The age at first use of the randomly selected donor was then transferred to the recipient.

6.7.1.6 Constraints on Univariate Predictive Mean Neighborhoods

As with all other drug use measures, imputations were conducted separately within each age group: 12 to 17, 18 to 25, and 26 or older. This could be considered a likeness constraint based on age, which was never loosened. In fact, recipients and donors were required to be the same age, if possible. If a donor of the same age was not found, the constraint eventually was reduced to a logical constraint, where the imputed age at first use was less than the recipient's age. A small delta also could be considered a likeness constraint, which could be loosened by enlarging or removing delta. Initially, the relative distance for determining age-at-first-use imputation neighborhoods (delta) was set so that any potential donor's predicted age at first use was within 5 percent of the recipient's predicted age at first use. Another likeness constraint, in addition to the match on age, required an approximate match on recency of use. The match was approximate because recipients who were past year users could have had donors who had used at any time in the past year (no distinction was made between past month and past year but not past month use). Finally, an attempt was made to require donors and recipients to be from States with similar usage levels, where usage was defined in terms of the prevalence of past month usage of the drug in question.

These likeness constraints for age at first use were more stringent than those for the other drug use measures. Therefore, it was often necessary to loosen the constraints. The order of loosening constraints occurred as follows: (1) removed the State-rank group; (2) abandoned the neighborhood and chose the donor with the closest predicted mean; (3) loosened the restriction requiring an approximate match on recency of use and instead required only that recipients who did not use in the past year had donors who also did not use in the past year (tobacco recipients who did not use in the past 3 years had donors who did not use in the past 3 years); (4) loosened the restriction that donors and recipients had to be the same age and instead required that the donor's age had to be greater than or equal to the recipient's age and the donor's age at first use had to be less than or equal to the recipient's age at first use;[95] and (5) loosened the "same-age" restriction even further so that the donor's age at first use was only required to be less than or equal to the recipient's age. A summary of the above constraints and the number of respondents with sufficient donors corresponding to each likeness constraint are listed for each drug in Appendix G.

[95] With the loosening of the recency constraint, it was necessary to include a requirement that if the recipient was not a past year user, the age at first use could not have equaled the current age.

For drugs with no multivariate assignment, there were several logical constraints. For those respondents with an age at first use that equaled the recipient's current age, donors were excluded under the following circumstances. First, if the recipient's 12-month frequency was greater than the number of days since his or her last birthday, donors whose age at first use was equal to the recipient's current age were excluded. For example, suppose an item nonrespondent's birthday was on March 1 and the interview date was June 30. Then the number of days between the interview date and the respondent's birthday (inclusive) is 122. If the respondent had a 12-month frequency of 140 (either reported or imputed), his or her age at first use could not be his or her current age. Second, if the respondent's recency of use indicated that he or she did not use in the past month, but the number of days since his or her last birthday was fewer than 30, the recipient's age at first use could not be equal to his or her current age. And third, if the respondent was not a past month user, but the difference between his or her 12-month frequency and the days since his or her last birthday was fewer than 30, the recipient's age at first use could not be equal to his or her current age. Consider again the example where the respondent's birthday was on March 1, the interview was on June 30, and the number of days between the interview date and the respondent's birthday (inclusive) is 122. If the respondent's recency of use indicated past year but not past month use, but his or her 12-month frequency was 111, some of those 111 days had to have occurred before his or her birthday, and the respondent's age at first use could not have equaled his or her current age. In addition, respondents with age-at-first-use values of 1 or 2 years were not eligible to be donors. Finally, cigarettes had yet another logical constraint: if the recipient was a daily cigarette user and his or her age at first daily use was not missing, the donors were prevented from having an age at first use later than the preexisting age at first daily use.

6.7.1.7 Multivariate Assignments

For smokeless tobacco (chewing tobacco and snuff), cocaine (crack), hallucinogens (LSD, PCP, and Ecstasy), pain relievers (OxyContin), and stimulants (methamphetamine), more than one age-at-first-use variable was associated with a single predicted age at first use. This led to a multivariate assignment of the imputed values. Drugs where multivariate assignments were necessary are discussed in the following sections.

6.7.1.7.1 Smokeless Tobacco (Chewing Tobacco and Snuff)

For reasons discussed in Section 6.3.7.1, one model for smokeless tobacco was fitted rather than individual models for chewing tobacco and for snuff. The nearest neighbor hot-deck neighborhood was then based on the overall smokeless tobacco predicted age at first use. Missing age-at-first-use values for chewing tobacco and/or snuff were replaced with the values from a donor within this neighborhood. Only missing values were replaced, and, if both chewing tobacco and snuff were missing, imputed values came from the same donor. As for the constraints on the neighborhoods, all the constraints listed in the previous section were applied to both snuff and chewing tobacco separately. The likeness constraints also were applied to both chewing tobacco and snuff separately, but when loosened, they were loosened for chewing tobacco and snuff simultaneously. It is important to note that, for both chewing tobacco and snuff, lifetime usage was considered known (employing the lifetime usage imputation) so that there was no question of use versus nonuse. If age at first use was missing for chewing tobacco or snuff in the original data, but the respondent was imputed to be a nonuser of chewing tobacco

or snuff in the lifetime imputation, the respondent's age at first chewing tobacco use or age at first snuff use would be adjusted to reflect the situation. Age at first use for smokeless tobacco was obtained by taking the minimum age at first use from chewing tobacco and snuff.

6.7.1.7.2 Cocaine and Crack

Even though cocaine and crack were in distinct modules, an age-at-first-use model was fitted for only cocaine. The nearest neighbor hot-deck neighborhood was then based on the overall predicted age at first use for cocaine. Missing age-at-first-use values for cocaine and/or crack were replaced with the values from a donor within this neighborhood. Only missing values were replaced, and, if both cocaine and crack were missing, the imputed values came from the same donor. As for the constraints on the neighborhoods, all the constraints listed in the previous Section 6.7.1.6 were applied to both cocaine and crack separately. For example, donors for cocaine were logically restricted so that, if the recipient's 12-month cocaine frequency was greater than the number of days since his or her last birthday, donors whose ages at first cocaine use were equal to the recipient's age were excluded. The same was true for crack. The likeness constraints also were applied to both cocaine and crack separately, but, when loosened, they were loosened for cocaine and crack simultaneously. It is important to note that, for both cocaine and crack, lifetime usage was considered known (employing the lifetime usage imputation) so that there was no question of use versus nonuse. If age at first use was missing for crack in the original data, but the respondent was imputed to be a nonuser of crack in the lifetime imputation, the respondent's age at first crack use would be adjusted to reflect the situation.

Because crack is a type of cocaine, additional logical constraints were required so that donated values would be consistent with preexisting nonmissing values. Specifically, if the crack age at first use was missing, but cocaine age at first use was not, the donated crack age at first use could not be earlier than the preexisting cocaine age at first use. Conversely, if the cocaine age at first use was missing and crack age at first use was not, the donated cocaine age at first use could not be later than the preexisting crack age at first use. Finally, if crack age at first use was missing, but the respondent was a crack user, the donor had to be a crack user.

6.7.1.7.3 Hallucinogens (LSD, PCP, Ecstasy, and Other Hallucinogens)

The hallucinogens module consisted of many subgate questions, and three substances—LSD, PCP, and Ecstasy—were child drugs. One model was fitted for hallucinogens' age at first use from which a single neighborhood was created for LSD, PCP, Ecstasy, and hallucinogens as a whole. The nearest neighbor hot-deck neighborhood was then based on the overall hallucinogens' predicted age at first use. Missing ages at first use for any or all of LSD, PCP, Ecstasy, and hallucinogens as a whole were replaced with the values from a donor within this neighborhood.[96] Only missing values were replaced, and, if any of the four ages at first use were missing, the imputed values came from the same donor. As for the constraints on the neighborhoods, the constraints listed in the previous section were all applied to hallucinogens as a whole. Because no 12-month frequency was available for the child drugs, it was not possible to implement any constraints on these drugs involving the 12-month frequency.

[96] A donor could not be found within the respondent's age group (26 or older) who met all the logical constraints after all the likeness constraints had been loosened. Therefore, the age at first use was randomly assigned within the appropriate range.

Because of the parent/child relationship, additional logical constraints were required so that donated values were consistent with preexisting nonmissing values. For example, if the ages at first use for LSD and PCP were missing, but the ages at first use for overall hallucinogens and Ecstasy were not, the donated LSD and PCP ages at first use could not be earlier than the preexisting overall hallucinogens age at first use (but the LSD and PCP ages at first use could be earlier than the Ecstasy age at first use). Another example is if the age at first use for hallucinogens was missing and the LSD age at first use was not (and the respondent was a nonuser of both PCP and Ecstasy), then the donated overall hallucinogens age at first use could not be later than the preexisting LSD age at first use. In addition, if any of the child ages at first use were missing, but the respondent was a user, then the donor also had to be a user. Finally, if the respondent used one or more of the child drugs, but used no "other" type of hallucinogen, then his or her overall hallucinogens age at first use was imputed (or assigned) to be equal to the minimum of the child ages at first use.

All of the constraints applied specifically to the child drugs were logical constraints. It is important to note that, for both the parent and child drugs, lifetime usage was considered known (employing the lifetime usage imputation) so that there was no question of use versus nonuse. If an age at first use was missing for one or more of the child drugs in the original data, but the respondent was imputed to be a nonuser of any of these drugs in the lifetime imputation, then the respondent's age at first use would be adjusted to reflect the situation.

6.7.1.7.4 Pain Relievers (OxyContin and "Other" Pain Relievers)

For pain relievers, OxyContin was a child drug. One model was fitted for age at first use of pain relievers from which a single neighborhood was created for both OxyContin and overall pain relievers. The nearest neighbor hot-deck neighborhood was then based on the overall pain relievers' predicted age at first use. Missing ages at first use for OxyContin and/or overall pain relievers were replaced with the values from a donor within this neighborhood. Only missing values were replaced, and, if both OxyContin and overall pain relievers were missing, the imputed values came from the same donor. As for the constraints on the neighborhoods, the constraints listed in the previous section were all applied to overall pain relievers.

As for hallucinogens, additional logical constraints were required to account for the parent/child relationship. Specifically, if the age at first use for OxyContin was missing, but overall age at first use of pain relievers was not, then the donated age at first use of OxyContin could not be earlier than the preexisting age at first use of pain relievers. Conversely, if the age at first use of pain relievers was missing and the age at first use of OxyContin was not, then the donated age at first use of pain relievers could not be later than the preexisting age at first use of OxyContin. In addition, if the age at first use of OxyContin was missing, but the respondent was an OxyContin user, then the donor had to be an OxyContin user. Finally, if the respondent used OxyContin, but used no "other" type of pain reliever, then the overall pain reliever age at first use was imputed (or assigned) to be the same value as the OxyContin age at first use. All of the constraints applied specifically to OxyContin were logical constraints. It is important to note that, for both pain relievers and OxyContin, lifetime usage was considered known (employing the lifetime usage imputation) so that there was no question of use versus nonuse. If age at first use was missing for OxyContin in the original data, but the respondent was imputed to be a

nonuser of OxyContin in the lifetime imputation, then the respondent's age at first use of OxyContin would be adjusted to reflect the situation.

6.7.1.7.5 Stimulants (Methamphetamine and "Other" Stimulants)

The handling of the age-at-first-use variables for methamphetamine and overall stimulants was very similar to the procedures for OxyContin and overall pain relievers, as described in the previous section.

6.7.1.7.6 Core-Plus-Noncore Methamphetamine

The edited age-at-first-use variables MTHAGE07, MTHYFU07, and MTHMFU07 were created in a similar manner to the variables METHAGE, METHYFU, and METHMFU. These variables were used in subsequent steps instead of METHAGE, METHYFU, and METHMFU.

6.7.1.8 Year-of-First-Use, Month-of-First-Use, and Day-of-First-Use Assignments

After the age-at-first-use imputations, all lifetime users of a given drug had nonmissing age-at-first-use values. Using this age at first use (AFU), users were assigned year/month/day of first use values. Recent initiates, or those respondents whose AFU was within 1 year of his or her age, were asked for their year of first use (YFU) and month of first use (MFU). The day of first use (DFU) was not collected in the questionnaire and was missing for all respondents. The YFU, MFU, and DFU data contained four patterns of missingness:

1. *Recent initiates*: missing day of first use only;
2. *Recent initiates*: missing month/day of first use;
3. *Recent initiates*: missing year/month/day of first use; and
4. *Less recent initiates*: missing year/month/day of first use.

For each missingness pattern, bounds on both the earliest possible date of first use and the latest possible date of first use were determined. The final earliest possible date of first use was equal to the maximum of its bounds, and the final latest possible date of first use was equal to the minimum of its bounds. Once the earliest and latest possible dates of first use were determined, a day was randomly selected from this interval. The imputation-revised month/day/year values were then extracted from this date of first use.

6.7.1.8.1 Missingness Pattern 1

In this missingness pattern, the respondent provided all the information asked by the questionnaire (i.e., both the MFU and YFU). However, to obtain a complete date of first use, a DFU also was needed. Thus, a DFU was randomly assigned, given the respondent's month and year of first use, in a way that was consistent with both the 12-month frequency/recency and age at first use. Below is a brief description of the process used to obtain a date of first use in such cases. The imputed YFU, MFU, and DFU were extracted from the date, as defined below:

Final date of first use = Earliest possible date + [(*Days between earliest and latest date*)*(a random number generated from a Uniform (0,1) distribution)],

where

Days between earliest and latest = *Latest possible date – Earliest possible date* + 1;

Earliest possible date = maximum [(AFUth birthday), (first day of the month indicated by MFU/YFU)]; and

Latest possible date =

- minimum [(Interview date – 12-month frequency + 1), (1 day before the (AFU + 1)th birthday), (last day of the month indicated by MFU/YFU)], *if recency* = 1;
- minimum [(Interview date – 29 – 12-month frequency), (1 day before the (AFU + 1)th birthday), (last day of the month indicated by MFU/YFU)], *if recency* = 2; or
- minimum [(Interview date – 1 year), (1 day before the (AFU + 1)th birthday), (last day of the month indicated by MFU/YFU)], *if recency* = 3.

Note that it is impossible for recent initiates to have recency = 4 (lifetime but not past 3 years). Recent initiates had to have begun using the drug no earlier than their (AFU)th birthday. Because AFU = current age, or AFU = current age – 1, their (AFU)th birthday was within the past 2 years. Respondents who had begun using the drug within the past 2 years must logically have last used the drug within the past 2 years, and therefore could not have had recency = 4.

In rare cases, the *earliest possible date* was set to 29 days before the interview. This occurred for respondents meeting all of the following conditions:

1. The *latest possible date* was within 29 days of the interview.
2. The *earliest possible date* determined by the above rule was within a year of the interview.
3. The recency = 1.
4. The 12-month frequency = 30-day frequency (if applicable), or the 12-month frequency = 1.

Logically, all the lifetime usage of the drug for these respondents occurred in the past 30 days (including the interview date). The first condition ensures that the application of this rule will not cause an inconsistency. The second condition implies that the drug was not used by these respondents more than 1 year ago. The third and fourth conditions imply that the drug was not used by these respondents in the interval (1 year before the interview, or 1 month before the interview). Therefore, these respondents did not use the drug more than 1 month ago. All their lifetime use must have occurred in the past month.

6.7.1.8.2 Missingness Pattern 2

The second missingness pattern occurred when a recent initiate provided his or her YFU, but did not provide an MFU. In such cases, a month and day were randomly assigned that were

consistent with both the respondent's frequency/recency and with the age-at-first-use range. The imputed MFU and DFU were derived in the same manner as the date of first use in Missingness Pattern 1, except with the following changes:

- For the *earliest possible date*, replace "first day of the month indicated by MFU/YFU" with "January 1st of the YFU."
- For the *latest possible date*, replace "last day of the month indicated by MFU/YFU" with "December 31st of the YFU."

6.7.1.8.3 Missingness Pattern 3

Similar to Missingness Pattern 2, the third missingness pattern occurred when recent initiates provided neither an MFU nor a YFU value. In these cases, the year/month/day of first use were randomly assigned from a uniform distribution in a way that was consistent with both the 12-month frequency/recency and the age at first use. Again, the imputed YFU, MFU, and DFU were derived in the same manner as described in Missingness Pattern 1.

6.7.1.8.4 Missingness Pattern 4

The fourth missingness pattern occurred when the respondent reported, or was imputed to, an age at first use at least 2 years less than his or her age. This case is analogous to data prior to the 1999 survey, where month and year of first use were not asked in the questionnaire. In this missingness pattern, the frequency (or frequencies) was immaterial to the final date of first use because the respondent could not have begun using in the past year:

Earliest possible date = AFUth birthday; and

Latest possible date =

- 1 day before the (AFU + 1)th birthday, *if recency* < 4; or
- minimum [(Interview date – 3 years), (1 day before the (AFU + 1)th birthday)], *if recency* = 4.

6.7.1.8.5 Exceptions to the Standard Assignment of the Date of First Use

Although most of the drugs followed the standard assignment of the date of first use, a few exceptions occurred. The tobacco products (cigarettes, cigars, chewing tobacco, and snuff) did not have a 12-month frequency. As a result, the 30-day frequency was used whenever possible. This affected only the *latest possible date*, which was defined for these drugs as follows:

Latest possible date =

- minimum [(Interview date – 30-day frequency + 1), (1 day before the (AFU + 1)th birthday)], *if recency* = 1;
- minimum [(Interview date – 30), (1 day before the (AFU + 1)th birthday)], *if recency* = 2;

- minimum [(Interview date – 1 year), (1 day before the (AFU + 1)th birthday)], *if recency* = 3; and
- minimum [(Interview date – 3 years), (1 day before the (AFU + 1)th birthday)], *if recency* = 4.

Another variation occurred with the smokeless tobacco date of first use. In this case, the minimum of the chewing tobacco and snuff dates was used to produce the smokeless tobacco date of first use.

For all child drugs (daily cigarettes, LSD, PCP, ecstasy, OxyContin, methamphetamine, and crack), the corresponding parent drug's date of first use was assigned first. Then, in the setting of the *earliest possible date* for the child drug, the parent drug's date of first use was used as an additional bound. This was done to ensure that the child drug's date of first use was never earlier than the parent drug's date of first use.

For all parent drugs whose child drugs had recency and frequency information (hallucinogens, pain relievers, stimulants, and cocaine), the child drug recency and frequency information was used to bound the *latest possible date*. For example, respondents with LSD recency = 3 could not have first used hallucinogens within the past year, regardless of the hallucinogens recency value. The bound effected by the child drug recency and frequency was calculated in exactly the same way as for the parent recency and frequency information (see Section 6.7.1.8.1).

For hallucinogens, pain relievers, and stimulants, an indicator of lifetime use of drugs other than the child drugs was created (see Table 6.2). For pain relievers and stimulants, if the respondent was not a lifetime user of the "other" drugs, then the child drug's date of first use was logically assigned to the parent drug's date of first use. The handling of the child drugs for hallucinogens was more complex, because there was more than one of them. The algorithm follows:

1. The date of first use was assigned for overall hallucinogens.
2. The *earliest possible date*, *latest possible date*, and the final date of first use for each child drug for which the respondent was a lifetime user were assigned.
3. For respondents who were lifetime nonusers of other hallucinogens:
    a. It was determined which, if any, child drug could have had the same date of first use as hallucinogens. Specifically, it was determined whether the date of first use for hallucinogens was between *earliest possible date* and *latest possible date* for each child drug.
    b. If none of the child drugs were eligible to receive the hallucinogens date of first use, nothing was done. Otherwise, one of the eligible child drugs was chosen at random, and its date of first use was overwritten with the hallucinogens date.

6.7.1.8.6 Final Date-of-First-Use Variables

As with all other imputation-revised variables, the final imputation-revised date-of-first-use variables were identified with the prefix IR, followed by a six-letter identifier, where a three-letter code identified the drug[97] and the final three letters identified the measure (AGE = age at first use, YFU = year of first use, MFU = month of first use, DFU = day of first use). Each IR variable was accompanied by an imputation indicator with the requisite II prefix. The levels for the imputation indicators were the standard levels used for all imputation-revised variables: 1 = questionnaire data; 2 = logically assigned; 3 = statistically imputed; and 9 = legitimate skip (not a lifetime user).

6.7.2 Imputations for Age at First Daily Cigarette Use

In addition to age at first use, the cigarettes module also included a question asking for the respondent's age at first daily cigarette use, where a daily user was defined as someone who reported having at some time smoked cigarettes every day for a period of at least 30 days. Imputation procedures for age at first daily cigarette use were similar to age at first use, with two key exceptions.

The first exception involved the domain of the age-at-first-use variable. Whereas the age-at-first-use question was asked of all cigarette users, the age-at-first-daily-use question was asked of only daily users. The "daily use" indication came from two sources. If a respondent answered either the 30-day frequency or estimated 30-day frequency with a "30," or if the respondent had a "yes" value for the edited variable associated with the "ever daily used" question (CIGDLYMO), then he or she was considered a daily user. For more information about CIGDLYMO, see Kroutil and Handley (2008) and Kroutil et al. (2008). At this stage in the process, there should be no missing responses to the 30-day frequency question. Daily users, based on 30-day frequency, should be either known (based on a response in the survey) or imputed. However, missing responses for the ever-daily-used question also should be imputed. The second exception involved the predicted means. Because of the high correlation between age at first use and age at first daily use, models for age at first use were used to define the imputation neighborhoods for age at first daily use.

Thus, the age-at-first-daily-use imputation involved two parts: The first part involved missing values in the ever-daily-used variable (CIGDLYMO). The second part involved all missing age-at-first-daily-use values for eligible daily users, including those that were imputed to "ever daily used."

6.7.2.1 Setup for Model Building—Ever-Daily-Used Variable (CIGDLYMO)

Because age at first daily use was asked of all persons who answered the ever-daily-used question with a "yes," it was necessary to ensure that this question had no missing values. As with all other drug use imputations, the file was broken into three age categories (12 to 17, 18 to

[97] Exceptions to this rule occurred with marijuana and cigarette daily use. For historical reasons, marijuana contained a two-letter code (MJ). Marijuana variables therefore ended with a five-letter identifier, rather than a six-letter identifier. The code for cigarette daily use was CDU, which differed from the general cigarette code of CIG. Details about cigarette daily use are provided in Section 6.7.2.9.

25, and 26 or older), and all subsequent procedures were performed separately within these age groups. To impute for missing values in the ever-daily-used variable, it was necessary to define the eligible population—respondents who had an imputation-revised 30-day frequency[98] of fewer than 30 days (includes legitimate skip codes for lifetime, but not past month users). If a valid response was provided in the ever-daily-used variable, the person was deemed an item respondent. Before modeling, the item respondent weights were adjusted to match the entire eligible population. This adjusted weight was computed using a response propensity model and included the following categorical covariates: imputed recency of use for cigarettes; the lifetime indicators for smokeless tobacco, cigars, pipes, alcohol, marijuana, cocaine, crack, heroin, hallucinogens, inhalants, pain relievers, tranquilizers, stimulants, and sedatives; categorical age; race/ethnicity; gender; census region; and a CBSA indicator.

6.7.2.2 Model Building—Ever-Daily-Used Variable (CIGDLYMO)

After the weights were adjusted, the ever-daily-used variable was modeled using weighted logistic regression in SUDAAN. The predicted mean from this model was the predicted probability of ever smoking cigarettes daily. Variables included in the initial regression equation were a revised 30-day cigarette frequency variable (in the same format as used in the age-at-first-use models; see Section 6.7.1.3); the imputation-revised cigarette age at first use; imputed recency of use for cigarettes; the lifetime indicators for smokeless tobacco, cigars, pipes, alcohol, marijuana, cocaine, crack, heroin, hallucinogens, inhalants, pain relievers, tranquilizers, stimulants, and sedatives; centered age; centered age squared; centered age cubed; gender; race/ethnicity; State rank (based on the recency variable); first-order interactions of centered age, centered age squared, gender, and race/ethnicity; census region; a CBSA indicator; marital status; education level; and employment status.

6.7.2.3 Computation of Predicted Means and Univariate Predictive Mean Neighborhoods—Ever-Daily-Used Variable (CIGDLYMO)

From the final model, a predicted mean of the ever-daily-used variable was computed for each eligible respondent. The assignment of imputation-revised ever-daily-used values was conducted using UPMN imputation, where the "predicted mean" was the predicted probability of daily use at some point in the respondent's lifetime, given that the respondent was a lifetime user, but not a current daily user. Again, the procedure defined a "neighborhood" of respondents (i.e., potential donors) by requiring that a respondent's predicted ever-daily-used probability be within a certain relative distance, delta, of the nonrespondent's predicted probability. Delta was set so that donors were required to have a predicted probability within 5 percent of that of the item nonrespondent.

6.7.2.4 Assignment of Imputed Values—Ever-Daily-Used Variable (CIGDLYMO)

Separate assignments were performed within each of the three age groups, subject to the constraints described in the next section. The ever-daily-used response of the randomly selected donor was then transferred to the recipient.

[98] The imputation-revised 30-day frequency included responses from the 30-day frequency question (CG07), as well as the estimated 30-day frequency question (CG07DKRE).

6.7.2.5 Constraints on Univariate Predictive Mean Neighborhoods—Ever-Daily-Used Variable (CIGDLYMO)

As with all other drug use measures, neighborhoods for the ever-daily-used variable were restricted so that candidate donors and recipients were in the same age group (12 to 17, 18 to 25, and 26 or older). Models were built separately within these three groups, so this likeness constraint was never loosened. The likeness constraints were nearly identical to those of age at first use (see Section 6.7.1.6). The only difference was in the definition of the predicted mean, the determination of which was described in Section 6.7.2.2.

6.7.2.6 Computation of Predicted Means and Univariate Predictive Mean Neighborhoods—Age at First Daily Cigarette Use

Instead of separately modeling age at first daily cigarette use, the predicted means from the age-at-first-cigarette-use models were used to determine neighborhoods. The imputation-revised age-at-first-daily-use assignment was conducted using UPMN imputation. The procedure defined a "neighborhood" of respondents by requiring that the respondent's predicted mean be within a certain relative distance (delta) of the nonrespondent's predicted mean.

6.7.2.7 Assignment of Imputed Values—Age at First Daily Cigarette Use

Separate assignments were performed within each of the three age groups, subject to the constraints described in the next section. The age at first daily use of the randomly selected donor was then transferred to the recipient.

6.7.2.8 Constraints on Univariate Predictive Mean Neighborhoods—Age at First Daily Cigarette Use

As with all other drug use measures, neighborhoods for age at first daily use were restricted so that candidate donors and recipients were in the same age group (12 to 17, 18 to 25, and 26 or older). The likeness constraints were nearly identical to those used for age at first use (see Section 6.7.1.6). There was only one difference: An additional step was employed if no donor was found after loosening all of the likeness constraints. In particular, if the age at first use and age at first daily use were both initially missing, the imputed age at first use was set back to missing and reimputed simultaneously with the age at first daily use so that both were mutually consistent.[99] A summary of the above constraints and the number of respondents who fitted into each one are listed for each drug in Appendix G.

All the logical constraints applied to age at first cigarette use also were applied to age at first daily cigarette use. In other words, simply replace the words "age at first use" with "age at first daily use" in Section 6.7.1.6. Besides those logical constraints, an additional logical constraint was applied specifically to age at first daily cigarette use. If the age at first use for a recipient with a missing age at first daily use was not missing, the donors were prevented from having an age at first daily use earlier than the preexisting age at first use.

[99] In the 2007 survey, the situation where no donors were available, even after loosening all constraints, never occurred. It has occurred in past NSDUHs, however, and the programming code still exists in case the situation occurs in future NSDUHs.

6.7.2.9 Asssignments of Date of First Daily Cigarette Use

After the imputation-revised age at first daily cigarette use was created, all daily cigarette users had a valid age at first daily cigarette use. From this age, a year/month/day of first daily use was assigned. The date assignment procedure was identical to the procedure described in Section 6.7.1.8, using the same exceptions noted in Section 6.7.1.8.5 for tobacco products and child drugs.

6.8 Recodes

Numerous recoded variables were created from imputation-revised versions of the drug measures. Many of these were created from the recency-of-use variables, and several were also created from the age-at-first-use and date-of-first-use variables. These variables were created mainly to facilitate the creation of the 2007 detailed tables.

6.8.1 Prevalence Recodes

From every imputation-revised recency variable, three dichotomous recodes were created: one for lifetime use, one for past year use, and one for past month use. The first part of the variable name was a three-letter abbreviation for the drug. The second part of the variable name identified the measure: FLAG for lifetime recodes, YR for past year recodes, and MON for past month recodes. The creation of these variables was straightforward. For example,

INHYR =

- 1, if IRINHRC was equal to past month use or past year but not past month use; else
- 0.

Several other prevalence recodes, which covered the same three measures, were created to incorporate information from several different drugs. Table 6.9 lists these recodes and the recency variables that were used to create them. The creation of these variables was also straightforward. If the respondent was a lifetime user of any of the drugs, then the FLAG variable was set to 1; else it was set to 0. The YR and MON variables were similar.

Table 6.9 Prevalence Recodes Incorporating More than One Recency Variable

General Drug Category	Variable Names	Source Recency Variables
Tobacco	TOBFLAG, TOBYR, TOBMON	Cigarettes, smokeless tobacco, cigars, pipes
Psychotherapeutics	PSYFLAG2, PSYYR2, PSYMON2[1]	Pain relievers, tranquilizers, stimulants, sedatives
Illicit Drugs Other than Marijuana	IEMFLAG, IEMYR, IEMMON	Psychotherapeutics, plus inhalants, hallucinogens, cocaine, and heroin
Illicit Drugs, but Only Marijuana	MJOFLAG, MJOYR, MJOMON	Same as MRJFLAG, MRJYR, and MRJMON, except set to 0 if the corresponding IEM variable is equal to 1
Illicit Drugs	SUMFLAG, SUMYR, SUMMON	Illicit drugs other than marijuana, plus marijuana

[1] These variable names include a suffix of "2" to distinguish them from earlier versions of psychotherapeutics recodes.

Similar recodes also were created from the core-plus-noncore methamphetamine and stimulants recency-of-use variables described in Section 6.6.3. The core-plus-noncore methamphetamine recodes were CPNMTHFG, CPNMTHYR, and CPNMTHMN. The core-plus-noncore stimulants recodes were CPNSTMFG, CPNSTMYR, and CPNSTMMN. The core-plus-noncore psychotherapeutic recodes were CPNPSYFG, CPNPSYYR, and CPNPSYMN. No core-plus-noncore versions of the IEM or SUM recodes described in Table 6.9 were created for use in the detailed tables, even though the prevalence estimates would likely increase slightly if the noncore methamphetamine data were incorporated.

6.8.2 Incidence Recodes

Incidence recodes were created for PSY, IEM, and SUM using the age-at-first-use variables and date-of-first-use variables for the same specific drugs described in Table 6.9. The age-at-first-use recodes were simply set to the minimum of the source age-at-first-use variables, and they were named with the suffix AGE: PSYAGE2, IEMAGE, and SUMAGE. For example,

PSYAGE2 = minimum of IRANLAGE, IRTRNAGE, IRSTMAGE, and IRSEDAGE.

To set the date-of-first-use variables, the earliest date of first use was found among the source variables for which the respondent was a lifetime user, and the new YFU, MFU, and DFU variables were determined using the YEAR, MONTH, and DAY functions in SAS®. For example,

PSYYFU2 = YEAR (minimum of dates of first use of pain relievers, tranquilizers, stimulants, and sedatives).

7. Nicotine Dependence

7.1 Introduction

The method used to measure dependence on nicotine in the 2007 National Survey on Drug Use and Health (NSDUH)[100] was first introduced in the 2001 survey and also was used in the 2002-2006 NSDUHs. The questions used in the 2007 survey were the same as those asked in other surveys since the 2001 NSDUH. As in the 2006 survey, only respondents who reported use of cigarettes in the past 30 days were asked these questions.

The method for determining nicotine dependence involved the calculation of a continuous scale, called the Nicotine Dependence Syndrome Scale (NDSS) (Shiffman, Hickcox, Gnys, Paty, & Kassel, 1995; Shiffman, Waters, & Hickcox, 2003). This scale was calculated from 17 NSDUH questionnaire items (Table 7.1) that were asked of respondents who used cigarettes in the past 30 days. Valid responses, which are described in Section 7.2, were required for each of the 17 questions. The scale was the mean value (appropriately adjusted where necessary) of the responses, provided all 17 responses were nonmissing.

Of the eligible respondents who did not answer all of the questions, the majority either were missing a response from only one of the questions or did not answer any of the questions. For the respondents missing only 1 of the 17 variables, imputation was used to fill in the values for the missing variable, using the information from the other 16 nonmissing variables, through weighted least squares regression models. This resulted in 17 regression models, 1 for each variable. Weighted least squares regression was not entirely appropriate for these data, because both the response variable and the covariates were ordinal variables, and least squares methods generally require the data to be continuous. However, the scale was calculated as a mean from ordinal variables, and the imputed values were used as only 1 value out of 17 in the calculation of an arithmetic mean. Any bias that might have resulted from using an inappropriate type of model would have had a minimal effect on the resulting NDSS.

The imputations described in this chapter are unique in this report because they were not performed using the predictive mean neighborhood (PMN) technique described in Appendix C. Also, the NDSS mean value was calculated from edited versions of the 17 nicotine dependence questionnaire variables. The majority of the editing procedures for these variables are described elsewhere (Kroutil & Handley, 2008).

[100] This report presents information from the 2007 National Survey on Drug Use and Health (NSDUH), an annual survey of the civilian, noninstitutionalized population of the United States aged 12 or older. Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA).

7.2 Edited Nicotine Dependence Variables

Table 7.1 shows the correspondence between the 17 questionnaire items used in the NDSS and the corresponding edited variables. Among eligible respondents (those who had used cigarettes in the past 30 days), the following valid responses for the edited variables, as with the raw variables, were required: 1 = Not at all true, 2 = Somewhat true, 3 = Moderately true, 4 = Very true, or 5 = Extremely true. For most nicotine dependence variables, "dependence" was marked by the "Extremely true" response. However, for question variables DRCGE04, DRCGE12, DRCGE13, and DRCGE14, "dependence" was marked by "Not at all true."

Table 7.1 Mapping of Raw Nicotine Dependence Question Variables to Edited Variables

Question Variable	Question Text	Edited Variable
DRCGE01	After not smoking for a while, you need to smoke in order to feel less restless and irritable.	CIGIRTBL
DRCGE02	When you don't smoke for a few hours, you start to crave cigarettes.	CIGCRAVE
DRCGE03	You sometimes have strong cravings for a cigarette where it feels like you're in the grip of a force you can't control.	CIGCRAGP
DRCGE04	You feel a sense of control over your smoking - that is, you can "take it or leave it" at any time.	CIGINCTL
DRCGE05	You tend to avoid places that don't allow smoking, even if you would otherwise enjoy them.	CIGAVOID
DRCGE07	Even if you're traveling a long distance, you'd rather not travel by airplane because you wouldn't be allowed to smoke.	CIGPLANE
DRCGE08	You sometimes worry that you will run out of cigarettes.	CIGRNOUT
DRCGE09	You smoke cigarettes fairly regularly throughout the day.	CIGREGDY
DRCGE10	You smoke about the same amount on weekends as on weekdays.	CIGREGWK
DRCGE11	You smoke just about the same number of cigarettes from day to day.	CIGREGNM
DRCGE12	It's hard to say how many cigarettes you smoke per day because the number often changes.	CIGNMCHG
DRCGE13	It's normal for you to smoke several cigarettes in an hour, then not have another one until hours later.	CIGSVLHR
DRCGE14	The number of cigarettes you smoke per day is often influenced by other things - how you're feeling or what you're doing, for example.	CIGINFLU
DRCGE15	Your smoking is not affected much by other things. For example, you smoke about the same amount whether you're relaxing or working, happy or sad, alone or with others.	CIGNOINF
DRCGE16	Since you started smoking, the amount you smoke has increased.	CIGINCRS
DRCGE17	Compared to when you first started smoking, you need to smoke a lot more now in order to be satisfied.	CIGSATIS
DRCGE18	Compared to when you first started smoking, you can smoke much, much more now before you start to feel anything.	CIGLOTMR

7.3 Imputation-Revised Nicotine Dependence Variables

7.3.1 Setup for Model Building

In general, imputation models for variable types other than nicotine dependence in the 2007 survey were modeled sequentially so that variables that were modeled early in the sequence could be used as covariates in models for variables later in the sequence. This was done to avoid fitting separate models for each missingness pattern. In the case of nicotine dependence, however, no imputation was performed if more than one NDSS variable was missing. As a result, for each respondent where imputation could be performed, all 16 nonmissing NDSS variables could be used as covariates in the model for the 17th missing variable. Therefore, no sequential modeling was necessary. Item respondents had to have complete data for all 17 of the NDSS questions used in the models, and logically they had to have used cigarettes in the past 30 days. Item nonrespondents were those who used cigarettes in the past 30 days and answered only 16 of the 17 NDSS questions with valid nonmissing responses. Respondents who had used cigarettes in the past 30 days and were therefore eligible to answer the NDSS questions but answered only 15 or fewer of those questions were left out of the modeling process. The missing values in the NDSS variables for these respondents remained missing in the imputation-revised variables. No response propensity adjustments were performed for the item respondent weights used in any of the models. However, the ratio-adjusted design-based weights were used in the imputation models. The variables included in the models are discussed in the next section.

7.3.2 Model Building

In the 2007 survey, one model was created for each NDSS variable. The response variable for each model was the edited variable that corresponded to the question text shown in Table 7.1. The covariates in each model were the remaining NDSS variables. For example, if CIGIRTBL was the response variable, then the covariates would be the remaining 16 NDSS variables: CIGCRAVE, CIGCRAGP, CIGINCTL, CIGAVOID, CIGPLANE, CIGRNOUT, CIGREGDY, CIGREGWK, CIGREGNM, CIGNMCHG, CIGSVLHR, CIGINFLU, CIGNOINF, CIGINCRS, CIGSATIS, and CIGLOTMR.

7.3.3 Computation of Predicted Means

If a respondent was missing only 1 of the 17 NDSS items, the predicted mean for this item was obtained using the coefficients corresponding to the other 16 nonmissing covariates from the appropriate weighted least squares regression. The covariates and the response variables were all ordinal, so it was possible for a predicted mean to exceed 5 or be less than 1. Section 7.2 describes the five valid responses.

7.3.4 Assignment of Imputed Values

For those respondents missing only 1 of the 17 NDSS items, the missing value was replaced by the predicted mean in the imputation-revised variable. No attempt was made to round the predicted mean, and no attempt was made to add a residual. The nicotine dependence imputation-revised variables were unique in that missing values remained as missing values if the respondent was eligible to answer the nicotine dependence questions but two or more NDSS

items were missing. The edited valid response was assigned for the remainder of respondents who answered all 17 nicotine dependence questions.

7.4 Summary Information for Nicotine Dependence Variables

Imputations were necessary for the nicotine dependence variables to create an NDSS score for as many eligible persons as possible. The imputation method was devised to be easy to implement, given the complexities of handling this type of missing data. To avoid complicated models, imputations were limited to cases where the respondent answered 16 of the 17 questions. If an eligible respondent answered fewer than 16 questions, no imputations were performed. It was possible that the respondent was eligible to answer the questions about nicotine dependence because he or she was imputed to be a past month cigarette user. Table 7.2 summarizes the eligibility of respondents to answer the nicotine dependence questions and reasons why respondents were eligible or not eligible. Furthermore, this table gives details about the amount of nicotine dependence data that was missing for eligible respondents. It also provides information on whether the respondent was imputed to be a past month cigarette user. Consequently, the respondent would be eligible to have nicotine dependence data but would have missing data for all the nicotine dependence variables.

Table 7.2 Summary of Response Patterns for NDSS Variables

Number of Missing NDSS Variables	Past Month Smoker	Past Month Smoker Status Imputed	Eligible to Answer NDSS Questions	NDSS Variables Imputed	Frequency
17	No	No	No	N/A[1]	50,795
17	No	Yes	No	N/A[1]	22
Subtotal					50,817
17	Yes	No	Yes	No	16
17	Yes	Yes	Yes	No	13
2-16	Yes	No	Yes	No	146
Subtotal					175
1	Yes	No	Yes	Yes	226
0	Yes	No	Yes	N/A[1]	16,652

N/A = not applicable; NDSS = Nicotine Dependence Syndrome Scale.

[1] The NDSS variables for this scenario were not missing.

8. Household Composition (Roster)

8.1 Introduction

This chapter describes the techniques used to edit inconsistent values in the household roster and the techniques used to create and impute missing values in the roster-derived household composition variables for the 2007 National Survey on Drug Use and Health (NSDUH).[101] The procedures used to create respondent-level detailed roster variables, roster-derived household composition variables, and roster-based proxy variables are included. The proxy variables allowed the selection and identification of a relative of the respondent who lived in the respondent's household (according to the household roster), who was aged 18 years or older, and who answered the health insurance coverage and income questions for the respondent. Imputations were accomplished using the predictive mean neighborhood (PMN) technique, which is described in Appendix C.

8.2 Household Roster Edits

8.2.1 Description of Household Composition (Roster) Section of Questionnaire

The introductory question for the household roster portion of the questionnaire (QD54) was interviewer administered. This question asked the respondent for information regarding the number of persons living in his or her household, where allowable entries ranged from 1 to 25. If either the interviewer indicated that the respondent lived alone or the question was unanswered, the household composition (roster) section was skipped. However, if the interviewer indicated a household size greater than 1, the interviewer was then prompted to ask the respondent questions about the age, gender, and relationship to the respondent of every member of the household, starting with the household's oldest member and including the respondent. If a pair of respondents was selected in a household, the interviewer indicated which member of a respondent's household roster corresponded to the other selected pair member. The roster entry for the respondent was referred to as the "self" entry. In effect, the respondent completed a grid with the number of rows corresponding to the value entered in QD54. Table 8.1 shows an example grid where the number of persons in the household is four. In this example, the roster of the wife/mother is shown and the indicator variable shows that the son was selected as the other pair member. The possible relationship codes and specific relationship details are listed in Table 8.2.

[101] This report presents information from the 2007 National Survey on Drug Use and Health (NSDUH), an annual survey of the civilian, noninstitutionalized population of the United States aged 12 or older. Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA).

Table 8.1 Household Composition (Roster) Grid Example Where Number of Persons in Household (QD54) Equals 4

Person #	Relationship to Respondent	Age in Years	Other Member Selected for Pair[1]
1	Self (Wife/Mother)	44	0 (No [Impossible])
2	Husband	42	0 (No)
3	Son	16	1 (Yes)
4	Boarder/Roomer	16	0 (No)

[1] This indicator variable applied only to respondents who were part of a pair selection. The other member selected could not have been the self because respondents were not interviewed twice. The other member selected was the roster member who had a value of "1" for this variable.

Table 8.2 Household Composition (Roster) Relationship Codes

Relationship Code #	Relationship to Respondent	Details about Relationship
1	Self	
2	Parent	Biological, Step, Adoptive, or Foster
3	Child	Biological, Step, Adoptive, or Foster
4	Sibling	Full, Half, Step, Adoptive, or Foster
5	Spouse	
6	Unmarried Partner	
7	Housemate or Roommate	
8	Child-in-Law	
9	Grandchild	
10	Parent-in-Law	
11	Grandparent	
12	Boarder or Roomer	
13	Other Relative	
14	Other Nonrelative	

8.2.2 Household Roster Consistency Checks

To reduce the amount of editing required during the data processing stage, consistency checks were included in the Blaise program code.[102] Two types of consistency checks were employed in the household roster section of the questionnaire. These checks (1) compared a roster entry corresponding to the respondent with previously entered questionnaire information or (2) compared a roster entry against other roster entries or the respondent's roster age for internal consistency.

[102] The Blaise program is the computer program within the computer-assisted interviewing (CAI) instrument that was used to direct the respondent and interviewer through the questionnaire.

8.2.2.1 Comparisons with Previously Entered Questionnaire Information

In the 2001 survey, a consistency check was added to the household roster section of the computer-assisted interviewing (CAI) instrument. This check was triggered if the interviewer reported a different gender for the respondent in the household roster than was previously recorded in the interview (question QD01). The interviewer was required to change either the roster entry or the gender that had been entered at the beginning of the interview. In the 2002 survey, a new consistency check involving the respondent's age was added. Not only was it necessary for the respondent's gender in the roster to match the questionnaire gender but also for the respondent's age in the roster to match the age that had been entered in the nonroster part of the questionnaire (the Blaise variable CURNTAGE). For the age check, the interviewer could either change the respondent's age entered in the roster or override the consistency check and provide an explanation as to why the roster age did not match CURNTAGE. Both of these consistency checks involved the respondent's own entry in the roster (the "self" entry). If the consistency check for age was overridden, the value for age corresponding to the self may not match the questionnaire-edited age. Explanations given by the interviewer for overriding this particular consistency check were carefully reviewed. In rare cases, the final value for age (AGE) was set to the age of the self in the questionnaire roster (the "roster age") based on these explanations, as well as other evidence. Additional details about how roster age was used are described in Chapter 4. Strategies for the more common situation where the original value for AGE was not set to the roster age are discussed in Section 8.2.4.

8.2.2.2 Internal Consistency Checks

Since the 2002 survey, internal consistency checks have been implemented for the household roster. These checks were triggered if any of the following conditions occurred:

1. The interviewer reported that the respondent had more than one spouse or unmarried partner or reported a spouse and an unmarried partner.
2. The interviewer reported that a household member was a parent or grandparent of the respondent and the respondent was older than the household member.
3. The interviewer reported that a household member was a child or grandchild of the respondent and the respondent was younger than the household member.
4. The interviewer reported that a household member was a spouse or an unmarried partner of the respondent and the household member was 16 years old or younger.
5. The interviewer reported that the respondent had a spouse or unmarried partner and the respondent was 16 years old or younger.
6. The interviewer reported that the respondent was either a child-in-law or a parent-in-law and the respondent was 16 years old or younger.
7. The interviewer reported that a household member was a child-in-law of the respondent and the household member was the same age or older than the respondent.
8. The interviewer reported that a household member was a parent-in-law of the respondent and the household member was the same age or younger than the respondent.

9. The interviewer reported that a household member was a biological parent of the respondent and the household member was less than 13 years older than the respondent.
10. The interviewer reported that a household member was a biological child of the respondent and the household member was less than 13 years younger than the respondent.
11. The interviewer reported that a household member was a biological sibling of the respondent and the household member was more than 24 years older or younger than the respondent.
12. The interviewer reported that a household member was a grandchild of the respondent and the respondent was 30 years old or younger.[103]
13. The interviewer reported that a household member was a grandparent of the respondent and the respondent was 60 years old or older.[3]

In the 2005 survey, a new consistency check was added that replaced checks #12 and #13 above. This new check was triggered if the following conditions occurred:

14. The interviewer reported that a household member was a grandparent or grandchild of the respondent and the age difference was less than 30 years.

In most cases, if the consistency check was triggered, the interviewer changed either an age code or a relationship code in the roster to a more appropriate value. Because new consistency checks may be introduced each year, fewer edits to the roster may be implemented each survey year. Nevertheless, any edit that was invoked because of an override to a consistency check was carefully scrutinized. The relevant household roster, as well as the explanation given by the interviewer for the override, was carefully examined to determine whether the override was legitimate. If the override was deemed legitimate (e.g., a father marries a woman, listed as [step] mother, who is younger than the respondent), the original answer was allowed to remain and the edit was not applied. If the interviewer's explanation was not considered legitimate, then the edit was applied. More details about roster edits are provided in Section 8.2.5. Explanations given by the interviewers for the overrides and evaluations of their legitimacy are provided in Appendix I.

8.2.3 Preliminary Roster Edits

To facilitate processing of the roster variables, a "roster level" file was created in which the number of records per respondent was given by the household size in question QD54. If the respondent quit the interview after the household size question or in the middle of the roster questions, "dummy" records were created that corresponded to the missing household members.

8.2.4 Roster Edits Involving the Self

The Blaise program code required the interviewer to identify exactly one "self" and a corresponding age and gender in the household roster. In theory, these values should have

[103] Consistency check #14 was added in 2005 to replace consistency checks #12 and #13.

matched CURNTAGE and QD01, respectively. Because the check involving gender was not allowed to be overridden, the gender for the self in the roster always matched QD01, which was equivalent to IRSEX (see Chapter 4). For the consistency check comparing the respondent's roster age against CURNTAGE, the age of the self in the roster should be close to the questionnaire-edited age, AGE (see Chapter 4 for a description of the methodology used to create AGE), especially if the respondent age was set to the roster age. Moreover, the interviewer was required to confirm with the respondent that the respondent was in fact the identified self. However, it was possible to have problems matching AGE with the age of the self in the roster. The interviewer was able to override the consistency check for age of the self for one of two reasons: (1) the self was misidentified and another roster member was the true self, but the interviewer insisted on not changing the entries, or (2) the interviewer correctly identified the self, but insisted that the correct age for the respondent was different than CURNTAGE, and other evidence did not support this insistence (AGE was not set to the roster age, as discussed in Section 8.2.2.1). In the case of a misidentified self, a second roster member in the household was selected whose gender matched IRSEX and whose age was within 1 year of AGE. The second roster member who replaced the original self had an age and gender that matched IRSEX and AGE, respectively.

If the consistency check was overridden, a misidentified self was diagnosed if (1) the roster age of the self differed from AGE by more than 1 year, and (2) another roster member of the same gender as QD01 (and IRSEX) had a roster age within 1 year of AGE.[104] If a misidentified self was diagnosed, it was assumed that the interviewer used the roster member identified as the self, rather than the respondent, as the point of reference. Using the example shown in Table 8.1, if the respondent's son was used as the reference point, the relationship for the respondent became "mother" instead of "self," and the "husband" became "father." Under these circumstances, the self code was set to missing, and the respondent's roster entries did not include a self. The remaining relationship codes in the roster also were set to missing. In some cases, the original relationship codes were salvaged, depending upon the roster member who was used as a reference point.

8.2.4.1 Original Self Misidentified: Identifying the Real Self

If the self was misidentified in the roster, an attempt was made to identify a self among the roster members corresponding to the respondent. A roster member was selected as the self under one of two possible circumstances: (1) the roster member's age, gender, and relationship data were missing, or (2) the roster member was of the respondent's gender and was within 1 year of the respondent in age. If more than one roster member met the above criteria, the roster members who met the criteria, but were not assigned the self code, were given a bad data code; that is, the original relationship code would no longer make sense because the reference person had been changed.

[104] A 1-year difference was allowed because the respondent's age might have changed during the interview. In this instance, the values of AGE and CURNTAGE may have differed by 1 year.

8.2.4.2 Salvaging Relationship Codes with a Misidentified Self

As stated earlier, if the self was misidentified, all other relationship codes were set to missing because the reference person was someone other than the respondent. In some cases, however, the original relationship codes were salvaged, depending upon the roster member who was used as a reference point. Relationship codes were salvaged under the following circumstances:

1. If the reference person was the respondent's sibling, the roster member listed as "self" was actually a sibling, and all other relationship codes were salvaged. (Generally, relationships between the respondent and other household members would be the same with a sibling. For example, the respondent's parents are also the respondent's sibling's parents.)
2. If the reference person was the respondent's spouse or unmarried partner, the roster member listed as "self" was actually a spouse or unmarried partner, and the children relationship codes were salvaged.
3. If all the roster members other than the misidentified self were either roommates, boarders, or other nonrelatives, then the reference person was the respondent's roommate, boarder, or other nonrelative. All other relationship codes were salvaged. This occurred once in the 2007 survey.

8.2.5 Roster Edits for Other Household Members

Relationship codes were edited if the relationship of the roster member was impossible based on age and gender in relation to the self. Edits of household roster ages, genders, and/or relationship codes were performed that either changed the reported value to another value or changed the reported value to bad data. It is important to note that in some cases, two members were selected in a household, which greatly increased the ability to edit the roster for those respondents. Some edits were associated with consistency checks. Interviewers' explanations for overrides to these consistency checks were carefully examined to assess the legitimacy of the override as explained in Section 8.2.2. Some edits were "automatic" in the programming code, which meant that the interviewer was assumed to have been incorrect when the override was implemented. These edits were undone if the interviewer's explanation for the override was considered legitimate. In other situations, the default strategy was to assume that the override of the consistency check was correct and, therefore, that the edit was applied only if the interviewer's explanation was suspicious. Interviewers' explanations for overrides to consistency checks and evaluations of their legitimacy are provided in Appendix I.

For all of the edits described below, the frequency of the application of each edit in the 2007 survey is listed. In some cases, this frequency is given for special cases within the description of the edit. The total number of applications in the 2007 survey is provided in parentheses after the description of each edit. The frequency in parentheses does not include cases where an override to a consistency check occurred and the explanation to the override was considered legitimate.

8.2.5.1 Edits to Roster Age, Gender, and Relationship Codes: Changes to Different Values (Correct Reference Person)

The following edits were performed on the roster age, gender, and relationship code values when the recorded age, gender, and/or relationship code was either missing or internally inconsistent and replaced by internally consistent values. In these cases, even though the relationship code was incorrect, the reference person for the relationship code was still the respondent.

1. When typing on a computer keyboard, it was possible for a double-digit age to have been entered as a single-digit age ("5" instead of "55"), or vice versa ("55" instead of "5"). If the relationship code still was believable even with the incorrectly entered age (e.g., "other relative"), this type of error was difficult to detect. On the other hand, if an age entered this way triggered one of the consistency checks discussed in Section 8.2.2.2, the interviewer had an opportunity to correct the entry error. On those occasions where the age did not trigger a consistency check, detection of the error was still possible among selected pairs. If two pair members were selected in the household, this error could be observed by examining the roster entries of the other pair member. If one pair member had an x-year-old and no xx-year-olds, and the other had an xx-year-old and no x-year-old, where x denoted a single-digit number, it was highly probable that an error had occurred. By comparing the number of children younger than 12 years old in each roster with the number of children on the screener roster, it was apparent how a correction should be made. In this instance, the offending age was replaced with the value given by the pair member whose roster age and screener age agreed. (2007 survey: was not applied)
2. If two members were selected in a household, the roster age for the other member selected was commonly not the same as the questionnaire-edited age (AGE, defined in Chapter 4) of the other pair member. In this case, the roster age for the other member selected was changed to this questionnaire-edited age value. (2007 survey: applied 3,010 times where age differences were only 1 or 2 years or applied 2,772 times where missing values were replaced)
3. If two members were selected in a household, the roster gender for the other member selected was often not the same as the imputation-revised gender (IRSEX, defined in Chapter 4) of the other pair member. In this case, the roster gender for the other member selected was changed to this imputation-revised gender value. (2007 survey: applied 35 times)
4. In previous survey years, the relationship codes for grandchild (9) and grandparent (11) were commonly confused. Because of the introduction of consistency checks (consistency checks #2 and #3 in Section 8.2.2.2), this did not occur in the 2007 survey. The following edit, which was used in previous survey years, was maintained in case of overrides: If the age of the respondent was at least 20 years older than that of the roster member, but the roster member was identified as a grandparent, the relationship code was changed to grandchild. Conversely, if the age of the respondent was at least 20 years younger than that of the roster member, but the roster member

was identified as a grandchild, then the relationship code was changed to grandparent. (2007 survey: was not applied)

8.2.5.2 Edits to Relationship Codes: Changes to Missing Codes

The following edits were performed on the roster relationship code values, where the relationship code given was internally inconsistent and no internally consistent value could be used to replace it. These edits were performed before the edits listed in Section 8.2.5.1 were completed. For respondents who had changes to their rosters that were due to the edits described below, changes to age and gender that were due to the edits in Section 8.2.5.1 were checked to make sure that they did not impact the decision to implement the edits below. The relationship code in these instances, as listed below, was set to a bad data code.

1. More than one roster member aged 15 years or older was listed as the respondent's unmarried partner or as the respondent's spouse. This situation should have been covered by consistency check #1 in Section 8.2.2.2. One override to this consistency check was observed in the 2007 survey. (2007 survey: was not applied)
2. A roster member aged 15 years or older was identified as a spouse and another was identified as an unmarried partner. In this case, the spouse code was maintained and the unmarried partner code was set to bad data. This situation should have been covered by consistency check #1 in Section 8.2.2.2. No overrides were observed in the 2007 survey. (2007 survey: was not applied)
3. The roster member was the respondent's parent, but was younger than the respondent. This situation should have been covered by consistency check #2 in Section 8.2.2.2. No overrides to this consistency check were observed in the 2007 survey. This edit would have been automatic for respondents younger than 15 years old. (2007 survey: was not applied)
4. The roster member was the respondent's child, but was older than the respondent. This situation should have been covered by consistency check #3 in Section 8.2.2.2. Two overrides that were considered legitimate did occur in the 2007 survey, though not with a respondent younger than 15 years old. This edit would have been automatic for respondents younger than 15. (2007 survey: applied once)
5. The roster member was the respondent's biological parent, but was less than 13 years older than the respondent. This situation should have been covered by consistency check #9 in Section 8.2.2.2. Three overrides to this consistency check occurred in the 2007 survey. (2007 survey: applied once)
6. The roster member was the respondent's biological mother, but was more than 60 years older than the respondent. (2007 survey: was not applied)
7. The roster member was the respondent's biological child, but was less than 13 years younger than the respondent. This situation should have been covered by consistency check #10 in Section 8.2.2.2. No overrides to this consistency check occurred in the 2007 survey. (2007 survey: was not applied)
8. A respondent had a biological sibling older than a biological parent, where the biological parent was at least 13 years older than the respondent. If this situation

occurred, the relationship code of the "sibling" was set to missing. If the age difference between the biological sibling and the respondent was more than 25 years, then a consistency check was triggered (consistency check #11 in Section 8.2.2.2). (2007 survey: was not applied)

9. A respondent had a biological parent younger than a biological sibling, where the biological parent was less than 13 years older than the respondent. If this situation occurred, the relationship code of the "parent" was set to missing. As with the previous edit, this edit was partially covered by consistency check #11 in Section 8.2.2.2. (2007 survey: was not applied)
10. The roster member was the respondent's child-in-law, but was at least 10 years older than the respondent. This situation should have been covered by consistency check #7 in Section 8.2.2.2. No overrides to this consistency check occurred in the 2007 survey. (2007 survey: was not applied)
11. The roster member was the respondent's parent-in-law, but was at least 10 years younger than the respondent. This situation should have been covered by consistency check #8 in Section 8.2.2.2. No overrides to this consistency check occurred in the 2007 survey. (2007 survey: was not applied)
12. The roster member was the respondent's parent-in-law or child-in-law, but either the roster member or the respondent was younger than 15 years old. This situation should have been covered by consistency check #6 in Section 8.2.2.2. Seven overrides to this consistency check occurred in the 2007 survey. (2007 survey: was not applied)
13. The respondent had two or more children-in-law, but had no children in the household. The in-law codes were all set to missing. (2007 survey: applied once)
14. The roster member was the respondent's grandchild, but the respondent or respondent's spouse (if applicable) was 25 years old or younger. This situation should have been covered by consistency check #12 in Section 8.2.2.2. No overrides to this consistency check occurred in the 2007 survey. (2007 survey: applied once)
15. The roster member was the respondent's grandchild, but the respondent's parents lived in the household. Also, the respondent had no children in the household and was less than 24 years older than the roster member. As with the previous edit, if the grandchild was in fact older than the respondent, this error should have been covered by consistency check #3 in Section 8.2.2.2. No overrides to this consistency check occurred in the 2007 survey. (2007 survey: was not applied)
16. The roster member was the respondent's sibling and the previous roster member[105] was a parent, but the roster member's age was within 4 years of the age of the parent. If the sibling was a half- or step-sibling, an additional requirement was that there was only one parent. (2007 survey: was not applied)
17. The roster member was the respondent's grandparent or grandchild, but the age difference between the respondent or the respondent's spouse (if applicable) and the

[105] A "previous roster member" is the member who immediately precedes the member of interest in the roster.

roster member was less than 20 years. If the roster member was a "grandchild" who was older than the respondent, then this situation was covered by consistency check #3 in Section 8.2.2.2. Similarly, if the roster member was a "grandparent" who was younger than the respondent, then this situation was covered by consistency check #2 in Section 8.2.2.2. If the age difference was 30 or more years, this was covered by consistency check #14. One grandparent override occurred in the 2007 survey. (2007 survey: was not applied)

18. If the respondent had two parents, but both parents were listed as biological mothers or both parents were listed as biological fathers, the roster genders of both roster members were set to missing. (2007 survey: applied seven times)

8.2.5.3 Edits to Relationship Codes: Changes to Different Values (Incorrect Reference Person: Illogical Child Code)

In Section 8.2.5.2, illogical relationship codes were set to bad data. Often, this occurred because the interviewer used someone other than the respondent as the reference person for one or more roster members. In some of these cases, the structure of the roster could have been used to determine the appropriate relationship code for that individual. Edits where the illogical code was "child" are listed below:

1. The interviewer might have put a roster member after the respondent's parent in the household roster. If the relationship code for that roster member was given as "child," the relationship code was illogical if the age made it impossible for the roster member to be the respondent's child (see #4 in Section 8.2.5.2). In fact, if more than one "child" was listed after the respondent's parent, each would be listed as illogical. However, it was likely that the interviewer was making the reference to the respondent's parent rather than the respondent. In this case, if the child relationship was not a stepchild and the age difference between the respondent's parent and the "child" was at least 12 years, then the relationship code was changed to sibling. (2007 survey: was not applied)
2. In some cases, the interviewer's entry for a roster member listed as "child" might simply be a typographical error, for example, where the "3" should be a "4" (see Table 8.2 for relationship codes). Interviewers usually corrected such errors when a consistency check was triggered in cases where the child was older than the parent or the child was a biological child who was less than 12 years younger than the parent (see Section 8.2.5.2). However, in cases where the interviewer insisted on the code, or where the child was younger than the respondent, but was less than 12 years younger than the respondent and was not biological, these typographical errors were more difficult to detect. If the respondent was living with parent(s) and unmarried and not living with an unmarried partner, and the roster member was not 12 or more years younger than the respondent, then the relationship code was changed to sibling. (2007 survey: applied 12 times)

3. Both sides in a selected pair[106] were respondents aged 18 or younger, both sides identified parents in the household, and one side had an illogical child code. When the number of illogical child codes was added to the number of siblings on one side, the sum was equal to the number of siblings on the other side. If the age of the roster member was younger than 25 years, then the relationship code was changed to sibling. (2007 survey: was not applied)
4. A roster member was listed as the respondent's child who was not more than 12 years younger than the respondent and the respondent was 25 or younger. The previous roster member was listed as "grandparent." The "child" was in reference to the respondent's grandparent and was considered either the respondent's parent or the respondent's uncle or aunt. If the roster member's age was at least 12 years older than the respondent and there were no nonimmediate family codes (7, 12, 13, or 14 as described in Table 8.2), then no uncles or aunts lived in the household. If a pair was selected and no nonimmediate family codes were found in either pair member's roster, then in either of these cases the relationship code was set to parent. Otherwise, the relationship code was set to missing. (2007 survey: was not applied)

8.2.5.4 Edits to Relationship Codes: Changes to Different Values (Incorrect Reference Person: Illogical Spouse Code)

The interviewer also could have used an incorrect reference person with illogical spouse codes. This error occurred most frequently when a selected child had a parent with a spouse (the other parent) or unmarried partner. Rather than identifying this individual as a "parent" or "other nonrelative," the interviewer identified the roster member as a spouse or unmarried partner of the child, even though the interviewer intended that the point of reference be the child's parent rather than the child. This manifestation of the illogical spouse code, along with others, is described below. It should be noted that many of these edits were covered by consistency checks #4 and #5 in Section 8.2.2.2, provided either the respondent or the roster member was 16 or younger. If any of the edits below were applied because of an override to one of these consistency checks, then it is noted in the affected edit.

1. Both sides in a selected pair identified a spouse or unmarried partner, but were not part of a spouse-spouse pair. This legitimately could have occurred only if there were multiple spouse-spouse pairs in the household. In this edit, an attempt was made to identify cases with a single spouse-spouse pair in the household, where one pair member had a correctly identified spouse or unmarried partner and the other pair member had an incorrectly identified spouse or unmarried partner. If the younger respondent, who was 21 years old or younger and at least 10 years younger than the older respondent, indicated a parent, and the older respondent indicated neither parents nor parents-in-law, then the older respondent should be considered either the younger respondent's parent or the parent's spouse or unmarried partner. If the misidentified code was "spouse," then the code was changed to "parent." However, if the misidentified code was "unmarried partner," then the roster member may or may not be considered the parent of the respondent. In most cases where the misidentified

[106] A selected pair has two rosters where each respondent is from the same household. A "side" refers to one of the two rosters that make up a selected pair.

unmarried partner was the respondent's parent's unmarried partner, the code was changed to parent. The exception occurred when (1) the unmarried partner of this respondent's parent was the other respondent selected in a pair, and (2) the unmarried partner did not indicate that the other pair member selected was his or her child in the parenting experiences question, FIPE3. In this instance, the relationship code was changed to a special code indicating that the roster member was an unmarried partner of the respondent's parent. (2007 survey: applied six times)

2. As in the previous edit, both sides in a selected pair identified a spouse or unmarried partner, but were not part of a spouse-spouse pair, and there was only a single spouse-spouse pair in the household. In this edit, both sides incorrectly identified the spouse or unmarried partner. In most cases, the pair was a sibling-sibling pair. If both respondents were younger than 21, both indicated a parent in the household, and the age difference between the respondents and their respective "spouse or unmarried partner" was unusually large, then on each side the misidentified spouse or unmarried partner should have been considered a spouse or unmarried partner of the respondent's parent. If both misidentified codes were "spouse," then both codes were changed to "parent." As stated in the previous edit, if both misidentified codes were "unmarried partner," then it was not clear whether each misidentified code should have been "parent." The rules used to determine whether the roster member was the respondent's parent were the same as in edit #1. The same special code as in the previous edit was used to identify an unmarried partner of the respondent's parent. Hence, the incorrectly identified "spouse or unmarried partner" code was changed for each respondent in the pair to either "parent" or the aforementioned special code. (2007 survey: applied four times)

3. In this edit, only one side in a selected pair identified a spouse (not unmarried partner), but the spouse was identified even though either (1) the respondent was younger than 15; (2) the spouse was younger than 15 and the other pair member did not have a spouse; or (3) the respondent was younger than 18, but responded that he or she was "never married" in the core part of the questionnaire, and the respondent did not have any parents-in-law in the household. If the respondent listed one parent, but the other pair member listed two parents, then the pair was a sibling-sibling pair and the relationship code was in reference to the parent. If the respondent listed one fewer sibling than the other pair member, then the pair was a sibling-sibling pair and the spouse code was a typographical error (meant to be a sibling, with a code "4" instead of "5"). (2007 survey: was not applied)

4. Only one side in a selected pair identified an unmarried partner, but the unmarried partner was identified even though either (1) the respondent was younger than 15 or (2) the unmarried partner was younger than 15. If the respondent listed one parent, but the other pair member listed two parents, then the pair was a sibling-sibling pair and the relationship code was in reference to the parent's unmarried partner. In this case, the relationship code was changed to parent. If the respondent listed one fewer sibling than the other pair member and the age difference between the respondent and the roster member identified as the unmarried partner was less than 15 years, then the pair was a sibling-sibling pair and the unmarried partner code was changed to sibling.

No overrides to this consistency check occurred in the 2007 survey. (2007 survey: was not applied)

5. Both sides in a pair identified the same household member as spouse or unmarried partner. If the previous roster member on one of the sides was a sibling, then the spouse or unmarried partner should be considered the sibling's spouse or unmarried partner. The spouse or unmarried partner relationship code was changed to bad data. If both sides had a previous roster member who was a sibling, then it was not clear to which pair member the spouse or unmarried partner belonged. To maintain proper counts, the spouse or unmarried partner code for the youngest pair member was changed. (2007 survey: was not applied)
6. A spouse or unmarried partner was identified even though (1) the respondent had one parent in the household who was the roster member listed before the spouse or unmarried partner; (2) the respondent either was younger than 17 years old or was between 17 and 20 years old and the spouse or unmarried partner was older than the respondent's parent; and (3) the respondent was more than 15 years younger than the spouse or unmarried partner. In the case of the misidentified spouse, the "spouse" of the respondent was considered the respondent's other parent. In the case of the misidentified unmarried partner, the "partner" of the respondent was considered the unmarried partner of the respondent's parent. The code was changed to "parent." For a household member with a spouse code who was aged 16 years or younger, this edit should have been covered by consistency check #4 in Section 8.2.2.2. No overrides to this consistency check occurred in the 2007 survey. (2007 survey: was not applied)
7. In cases where the respondent was younger than 15 years old, he or she identified a spouse or unmarried partner, and the above edits did not apply, the relationship code was set to bad data. In cases where the roster member was younger than 15, the roster member was identified as a spouse or unmarried partner, and the above edits did not apply, the relationship code and roster member's age were set to bad data. This should have been covered by consistency checks #4 and #5 in Section 8.2.2.2. No overrides to these consistency checks were observed in the 2007 NSDUH data that were not already handled by other edits in this section. (2007 survey: was not applied)

8.2.5.5 Edits to Relationship Codes: Changes to Different Values (Incorrect Reference Person: Illogical Sibling Codes)

If the relationship code was identified as the respondent's sibling, but the age difference between the roster member and the respondent was at least 20 years, then the sibling relationship code was suspicious. If the previous roster entry was either the respondent's child or another sibling with the same characteristics, and either the respondent did not have parents in the household or the parent was a mother and the age difference between the mother and the sibling was more than 50 years, then the sibling relationship codes were referencing the respondent's children's relationships to each other. The relationship codes were therefore changed to "child." Age differences greater than 25 years among biological siblings would have been covered by consistency check #11 in Section 8.2.2.2. Three overrides to this consistency check were observed in the 2007 survey. The cases were checked individually, with particular scrutiny placed on age differences between 20 and 25 years. (2007 survey: applied four times)

8.2.5.6 Edits to Relationship Codes: Changes to Different Values (Incorrect Reference Person: Illogical Grandchild Codes)

If the relationship code was identified as the respondent's grandchild, but the respondent was too young to have a grandchild (25 or younger), it was possible that the roster member was a grandchild of a previous roster member. If two young respondents were selected where both identified the same grandparents and the same parents, and the respondent on the other side had siblings, then the grandchild should be considered the respondent's sibling. If this was not established, then the roster member could be the respondent's sibling or the respondent's cousin and the code was set to bad data. If the grandchild was older than the respondent, then this edit would have been covered by consistency check #3. If the age difference between the grandchild and the respondent was less than 30 years, then this edit would have been covered by consistency check #14 in Section 8.2.2.2. (2007 survey: was not applied)

8.2.5.7 Edits to Relationship Codes: Changes to Different Values (Incorrect Reference Person: Illogical In-Law Codes)

An incorrect reference code also occurred with in-laws. Either the child-in-law was the child of someone else in the roster other than the respondent or the respondent was referring to himself or herself as the parent-in-law of the roster member. An in-law code was deemed incorrect if a roster member was listed as the respondent's child-in-law who was not more than 12 years younger than the respondent and the respondent was 25 or younger. If the relationship code was listed as child-in-law, and the previous roster member was listed as grandparent, then the child-in-law was in reference to the respondent's grandparent and should have been considered either the respondent's parent or the respondent's uncle or aunt. If the roster member's age was at least 12 years older than the respondent and there were no nonimmediate family codes (7, 12, 13, or 14 as described in Table 8.2), then no uncles or aunts lived in the household. If a pair was selected, no nonimmediate family codes were found in either pair member's roster. In either of these cases, the relationship code was set to parent. Otherwise, no certainty was associated with the relationship code, and this code was set to missing. (2007 survey: was not applied)

8.3 Creation of Respondent-Level Detailed Roster Variables

The raw roster variables contained information for each roster member: age, gender, relationship to respondent, and a 0/1 variable that indicated whether the roster member was the other member selected in a pair. Each of these attributes had a multiple of 25 variables corresponding to the maximum of 25 members of a household. Separate variables were created for male and female household members and for household members with ages reported in years as opposed to months. When the edited versions of these variables were created, this information was brought together into four sets of variables, one set for each attribute. The edits listed in Section 8.2 were incorporated into the values of the detailed roster variables, called ROSAGE1-ROSAGE25 (roster age), ROSSEX1-ROSSEX25 (roster gender), ROSRLT1-ROSRLT25 (relationship to respondent), ROSMSL1-ROSMSL25 (0/1 indicator: other member selected, pair members only), PRNTYP1-PRNTYP25 (type of parent: biological, adoptive, etc.), SIBTYP1-SIBTYP25 (type of sibling: biological, adoptive, etc.), CHDTYP1-CHDTYP25 (type of child:

biological, adoptive, etc.), and TWNTYP1-TWNTYP25 (type of twin: identical, fraternal, or neither).

8.4 Creation of Household Roster-Derived Variables

After replacing faulty information in the roster with missing values, the number of individuals with various characteristics in each roster was determined. These counts were recorded in the household roster-derived variables shown in Table 8.3. If any information in the roster was missing, the roster-derived variable was set to missing. However, if some of the roster records for a respondent's household had missing data, then roster records with nonmissing data for that household were used to limit the possible values to which the missing roster-derived variable could have been imputed. Details on the imputation of the household roster-derived variables are provided in Section 8.5. If two respondents were selected in a single household as part of a pair, then the information from one pair member was not used to edit that of the other pair member. This was because the interviews for each pair member could have occurred at different times, resulting in possible differences in the household composition.

The respondent's household size was assumed to equal the total number of rostered persons in the household, TOTPEOP, as shown in Table 8.3. The value of TOTPEOP was expected to equal the value of QD54 in most cases. However, in some cases, the original self was misidentified and no other roster members were close to matching the respondent's age and gender. In these cases, an extra roster member was added to correspond to the respondent (the self) so that the value of TOTPEOP was 1 greater than the value of QD54. For other cases, the respondent did not enter a value for QD54, and thus TOTPEOP and all the roster-derived variables were missing. Finally, it was possible that duplicate entries were put into the household roster so that the value of TOTPEOP would be determined by excluding the duplicates from the roster. This latter situation was usually impossible to detect, unless the respondent had two biological fathers or two biological mothers of exactly the same age. In this instance, the extra biological parent of the same gender was dropped from the roster, and the value of TOTPEOP was reduced to 1 less than the value of QD54.

Table 8.3 Household Roster-Derived Variables

Variable Description	Variable Name
Total number of rostered persons	TOTPEOP
Number of persons in household aged 17 or younger	KID17
Number of persons in household aged 65 or older	HH65
Indicator of whether the respondent had family members in household	FAMSKIP
Number of respondent's family members in household (includes foster relationships)	FMLYSIZE
Number of respondent's family members in household aged 17 or younger (includes foster relationships)	KIDFMLY
Number of respondent's family members in household (excludes foster relationships)	FAMSIZE
Number of respondent's family members in household aged 17 or younger (excludes foster relationships)	KIDFAMSZ
Number of respondent's children in household aged 2 or younger	NRBABIES
Number of respondent's children in household aged 3 to 5 years old	NRPRESCH
Number of respondent's children in household aged 6 to 11 years old	NRYUNGCH
Number of respondent's children in household aged 12 to 17 years old	NRTEENS
Number of respondent's children in household aged 17 or younger	NRCH0_17
Number of respondent's children in household aged 18 to 20 years old	NROLDRCH
Number of respondent's children in household aged 21 or older	NROLDCH
Number of roommates/housemates in household	NROOMATE
Indicator of presence of mother in household (12- to 17-year-olds)[1]	IMOTHER
Indicator of presence of father in household (12- to 17-year-olds)[1]	IFATHER
Indicator of presence of foster child in household (12- to 14-year-olds)[2]	FSTRCHLD

[1] The IMOTHER and IFATHER indicators were not 0/1 indicators because levels were provided for "unknown" and "18 or older."

[2] This variable was required for the creation of a poverty variable for the 2003-2005 survey years.

The variables KID17 (number of persons in the household aged 17 or younger) and HH65 (number of persons in the household aged 65 or older) were simple counts based on the roster ages and did not account for the relationships of the individuals to the respondent. If some of the roster members had missing ages, the values of KID17 and HH65 also were missing, regardless of whether some of the roster members were eligible to be part of the count. In these instances, the imputed values for KID17 and HH65 were restricted based on the nonmissing information available in the roster, as explained in Section 8.5.6. However, if the roster member was missing a relationship code, but not an age, then that roster member was still eligible to be counted in these variables.

The variable FAMSKIP was an indicator of whether the respondent's household contained other family members. It was created based on the relationship codes of the roster

members. If one or more of the roster members had a missing relationship code, and no other family members were in the respondent's household, then the value of FAMSKIP was set to missing. However, if one of the nonmissing roster member's relationship codes indicated that the household contained one of the respondent's family members, then the value of FAMSKIP was not missing, even if other roster members had missing relationship codes.

The variables FMLYSIZE (number of respondent's family members in the household, including foster relationships), FAMSIZE (number of respondent's family members in the household, excluding foster relationships), KIDFMLY (number of respondent's family members in the household aged 17 or younger, including foster relationships), and KIDFAMSZ (number of respondent's family members in the household aged 17 or younger, excluding foster relationships) were simple counts based on the relationships of the individuals to the respondent and the ages in the respondent's household roster. FMLYSIZE and KIDFMLY were created to determine appropriate measures of poverty levels, using Federal poverty definitions starting in 2006. FAMSIZE and KIDFAMSZ were used in the 2003 to 2005 surveys. The definition of "family" for FAMSIZE and KIDFAMSZ was a little different from that used for other roster variables; foster relationships were not considered family relationships. If some of the roster members had missing ages or missing relationship codes, the values of FMLYSIZE, FAMSIZE, KIDFMLY, and KIDFAMSZ were set to missing, even though some of the roster members might have been eligible to be part of the count. In these instances, the imputed values were restricted based on the nonmissing information available in the roster, as explained in Section 8.5.6.

Eleven other roster-derived variables were created that used both the age and relationship codes of the roster members. All of the roster-derived variables and their definitions are summarized in Table 8.3. Each of these variables was missing if the age or relationship codes for at least one roster member in a respondent's household were missing.

8.5 Imputation of Household Roster-Derived Variables

Although 19 roster-derived variables were created from the edited roster, missing values were imputed for only 8 of these variables: TOTPEOP, KID17, HH65, FAMSKIP, FMLYSIZE, KIDFMLY, FAMSIZE, and KIDFAMSZ. The missing values in these variables were imputed using the univariate predictive mean neighborhood (UPMN) technique, as described in Appendix C.

8.5.1 Hierarchy of Household Roster-Derived Variables

After editing the roster variables, the next step in the imputation of household roster-derived variables was to determine the order in which the variables should be modeled. Each roster-derived variable was expected to have a high association with the other seven roster-derived variables. Hence, it was important to perform the imputations sequentially so that variables early in the series were used as covariates for subsequent variables, if needed. The order in which the roster variables were imputed is shown in Table 8.4.

Table 8.4 Household Roster-Derived Variables (in Order of Imputation)

Roster Variable	Edited Variable	Imputed Variable
Total number of rostered persons	TOTPEOP	IRHHSIZE
Total number of persons aged 17 or younger	KID17	IRKID17
Total number of persons aged 65 or older	HH65	IRHH65
Indicator of whether the respondent has family members in household	FAMSKIP[1]	IRFAMSKP
Total number of respondent's family members in household (includes foster relationships)	FMLYSIZE	IRFMLYSZ
Total number of respondent's family members in household aged 17 or younger (includes foster relationships)	KIDFMLY	IRKDFMLY
Total number of respondent's family members in household (excludes foster relationships)	FAMSIZE	IRFAMSZE
Total number of respondent's family members in household aged 17 or younger (excludes foster relationships)	KIDFAMSZ	IRKIDFAM

[1] FAMSKIP was set to 0 if the roster had relationship codes of 2, 3, 4, 5, 6, 8, 9, 10, 11, and 13 as described in Table 8.2. FAMSKIP was set to 1 if no relationship codes were missing and the roster had codes of 1, 7, 12, and/or 14 as described in Table 8.2.

8.5.2 Setup for Model Building

Once the hierarchy of the roster-derived variables was established, the next step was to define respondents, nonrespondents, and the item response mechanism. Imputations for all roster-derived variables were conducted separately within the four age groups: 12 to 17, 18 to 25, 26 to 64, and 65 or older. Response propensity adjustments were then computed for each age group to make the item respondent weights representative of the entire sample. (Because the modeling of the final weight adjustments was not completed at the time of the roster imputations, the person-level sample design weights were adjusted to account for nonresponse at the household level using a simple ratio adjustment.[107]) Item respondents were not defined across all roster categories. Hence, this adjustment was computed separately for each age group and for each variable. The covariates in the response propensity models were the same covariates as those used in the main model presented in the next section. The item response propensity model is a special case of the generalized exponential model (GEM).[108] Details of the GEM software are presented in Appendix B.

8.5.3 Sequential Model Building

The variables TOTPEOP, KID17, HH65, FMLYSIZE, KIDFMLY, FAMSIZE, and KIDFAMSZ were assumed to have a Poisson distribution, and the parameters for the models

[107] In subsequent text, the use of the word "weights" will refer to the ratio-adjusted design weights.

[108] The GEM macro, which was written in SAS/IML® software, was developed at RTI International (a trade name of Research Triangle Institute) for weighting procedures.

were estimated using the LOGLINK procedure in SUDAAN® software.[109] The binary variable FAMSKIP was modeled using weighted logistic regression. The covariates in each model were continuous centered age,[110] continuous centered age squared, gender, race/ethnicity, imputation-revised roster-derived variables earlier in the sequence, region, population density, percentage Hispanic/Latino households in segment, percentage of owner-occupied households in segment, and (for TOTPEOP only) number of persons in the household eligible for interviewing (from the pre-interview screener). There were also predictors that consisted of one-way interactions of centered age with race/ethnicity, centered age with gender, race/ethnicity with gender, centered age squared with race/ethnicity, and centered age squared with gender. For the three older age groups, the additional covariates of marital status, education status, and employment status also were included.

8.5.4 Computation of Predicted Means and Univariate Predictive Mean Neighborhoods

From the final models, a predicted mean was computed for every respondent. The assignment of imputed values for the roster-derived variables was conducted using the UPMN technique.

8.5.5 Assignment of Imputed Values

Separate assignments were performed within each of the four age groups. A univariate imputation was implemented for each of the roster-derived variables within each age group, using the predicted means from the appropriate models. Assignments were made within preset bounds, as discussed in the next section. If no imputed values were available within the preset bounds, a random imputation was performed within those bounds.

8.5.6 Constraints on Univariate Predictive Mean Neighborhoods

A univariate imputation was implemented on each variable within each age group after predicted means from the models had been determined. In a general UPMN imputation, the neighborhood is restricted by two types of constraints: (1) logical constraints (which cannot be loosened) to make imputed values consistent with a nonrespondent's preexisting nonmissing values of other variables, and (2) likeness constraints (which can be loosened) to make candidate donors in the neighborhood as similar to recipients as possible.

The logical constraints on the neighborhoods were sequentially based on the information already available in the roster and on roster-derived variables already imputed. The assignment of imputed values for KID17 was restricted within a lower and upper bound based on the value of IRHHSIZE and the nonmissing ages in the roster. For example, if a household roster had four members consisting of two members aged 18 or older, one member who was missing age, and another member aged 17 or younger, then KID17 would be missing. Thus, at least one child aged

[109] SAS®-callable SUDAAN® was used to fit all binary logistic regression models. Details about the LOGLINK procedure are discussed and additional references are provided in the *SUDAAN Language Manual Addendum, Release 9.0.3* (RTI International, 2007). SAS software is a registered trademark of SAS Institute, Inc. SUDAAN is a registered trademark of RTI International.

[110] The covariate age was centered within each age group to reduce the effects of multicollinearity, particularly with the squared age terms. For more information on "centering" and "multicollinearity," refer to Draper and Smith (1981).

17 or younger would be in the household and two adults would be in the household. Hence, the assignment of KID17 in this example would be restricted between the values of 1 and 2. Likewise, HH65 was restricted within bounds in the same manner, using the variables IRHHSIZE and IRKID17 and the nonmissing ages in the roster. FMLYSIZE also was restricted within bounds based on IRHHSIZE and the nonmissing ages in the roster. KIDFMLY was restricted within bounds using the variables IRHHSIZE, IRFMLYSZ, and IRKID17 and the nonmissing ages in the roster. FAMSIZE was restricted within bounds using the variables IRHHSIZE and IRFMLYSZ and the nonmissing ages in the roster. KIDFAMSZ was restricted within bounds using the variables IRHHSIZE, IRFMLYSZ, IRKDFMLY, IRFAMSZE, and IRKID17 and the nonmissing ages in the roster.

Likeness constraints also were applied to the imputation of missing values in TOTPEOP, KID17, HH65, FAMSKIP, FMLYSIZE, KIDFMLY, FAMSIZE, and KIDFAMSZ. The delta constraint (described in Appendix A) could be considered a likeness constraint and could be loosened by enlarging delta or abandoning the neighborhood altogether and taking the donor with the closest predicted mean. For TOTPEOP, delta was the only likeness constraint. If possible, donors and recipients for KID17 and HH65 were required to have the same household size (IRHHSIZE, the imputation-revised version of the household size variable). FAMSKIP donors and recipients were required to have the same values for IRKID17 (the imputation-revised version of KID17) and marital status. For FMLYSIZE, donors and recipients were required to have the same values for IRHHSIZE and IRKID17. For FAMSIZE and KIDFMLY, donors and recipients were required to have the same values for IRHHSIZE, IRKID17, and IRFMLYSZ (the imputation-revised version of FMLYSIZE). Also, KIDFAMSZ donors and recipients were required to have the same values for IRHHSIZE, IRKID17, IRFMLYSZ, and IRKDFMLY (the imputation-revised version of KIDFMLY). For KID17 and HH65, the household size likeness constraint was loosened after abandoning the neighborhood. For FAMSKIP, the marital status likeness constraint was never loosened, even after enlarging the neighborhood. For FMLYSIZE, KIDFMLY, FAMSIZE, and KIDFAMSZ, the neighborhood was abandoned before any other likeness constraints were loosened. The likeness constraints and the number of recipients with sufficient donors corresponding to each likeness constraint are summarized in Appendix G.

8.6 Proxy Variables

8.6.1 Introduction

The proxy portion of the questionnaire allowed the interviewer to determine whether there was another person in the household who was better suited than the respondent to answer the questions about health insurance coverage and income. As in previous survey years, for respondents in households with two or more members, respondents were asked to provide a roster of all persons living in the household (including the respondent) and the relationship of the respondent to the other household members. If the household contained at least one adult related to the respondent, the respondent was asked questions to determine whether this other person (or one of the other persons) might be a more suitable proxy. The questions concerned with proxy information in the 2007 survey were the same as those asked since the 2003 survey, but were slightly different from those asked in the 1999-2002 surveys. For all surveys since the 1999 NSDUH, whether or not a proxy could be selected was based on whether family members aged

18 or older were in the household roster. However, since the 2003 survey, the respondent was asked to choose a suitable proxy from a list of eligible family members based on the respondent's household roster. In the surveys prior to 2003, the respondents were allowed simply to provide the relationship of their proxy regardless of their answers in the household roster.

8.6.2 Editing of Proxy Variables

All survey respondents were allowed to choose someone to be their proxy as long as the following conditions were met:

1. There was more than one person in the household.
2. The eligible person was a relative (not a boarder, roommate, or some other nonrelative).
3. The eligible person was aged 18 or older.

Table 8.5 shows the correspondence between the five questionnaire items in the proxy section of the questionnaire and the corresponding edited variables. Except for QP02 and its edited variable PRXRELAT, the valid questionnaire responses were "1 = Yes" and "2 = No." QP02 and PRXRELAT had multiple responses ranging from 1 to 21 with each level representing the relationship of the proxy to the respondent.

Table 8.5 Mapping of Raw Proxy Information Variables to Edited Variables

Raw Variable	Text of Survey Question Associated with Raw Variable	Edited Variable
QP01	Is there anyone else who lives here who is 18 or older who would be better able to give me the correct information about your health insurance coverage and the kinds of income you receive?	PRXABLE2
QP02	Who is the person you think can help us get the correct information for these questions?	PRXRELAT
QP03	Is your [QP02 fill] available right now?	PRXHOME2
QP04	Would you ask your [QP02 fill] to join us to help with these last questions about health insurance and income?	PRXJOIN2
HASJOIN	Has the person's [QP02 fill] joined R?	PRXYANS2

8.6.2.1 Edited Indicator of Potential Proxies in Household (EDFAM18)

As described in Section 8.4, a binary variable (FAMSKIP) was created that indicated whether the respondent's household roster included other family members. If the presence or absence of other family members was ambiguous because of a missing household size or missing values in the roster, FAMSKIP could not be determined. As described in Section 8.5, missing values in FAMSKIP were imputed in the variable IRFAMSKP. A similar variable was created to identify households where the respondent's household roster included other family members aged 18 years or older ("adult" family members), any one of whom could potentially serve as a proxy for the respondent. The edited indicator was called EDFAM18, where "1" indicated that no potential proxy existed in the respondent's household and "0" indicated otherwise.

8.6.2.2 Editing of Proxy Variables when EDFAM18 = 1

In most cases, a value of EDFAM18 = 1 implied that the respondent was skipped out of the proxy questions because no potential proxy existed in the household. In these cases, all of the proxy variables were given a legitimate skip code (99). Two situations could occur, however, where adult family members were incorrectly identified in the household roster by the computer. In these cases, the respondent was allowed to answer the proxy questions even though the value of EDFAM18 was 1 (i.e., the final edited household roster indicated that no potential proxy existed in his or her household). The two situations were (1) the respondent had not identified any adult family members in the household, but had nonfamily members in the household whose ages were not known; and (2) the unedited household roster indicated that one potential proxy existed in the household, but editing changed the age of this single potential proxy to younger than 18. In these situations, the interviewer indicated that none of these household members who were incorrectly identified as adult family members were proxies. However, the "no" value in the first raw proxy variable (QP01) was replaced by a logically assigned legitimate skip (89) in the corresponding edited variable (PRXABLE2). For cases where PRXABLE2 was set to 89, all of the edited proxy variables corresponding to the raw proxy variables, which followed QP01, were given legitimate skip codes (99).

8.6.2.3 Editing of Proxy Variables when EDFAM18 = 0

If EDFAM18 was 0, the proxy variables were edited as follows:

1. If the raw proxy variables had legitimate nonmissing values (i.e., not replaced by a logically assigned legitimate skip), the edited proxy variables (except PRXRELAT) were set to those nonmissing values.
2. If any of the raw proxy variables (except PRXRELAT) had a value of 2 ("no"), then all of the variables that followed were edited to legitimate skips.
3. If any of the raw proxy variables had a value of "don't know" or "refused," then the corresponding edited variable and all the edited variables that followed were given a "don't know" or "refused" code (94 or 97).
4. If any of the raw proxy variables did not have a value and a legitimate skip code could not be applied, then the corresponding edited variable and all the variables that followed were given a "no answer" code (98).

In addition to the above edits, more detailed rules were used to assign values to PRXRELAT, the edited variable corresponding to QP02. The value of QP02, which identified the proxy for the respondent, was chosen directly from the respondent's household roster. To assign a code for QP02, a subset of the respondent's roster (called a proxy roster) was created that included only adult family members. In the cases where the proxy roster included a large number, only the first nine adult family members listed in this roster were allowed for selection. Once the proxy roster was established, the number selected in QP02 was matched to the corresponding person in the proxy roster. The definitions of the levels of PRXRELAT are shown in Table 8.6.

Table 8.6 Assignment of Values for PRXRELAT, Based on Proxy Member Relationship

PRXRELAT	Relationship of Proxy Member	Gender of Proxy Member
1 = Father	Parent	Male
2 = Mother	Parent	Female
3 = Son	Child	Male
4 = Daughter	Child	Female
5 = Brother	Sibling	Male
6 = Sister	Sibling	Female
7 = Husband	Spouse	Male
8 = Wife	Spouse	Female
9 = Male Unmarried Partner	Unmarried partner	Male
10 = Female Unmarried Partner	Unmarried partner	Female
11 = Son-in-law	Child-in-law	Male
12 = Daughter-in-law	Child-in-law	Female
13 = Grandson	Grandchild	Male
14 = Granddaughter	Grandchild	Female
15 = Father-in-law	Parent-in-law	Male
16 = Mother-in-law	Parent-in-law	Female
17 = Grandfather	Grandparent	Male
18 = Grandmother	Grandparent	Female
19 = Other Male Relative	Other relative	Male
20 = Other Female Relative	Other relative	Female

9. Income

9.1 Introduction

As with most of the imputation-revised variables discussed in the previous chapters of this report, imputations for the 2007 National Survey on Drug Use and Health (NSDUH)[111] were accomplished using the predictive mean neighborhood (PMN) technique, as described in Appendix C. The edits applied to the income variables are described in Kroutil and Chien (2008).

The imputation of income was separated into two phases. The first phase was known as the "binary variable phase" and involved the imputation of all the binary income variables, as well as the number of months on welfare. This included the "yes-no" questions about the following sources of income: Social Security, Supplemental Security Income, welfare cash assistance, welfare noncash assistance, wages, food stamps, child support, interest/investment income, and other income; the number of months in which welfare was received (the only nonbinary variable in the binary variable phase); and a "yes-no" question regarding whether the respondent's income or the respondent's family income (in the household) was $20,000 or more (including income from the sources referenced in the previous questions). If a household contained other family members, then separate questions were asked to ascertain personal-level responses and other-family-level responses. These responses were then combined to create family-level responses. The second phase of the imputation of income was known as the "finer category phase" and consisted of imputing more specific income categories for the respondent and the respondent's family in the household.

In the 2007 NSDUH, 3,262 (4.8 percent) of the sample of 67,870 respondents received a new reduced set of income questions to reduce the burden on respondents. This group of respondents was categorized as belonging to Sample B, while the group of remaining respondents who answered the original set of income questions was categorized as belonging to Sample A.

In the new set of income questions in the 2007 survey, questions covering child support, interest/investment income, and other income were omitted. In addition, separate questions to ascertain personal-level and other-family-level responses were no longer asked; all questions were asked at the family level only.

Respondents in both Sample A and Sample B were asked questions about binary and finer category actual annual income at personal and family levels, in the same way as in previous study years. A comparison between the original and reduced set of income questions in terms of questions asked is shown in Table 9.1.

[111] This report presents information from the 2007 National Survey on Drug Use and Health (NSDUH), an annual survey of the civilian, noninstitutionalized population of the United States aged 12 or older. Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA).

Table 9.1 Comparison between Original and Reduced Set of Income Questions

Income Questions in 2007 NSDUH	Original Set			Reduced Set		
	Personal Level	Other Family Member Level	Family Level	Personal Level	Other Family Member Level	Family Level
Social Security	x	x	—	—	—	x
Supplemental Security Income	x	x	—	—	—	x
Welfare Payments	x	x	—	—	—	x
Other Welfare Services	x	x	—	—	—	x
Investment Income	x	x	—	—	—	—
Child Support Payments	x	x	—	—	—	—
Wages	x	x	—	—	—	x
Other Income	x	x	—	—	—	—
Food Stamps	x	x	—	—	—	x
Months on Welfare	x	x	—	—	—	x
Binary Total Income	x	—	x	x	—	x
Finer Category Total Income	x	—	x	x	—	x

NOTE: The "x" symbol indicates that this question was asked and the "—" symbol indicates that this question was not asked in the 2007 NSDUH.

During each quarter, approximately 5 percent of 2007 NSDUH respondents were randomly selected to receive the new income questions (Sample B). This module was randomly assigned at the dwelling unit level. If two persons were selected in the household, both persons received the same income module treatment. A total of 3,262 new income module interviews were completed for the year.

9.2 Binary Variable Phase

9.2.1 Order of Modeling Income Variables

The first step in the imputation of income variables was to determine the order in which the variables would be modeled. A motivation for using a hierarchy in PMN is provided in Appendix C for drug use variables. For a model predicting whether a respondent had a given

source of income, other sources of income were useful covariates. Following a provisional imputation of missing income values in the binary variable phase, the indicators earlier in the sequence were used as covariates for income models later in the sequence. Any imputed values in the income variables were considered temporary at this stage. This was because the final imputation was not implemented for income indicators until the modeling was completed for all income variables in the binary variable phase. The order in which the income indicators were imputed is listed in Table 9.2.

Table 9.2 Order of Imputation of Income Variables in Binary Variable Phase and Edited Family Income Response Variables Used in Predictive Mean Models

Income Type	Variable Name
Family Social Security	FAMSOC
Family Supplemental Security Income	FAMSSI
Family Welfare Payments	FAMPMT
Family Other Welfare Services	FAMSVC
Family Investment Income	FAMINT
Family Child Support Payments	FAMCHD
Family Wages	FAMWAG
Family Other Income	FAMOTH
Family Food Stamps	FSTAMP
Family Months on Welfare	WELMOS
Total Family Income[1]	FINC1

[1] The model for total family income used all of the variables above as covariates except the variable indicating months on welfare.

9.2.2 Setup for Model Building

Once the hierarchy of income variables in the binary variable phase was established, the next step was to define respondents, nonrespondents, and the item response mechanism. Because of the changes in the 2007 NSDUH, those variables that were not in the reduced set of income questions were assigned a legitimate skip code in Sample B. The imputation-revised variables also were set to a legitimate skip code and delivered to the analytic file. However, internal variables were created with the legitimate skip codes replaced by imputed values; the internal variables were only used as covariates in models for subsequent processing. That is, prior to the modeling step, the data from Sample B were set up in the same way as for Sample A, with the exception of missing values for those internal variables that were not in the reduced set of income questions. A methodological study (Aldworth, Liu, & Copello, 2007) showed that the donor pool would still be sufficiently large to accommodate the substantial increase in cases requiring imputation. Imputations for all income indicators were conducted separately within the four age groups of respondents: 12 to 17, 18 to 25, 26 to 64, and 65 or older. For an individual to be considered an item respondent for income variables in the binary variable phase, he or she must have complete data for all of the questions included in this phase. These questions consist of Social Security, Supplemental Security Income, welfare payments and services, investments, child support, wages, other sources of income, food stamps, months on welfare, and total family

income (less than $20,000 vs. $20,000 or more). Respondents in Sample B were automatically considered as item nonrespondents because their internal values for those variables that were not in the reduced set of income variables were set to missing. Response propensity adjustments were then computed for each age group to make the item respondent weights representative of the entire sample. (As with health insurance, the final analysis weights were used as weights. See Chapter 10 for further discussion.) Because item respondents were defined across all the income variables in the binary variable phase, this adjustment was computed only once per age group and then used in the modeling of income indicators. The item response propensity model is a special case of the generalized exponential model (GEM), which is described in Appendix B. The covariates in the item response propensity model were the same as those included in the imputation model and are discussed in the next section.

9.2.3 Sequential Model Building

Beginning with Social Security, the probability that a family received income from a given source was modeled for item respondents within each age group using the nonresponse-adjusted weights. For the models, the parameters were estimated using logistic regression.[112] The response variable for each model was the edited combination of the pair of questionnaire variables associated with each income topic in the binary variable phase, the names for which are provided in Table 9.2. The covariates in each model were centered continuous age,[113] centered age squared, gender, race, provisional income indicators imputed earlier in the sequence, region, population density, percentage Hispanic/Latino households in the segment,[114] percentage non-Hispanic/Latino black/African-American households in the segment, percentage owner-occupied households, imputation-revised number of adults in household, imputation-revised number of children in household, imputation-revised number of adults aged 65 years or older in household, and a three-level State-rank variable. There were also predictors that consisted of one-way interactions of centered age with race, centered age with gender, race with gender, centered age squared with race, and centered age squared with gender. For the three older age groups, the additional covariates of marital status, education status, and employment status were used. For the State-rank groups, definitions were determined in terms of the proportion of a given State's residents having an income of $20,000 or more. The same covariates were used for both the months-on-welfare variable and the binary total family income variable. For the months-on-welfare variable, weighted least squares regression was used, where the dependent variable was a standard logit,[115] such that $Y = \text{logit}(p)$ and p = number of months on welfare divided by 12. The binary total family income variable was modeled using weighted logistic regression. For a complete summary of the income imputation models, see Appendix F.

---

[112] In the 2007 NSDUH, the logistic regression models were run in SAS®-callable SUDAAN® rather than SAS. Both SAS and SUDAAN yield the same predicted means given the same set of covariates, but because SUDAAN acknowledges the survey design, it gives correct values for the standard errors associated with each parameter estimate. Details about the logistic regression model and additional references can be found in the *SUDAAN Language Manual Addendum, Release 9.0.3* (RTI International, 2007). SAS software is a registered trademark of SAS Institute, Inc.; SUDAAN is a registered trademark of Research Triangle Institute.

[113] The covariate age was centered within each age group to reduce the effects of multicollinearity, particularly with the squared and cubed age terms. For more information on "centering" and "multicollinearity," refer to Draper and Smith (1981).

[114] Segments were the first-stage sample units in the multistage 2007 NSDUH sample. Each segment consisted of a set of U.S. Census Bureau blocks.

[115] The Cox empirical logit was used when a person was on welfare for all 12 months.

9.2.4 Computation of Predicted Means and Univariate Predictive Mean Neighborhoods

Following the modeling of each income variable in the binary variable phase, missing values were replaced by provisional imputed values. This was necessary so that these variables could be used as covariates in subsequent models. Although no provisional imputed values were used to build the models, it was necessary to calculate predicted means for all respondents, including item nonrespondents, using the parameter estimates from the models. This sometimes required the use of the provisional values for the covariates. The predicted probabilities from these models were used to assign provisional values using the univariate predictive mean neighborhood (UPMN) imputation method described in Appendix C.

9.2.5 Assignment of Provisional Imputed Values

Separate assignments of provisional values were performed within each of the four age groups (12 to 17, 18 to 25, 26 to 64, 65 or older) for all income variables. The final income imputations were multivariate across all the variables in the binary variable phase. These variables represented source of income, months on welfare, and total income. The multivariate imputation process is further described in Section 9.2.8.

9.2.6 Constraints on Univariate Predictive Mean Neighborhoods

After predictive mean values from the model had been determined, a univariate imputation was implemented on each variable within each age group. In general, the PMN is restricted by two types of constraints: (1) logical constraints (which cannot be loosened) to make imputed values consistent with a nonrespondent's preexisting nonmissing values of other variables, and (2) likeness constraints (which can be loosened) to make candidate donors in the neighborhood as similar to recipients as possible. As a logical constraint in the binary income variable imputations, donors were required to have the same value for the family skip variable (IRFAMSKP) as the recipient. The neighborhoods for the binary income indicators were restricted so that candidate donors and recipients would be within the same age group (12 to 17, 18 to 25, 26 to 64, 65 or older). Models were built separately within these four groups, so this likeness constraint was never loosened. A small delta also could be considered a likeness constraint, and it could be loosened by enlarging delta or abandoning the neighborhood altogether and taking the donor with the closest predicted mean. More details about delta are described in Appendix C. This was the only likeness constraint that could be loosened with the binary income provisional imputations.

9.2.7 Multivariate Assignments

The predicted means were calculated with edited family income variables (see Table 9.1) as the response variables. For each variable, neighborhoods were created using scalar predicted means from the appropriate model. With respect to these scalar predicted means, a univariate methodology was used to determine the neighborhood. In most cases, three edited variables were associated with each predicted mean so that missing values for these three variables required assignment of imputed values. Hence, even when determining the provisional imputed values using the univariate procedure, the assignment of imputed values was multivariate for all binary

phase variables with two exceptions: food stamps and months on welfare. Table 9.3 shows the variables associated with each of the income models.

Table 9.3 Imputation-Revised Personal and Family Income Variables

Income Model	Variables
Social Security	IRPSOC, IROFMSOC, IRFAMSOC
Supplemental Security Income	IRPSSI, IROFMSSI, IRFAMSSI
Welfare Payments	IRPPMT, IROFMPMT, IRFAMPMT
Welfare Services	IRPSVC, IROFMSVC, IRFAMSVC
Investment Income	IRPINT, IROFMINT, IRFAMINT
Child Support Payments	IRPCHD, IROFMCHD, IRFAMCHD
Wages	IRPWAG, IROFMWAG, IRFAMWAG
Other Income	IRPOTH, IROFMOTH, IRFAMOTH
Food Stamps	IRFSTAMP
Welfare Months	IRWELMOS
Total Family Income	IRPINC1, IRFINC1, IRFAMIN1

9.2.8 Multivariate Imputation

Sections 9.2.1 through 9.2.7 summarize the specifics of separating the set of binary income variables (in the 2007 NSDUH) into item respondents and item nonrespondents. These sections also describe model building, computation of predicted means, and the assignment of imputed values for these measures using a univariate predicted mean. In most cases, however, these univariate assignments were only provisional. The final imputed values for these income measures were obtained from neighborhoods built on a vector of predicted means using the multivariate predictive mean neighborhood (MPMN) technique, as described in Appendix C. Consistent with the univariate imputations, the multivariate assignments were performed separately within four age groups of respondents: 12 to 17, 18 to 25, 26 to 64, and 65 or older.

For these source-of-income variables, a single months-on-welfare variable, and the binary total income variables, the collective distance between their conditional predicted means for a given incomplete data respondent and the complete data respondents was determined using a Mahalanobis distance[116] within each age group. As with other applications of MPMN, the predictive mean vector used in the Mahalanobis distance calculation included only variables that were missing for a given item nonrespondent. For the recipient, only missing values among the variables were replaced by the donor's values. For example, if the respondent was missing only a response for the other-family welfare payments question, then only the donor's other-family welfare payments response was given to the recipient.

The predicted mean that results from the months-on-welfare model was a logit of the proportion of the year received. This logit was back-transformed into a proportion, which was

[116] See Appendix C for a definition of Mahalanobis distance. A definition can also be found in Manly (1986).

the predicted mean used to match donors to each recipient. This meant that the proportion could be treated as a probability, which in turn could be multiplied by the probability of receiving welfare in the past year. Hence, the matching predicted mean could be made conditional on the receipt of welfare in the past year, if necessary. More details about how the months-on-welfare predicted mean was made conditional on receipt of welfare in the past year are presented in Appendix G.

Candidate donors were restricted according to logical constraints, which could not be loosened. As with the univariate provisional imputations, donors and recipients were required, as a logical constraint, to have the same value for the family skip variable. In addition, if a respondent was missing the months-on-welfare question, but was not missing one of the feeders to this question, the donor and recipient were required to have the same values for the nonmissing feeder question variables. For months on welfare, the feeder questions were those involving welfare payments or welfare services. Missingness patterns and the logical constraints imposed for the binary income variables are presented in Appendix G.

A number of likeness constraints also were imposed on the multivariate neighborhood for the binary income variables. The donors were usually restricted to those who were the same age as the recipient or, if that constraint was too restrictive, an age within 5 years of the recipient. There was a high degree of association between respondents who received welfare payments, welfare services, and food stamps. There was also a high degree of association between respondents earning an income from investments and respondents who had high incomes, both of which were negatively associated with welfare payments, welfare services, and food stamps. Hence, if a recipient required imputation for one or more of these six variables (i.e., welfare payments, welfare services, food stamps, binary income, investment income, and months on welfare), but had information on at least one of these variables, the donors were restricted so that donors and recipients had the same values for these nonmissing variables. If one of the pair of income variables (personal and other-family-member source of income, or personal and family income) was missing, the donor and recipient were required to have the same value for the nonmissing variable.

Some other likeness constraints corresponded to covariates that were highly correlated with the response, but these constraints often were not included in SUDAAN® models. This was due to near-empty cells when the variables were cross-tabulated, causing instability in the estimates. In particular, this affected the following personal and/or other-family-member binary source-of-income variables: welfare payments, welfare services, child support, wages, and Social Security. The welfare and child support variables were strongly related to whether children were in the household. Because the variable representing the number of children younger than 18 in the household was included in the models, the following likeness constraint was added: both the donor and recipient had to have children either younger than 18 in the household or 18 or older in the household. This constraint was applied if one or more of the source-of-income variables (either personal or other family) for welfare payments, welfare services, or child support were missing. Likewise, new likeness constraints were added for the wages variable, which was highly correlated with employment status and, for respondents aged 65 or older, whether someone younger than 65 was in the household. Specifically, if the personal wages response was missing among respondents aged 15 or older, both the donor and recipient had to be either working or not working. Among respondents aged 65 or older, if personal wages or other-family-member wages

variables were missing, both the donor and recipient had to have someone either aged 18 to 64 in the household or not aged 18 to 64 in the household.

Finally, if the other-family-member Social Security value was missing, both the donor and recipient had to have someone either aged 65 or older in the household or not aged 65 or older in the household. If insufficient donors were present, the constraints were loosened in the following order: (1) abandoned the neighborhood and chose the donor with the closest predicted mean; (2) removed the requirement that donor and recipient needed to be of the same age, but required them to be within 5 years of each other; (3) removed the requirement that the donor and recipient be within 5 years of age of each other; (4) removed the constraint that incorporated the association between the welfare, food stamps, investment income, and total income questions; (5) removed the months-on-welfare constraints regarding personal and other-family-member welfare payments and services, and replaced it with a less strict requirement that the donor's and recipient's family welfare payments and services must be a match; and (6) removed the number-65-or-older constraint among respondents missing Social Security information and the number-younger-than-18 constraint among respondents missing welfare payments, welfare services, and child support information. The likeness constraints and the number of recipients with sufficient donors corresponding to each likeness constraint are summarized in Appendix G.

9.2.9 Binary Income Recode: GOVTPROG

The dichotomous recoded income variable GOVTPROG indicated whether the respondent participated in any government assistance programs. It was created from four imputation-revised variables: family Supplemental Security Income (IRFAMSSI), family food stamps (IRFSTAMP), family welfare payments (IRFAMPMT), and family welfare services (IRFAMSVC). Although a variety of recoded variables were created, only GOVTPROG is described here because it was used as a covariate in subsequent health insurance models. (See Chapter 10 for details on the imputation of missing values in the health insurance variables.)

9.3 Finer Category Phase

9.3.1 Hierarchy of Income Variables

Three income variables resulted from editing the questions in the finer income category phase: personal total income (PINC2), total family income if there are other family members (FINC2), and total family income (FAMINC2). These three variables were all considered simultaneously using a failure time model, which is described in Section 9.3.3. Because only one model was fit, no hierarchy was required.

9.3.2 Setup for Model Building

As with the variables in the binary variable phase, the imputations were conducted separately within the four age groups of respondents: 12 to 17, 18 to 25, 26 to 64, and 65 or older. For an individual to be considered an item respondent for income variables in the finer category phase, he or she must have complete data for both questions in this phase. Response propensity adjustments were computed for each age group to make the item respondent weights representative of the entire sample, and the appropriately adjusted weights were used in the

models. As with health insurance and the binary income variables, the final analysis weights were used as weights. The variables included in the model, which predicted the probability of item nonresponse, were the same as those included in the imputation model. Details are discussed in the next section.

9.3.3 Sequential Model Building

The finer categories of income were modeled using the LIFEREG procedure in SAS/STAT® software.[117] This procedure was used for regression modeling of continuous nonnegative random variables, such as survival times and income, by fitting models that are sometimes referred to as "failure time models." This particular type of model, which was assumed for the response variable representing income, can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$ is a vector of observed responses, $\mathbf{X}$ is the matrix of covariates, $\boldsymbol{\beta}$ is the parameter vector, and $\boldsymbol{\varepsilon}$ is a vector of error terms. Particularly, the error terms are assumed to come from a known multivariate distribution, such as the logarithm of a three-parameter generalized gamma model, or a more common two-parameter distribution, such as gamma, Weibull, lognormal, or log-logistic. Although the underlying random variable $\mathbf{y}$ is assumed to be continuous, the LIFEREG procedure allows the variable to be reported in interval categories, such as the NSDUH income intervals. The contribution of an individual with covariates in the matrix $\mathbf{X}$ to the overall likelihood is simply the probability mass assigned by the model to the interval $(l, u]$ containing the actual continuous income for that individual. For this interval, l represents the lower bound and u represents the upper bound. This contribution has the form $F(u|\mathbf{X},\boldsymbol{\beta},\sigma^2) - F(l|\mathbf{X},\boldsymbol{\beta},\sigma^2)$, where F is a cumulative distribution function and σ^2 represents the variance of the individual responses. The LIFEREG procedure uses standard likelihood methods of inference and incorporates the survey weights.

LIFEREG allowed several choices for the functional form of the parametric model that corresponded to the error distribution, including the two-parameter log-logistic, lognormal, gamma, and Weibull and the three-parameter generalized gamma. Each of these models was fit to each of the four age group-specific datasets. Compared with the other models, the gamma distribution provided a better overall fit, as indicated by likelihood techniques. Because the three-parameter generalized gamma did not significantly improve on its two-parameter special cases, when using the likelihood ratio tests as criteria for comparison, it was decided to use a two-parameter model.

Many of the covariates considered in the model for the finer category phase included the same covariates used in the binary variable phase. These covariates included centered continuous age, centered age squared, gender, race, region, population density, percentage Hispanic/Latino population, percentage non-Hispanic/Latino black/African-American population, percentage owner-occupied households, imputation-revised number of adults in household, imputation-revised number of children in household, imputation-revised number of adults aged 65 years or

[117] Details about the LIFEREG procedure are discussed in the *SAS/STAT User's Guide, Version 8* (SAS Institute, 1999).

older in household, and a three-level State-rank variable. As in the binary variable phase, the State-rank groups in the finer category group were defined in terms of the proportion of a given State's residents whose incomes were $20,000 or more. For both phases, there were also predictors that consisted of one-way interactions of centered age with race, centered age with gender, race with gender, centered age squared with race, and centered age squared with gender. For the three older age groups, the additional covariates of marital status, education status, and employment status were used for both the binary variable phase and the finer category phase. Also, all imputation-revised income indicators considered in the binary variable phase were used as covariates for the finer category phase.

9.3.4 Computation of Predicted Means and Univariate Predictive Mean Neighborhoods

As described in the previous section, the failure time model contained the term $\mathbf{X\beta}$, which was the predictive mean value. This value was a monotonic function of the conditional mean of the modeled income distribution at a given individual set of values of the regression covariates. Specifically, $\mathbf{X\beta}$ was a translation of the estimated mean of log income. Mean values were computed for both item respondents and item nonrespondents using the parameters from the failure time model. Subsequently, these values were used to assign imputed values using the UPMN imputation method, which is described in Appendix C.

9.3.5 Assignment of Imputed Values

Separate assignments of imputed values were performed within each of the four age groups for all finer category income variables. Only missing values were replaced by imputed values using the same donor for both personal and family finer income variables. The multivariate imputation process is further described in Section 9.3.7.

9.3.6 Constraints on Univariate Predictive Mean Neighborhoods

Donors and recipients were required to have the same values for both the binary personal and family income variables and the indicator of whether other family members were in the household (IRFAMSKP). In addition, if either of the personal income or family income finer category responses were nonmissing, donors and recipients were required to have the same values for the nonmissing variable. Finally, donors were required to have predictive mean values "close to" (within the delta distance) recipients' predictive mean values. If insufficient donors were available using these constraints, the constraint involving nonmissing personal or family income finer category responses was loosened to a logical constraint. This logical constraint required the recipient's nonmissing value to be consistent with the donor's value for the other variable. Finally, if no donors were available, the neighborhood was abandoned, and the donor with the closest predicted mean to the recipient was chosen, subject to the logical constraints. The likeness constraints and the number of recipients with sufficient donors corresponding to each likeness constraint are summarized in Appendix G.

9.3.7 Multivariate Assignments

The predicted means were calculated using the edited (finer category) family income variables (see Table 9.2) as the response variables. For each family income variable,

neighborhoods were created using scalar predicted means from the appropriate model. The methodology for determining the neighborhood was therefore univariate in terms of these scalar predicted means. Three edited variables were associated with each predicted mean so that the missing values for the three variables required assignment of imputed values. Hence, even when determining the provisional imputed values using the univariate procedure, the assignment of imputed values was multivariate for all but two of the variables. For the 2007 NSDUH, the imputation-revised variable for the personal income variable was called IRPINC2, the family income variable with legitimate skips was called IRFINC2, and the family income variable without legitimate skips was called IRFAMIN2.

9.3.8 Imputation-Revised Value Reassignments for Sample B

In the 2007 NSDUH, after the binary and finer category income internal variables were imputed, the values of the imputation-revised and the imputation indicator variables needed to be modified. As mentioned in Section 9.2.2., for Sample B, legitimate skip values were assigned to those variables that were not in the reduced set of income questions. The sole purpose of the internal variables was in their use as covariates for subsequent processing. Also, because the imputation-revised variables were based on the internal variables with missing values set at the beginning of the processing, they needed to be reassigned back to the legitimate skip code after the imputation processing had been completed. The imputation indicator variables were processed in a similar manner. As shown in Table 9.1, in the right three columns (i.e., the "Reduced Set" columns), all the imputation-revised and the imputation indicator variables corresponding to the "—" symbol were assigned a legitimate skip code and delivered to the analytic file. However, no changes were made to the variables in Sample A because the respondents in this group answered all the original income questions.

9.3.9 Finer Category Income Recodes: INCOME and INCOME5

The recoded variable INCOME classified the families of respondents into four income levels: less than $20,000; $20,000 to $49,999; $50,000 to $74,999; and $75,000 or more. Another recoded variable (INCOME5) was created to take advantage of an extra level of income. This variable had five levels: the first three levels were equivalent to INCOME, but the last level of INCOME was separated into two levels: $75,000 to $99,999; and $100,000 or more. Both INCOME and INCOME5 were recodes of the variable IRFAMIN2. A variety of recoded variables were created but are not discussed in this report. However, as with GOVTPROG, the variable INCOME is discussed here because it was used as a covariate in subsequent health insurance models (see Chapter 10 for details on the imputation of missing values in the health insurance variables). INCOME5, which is currently used for special requests, also is discussed because it is similar to the INCOME variable and because it might be used in place of INCOME in future NSDUHs.

10. Health Insurance

10.1 Introduction

For the National Survey on Drug Use and Health (NSDUH),[118] the health insurance imputations were divided into two methods: the "old method" and the "constituent variables method." The old method was used to impute three overall health insurance variables in a way that was consistent with previous iterations of NSDUH. The first variable, IRPINSUR, was simply an imputation-revised version of the edited "private health insurance" variable. The second and third variables, IRINSUR and IRINSUR3, were both indicators of "any health insurance" coverage. These different versions of health insurance coverage indicators were created because the question set changed between 1999 and 2001.

The constituent variables method was used to impute the specific health insurance variables in two stages. The first stage imputed four specific health insurance variables individually: IRMCDCHP, IRMEDICR, IRCHMPUS, and IRPRVHLT, which are indicators of coverage for Medicaid/CHIP, Medicare, CHAMPUS, and private health insurance, respectively. The second stage created the imputation-revised "any other health insurance" variable, IROTHHLT. The five constituent imputation-revised health insurance variables were then used to create the overall health insurance variable, IRINSUR4.

Regardless of whether the final health insurance variables were derived by the old method or the constituent variables method, imputations were performed using the same methodology, the predictive mean neighborhood (PMN) technique, as described in Appendix C.

10.2 Edited Insurance Variables

Table 10.1 shows the edited counterparts for some of the health insurance questionnaire (raw) variables. In the 2007 survey, the edited variables had the same values as the questionnaire variables, except that missing values were replaced by standard NSDUH missing value codes.

10.2.1 Edited Insurance Variables (Old Method)

Using the old method, three health insurance variables, INSUR, INSUR3, and PINSUR, were created from the six edited variables shown in Table 10.1. Two of them, INSUR and INSUR3, indicated whether the respondent had "any" health insurance. The third, PINSUR, indicated whether the respondent had any "private" health insurance. INSUR3, which was consistent with the variable of the same name created in the 2001 survey, was coded as "yes" if any one of the six variables listed in Table 10.1 were coded as "yes" and was coded as "no" if all six variables were coded as "no." The other overall health insurance indicator, INSUR, was created to maintain consistency with the 1999 survey. Because the questions associated with CHIPCOV (Children's Health Insurance Program) and HLTINNOS (covered by any kind of

[118] This report presents information from the 2007 National Survey on Drug Use and Health (NSDUH), an annual survey of the civilian, noninstitutionalized population of the United States aged 12 or older. Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA).

health insurance) did not exist in the 1999 questionnaire, these two variables were excluded from the determination of INSUR. The INSUR variable was coded as "yes" if any of the other four variables listed in Table 10.1 were coded as "yes" and was coded as "no" if all four variables were coded as "no."[119]

Table 10.1 Mapping of Raw Health Insurance Variables to Edited Counterparts

Question Number[1]	Question Text[2]	Edited Variable[3]
QHI01 QHI01v	Is the respondent covered by Medicare?	MEDICARE (1 = yes, 2 = no)
QHI02, QHI02v	Is the respondent covered by Medicaid or Medical Assistance?	MEDICAID (1 = yes, 2 = no)
QHI02A	Is the respondent currently covered by a Children's Health Insurance Program operated by your State of residence?[4] (asked only of respondents aged 12 to 19)	CHIPCOV (1 = yes, 2 = no)
QHI03	Is the respondent currently covered by CHAMPUS or TRICARE, CHAMPVA, the VA, or military health care?	CHAMPUS (1 = yes, 2 = no)
QHI06	Is the respondent currently covered by private health insurance?	PRVHLTIN (1 = yes, 2 = no)
QHI11	Is the respondent currently covered by any kind of health insurance, that is, any policy or program that provides or pays for medical care?	HLTINNOS (1 = yes, 2 = no, 99 = legitimate skip[5])

[1] The "v" questions were asked to verify the answer given in the previous question for respondents who were younger than 65 and a Medicare recipient or older than 65 and a Medicaid recipient.

[2] The questions provided in this table are abbreviated versions of those given in the questionnaire.

[3] Missing values in these edited values were represented by standard missing value codes. CHIPCOV was replaced in the final analytic file by CAIDCHIP, a combination of MEDICAID and CHIPCOV. See Section 10.2.2 for details.

[4] The questionnaire did not ask the question exactly in this way. It identified the specific program, depending upon the State of residence entered by the respondent.

[5] A respondent was assigned a legitimate skip for HLTINNOS if they answered "yes" or gave no answer to at least one of the other health insurance questions.

Only the edited variable PRVHLTIN (whether the respondent was covered by private health insurance at the time of the survey) was used to create PINSUR. Missing data for the edited variable PRVHLTIN were coded using the standard NSDUH missing data codes for "don't know," "refused," and "blank," whereas missing data for PINSUR were all coded as "98," which was a code for missing data. Except for the codes used to handle missing data, PINSUR and PRVHLTIN were equivalent. The variable PINSUR was created to maintain consistency with pre-1999 surveys, in which other variables also contributed to the indicator of coverage by private health insurance. All respondents with private health insurance were considered to have health insurance. Therefore, respondents with private health insurance were a subset of the respondents who had health insurance.

[119] In the 2000 survey, the variable INSUR2 was created to take advantage of the additional information provided by questions that did not exist in the 1999 questionnaire. However, because these additional questions were either replaced or reworded in later surveys, the variable INSUR2 has not been used in the surveys since 2000.

10.2.2 Edited Insurance Variables (Constituent Variables Method)

Using the constituent variables method, the editing process combined the variables MEDICAID (whether the respondent was covered by Medicaid or Medical Assistance) and CHIPCOV (whether the respondent was currently covered by a Children's Health Insurance Program) to create the variable CAIDCHIP, which indicated whether someone was covered by Medicaid or one of the State children's health plans. This variable and all the other edited variables in Table 10.1, except HLTINNOS, were used directly as base variables for imputation.

A respondent was routed to QHI11 (whether the respondent was covered by any kind of health insurance at the time of the survey) if they answered "no" to all the other health insurance questions. All other respondents were given a legitimate skip value to the variable HLTINNOS, as shown in Table 10.1. Therefore, it was possible that the imputation-revised versions of the four specific health insurance variables would all have had a value of "no," and the value of HLTINNOS would have been a legitimate skip, if one or more of the "no" values was imputed. In this instance, another variable was needed to reflect the fact that a respondent could have had a valid "yes" or "no" imputed value for "any other health insurance," even though the respondent was never asked QHI11 and was assigned a legitimate skip code. Thus, the variable ANYOTHER was created using HLTINNOS and an additional edited variable, SKHLCCOV, which indicated whether a respondent was covered by any health insurance. SKHLCCOV and ANYOTHER were defined as follows:

SKHLCCOV =

- 1 (or 3) if CAIDCHIP = 1, MEDICARE = 1, CHAMPUS = 1, or PRVHLTIN = 1[120]; else
- 2 if CAIDCHIP = 2, MEDICARE = 2, CHAMPUS = 2, and PRVHLTIN = 2; else
- missing value code if the nonmissing values of CAIDCHIP, MEDICARE, CHAMPUS, and PRVHLTIN are all "2," and at least one of these variables had a missing response.

ANYOTHER =

- legitimate skip code (99) if SKHLCCOV = 1 or 3; else
- SKHLCCOV if SKHLCCOV = 2 or a missing value code.

10.3 Imputation-Revised Health Insurance Variables (Old Method)

The old method of creating the final imputation-revised health insurance variables amounted to imputing missing values in the recoded variables (INSUR and INSUR3), as described in the previous section and in PINSUR. This resulted in the creation of three

[120] SKHLCCOV was coded as a 3 if the respondent was covered by a State children's health insurance program but was not covered by Medicaid, Medicare, CHAMPUS, or private health insurance. Respondents with SKHLCCOV = 3 were treated in the same manner as those with SKHLCCOV = 1.

imputation-revised variables: two for overall health insurance (IRINSUR and IRINSUR3) and one for private health insurance (IRPINSUR).

10.3.1 Order of Modeling Health Insurance Variables (Old Method)

The old method used multivariate predictive mean neighborhood (MPMN) imputation for private health insurance and overall health insurance. However, respondents who answered "yes" to the private health insurance question also were logically covered by overall health insurance. Therefore, it was not possible to use INSUR or INSUR3 as covariates in the PINSUR model, or vice versa.

10.3.2 Setup for Model Building (Old Method)

After determining the modeling order of the health insurance variables, the next step was to define respondents, nonrespondents, and the item response mechanism. Imputations for all three health insurance variables were conducted separately within four age groups: 12 to 17, 18 to 25, 26 to 64, and 65 or older.

One model was created for PINSUR and another for INSUR3. A respondent was considered an item respondent for health insurance only if his or her status was known for both private health insurance and overall health insurance as defined by INSUR3. To meet this criterion, the respondent must have had a valid "yes" or "no" response in PRVHLTIN (the edited variable corresponding to QHI06 [whether the respondent was currently covered by private health insurance]). In addition, he or she either must have answered QHI01, QHI02, QHI02A, QHI03, or QHI11[121] (see Table 10.1 for descriptions of these variables) with a valid "no" response or must have answered "yes" to at least one of the six questions (including QHI06). This ensured that the interview respondent's status with respect to both overall health insurance (INSUR3 definition) and private health insurance was completely known. For example, if the interview respondent did not answer QHI01, but answered "no" to the other five questions, his or her status with respect to overall health insurance depended on the missing response to QHI01. However, if the respondent answered "yes" to any of the other five questions, the value of INSUR3 was already known to be "yes."

Note that it was possible for a respondent to be defined as an item nonrespondent for INSUR3, but as an item respondent for INSUR. This occurred if a respondent gave valid "no" answers to QHI01, QHI02, QHI03, and QHI06, but he or she did not answer QHI02A or QHI11 (and did not give a valid "yes" answer to either of these). On the other hand, because the variables making up INSUR constituted a subset of those corresponding to INSUR3, an item nonrespondent for INSUR was necessarily an item nonrespondent for INSUR3. Moreover, an item nonrespondent for PINSUR was necessarily an item nonrespondent for INSUR3. Because missing values in all three variables (PINSUR, INSUR, and INSUR3) were imputed, an item respondent was defined based on the response to INSUR3.

[121] References to QHI01 and QHI02 naturally imply that if the respondent was younger than 65 and answered "yes" to QHI01, then he or she also answered QHI01v. Moreover, if the respondent was 65 or older and answered "yes" to QHI02, then he or she also answered QHI02v.

To ensure that the weights adequately represented the population, the weights for item nonrespondents (as defined by INSUR3) were reallocated to item respondents using item response propensity models within each age group for the pair INSUR3 and PINSUR. (Because the modeling of the final weight adjustments was not completed at the time of the health insurance imputations, the person-level sample design weights were adjusted to account for nonresponse at the household level using a simple ratio adjustment.)[122] The item response propensity model is a special case of the generalized exponential model (GEM),[123] which is described in Appendix B. The variables included in the model predicting the probability of item nonresponse were the same as those included in the main model, which is discussed in the next section.

10.3.3 Sequential Model Building (Old Method)

The probability that the respondent had health insurance (as defined by INSUR3) and the probability that the respondent had private health insurance were both modeled for item respondents, within each age group, using the nonresponse adjusted weights. The private health insurance model was created only for respondents who were known to have overall health insurance so that the predicted probability modeled was $P(\text{PINSUR} = 1 \mid \text{INSUR3} = 1)$. For the models, the parameters were estimated using logistic regression.[124] Each response propensity model included the following pool of predictors: centered age,[125] race/ethnicity, centered age squared, centered age cubed, gender, population density, percentage of housing in segment that was owner-occupied, percentage of Hispanics/Latinos in the segment, percentage of non-Hispanic/Latino blacks/African Americans in the segment, and household size. There were also predictors that consisted of one-way interactions of centered age with race/ethnicity, centered age with gender, race/ethnicity with gender, centered age squared with race/ethnicity, and centered age squared with gender. For the three older age groups (18 to 25, 26 to 64, and 65 or older), the additional predictors of marital status, education level, and employment status also were considered in each model.

10.3.4 Computation of Predicted Means (Old Method)

Using the parameter estimates from models for overall and private health insurance, predicted probabilities of having insurance were computed for both item respondents and nonrespondents. In other multivariate imputations, a hierarchy was required, where provisional imputations were performed on variables earlier in the hierarchy to be used as covariates in variables further down the hierarchy. A final multivariate imputation was then performed on all variables in the hierarchy. However, because neither variable could be used as a covariate in the model for the other variable, no provisionally imputed values were required.

---

[122] In subsequent text, the use of the word "weights" will refer to the ratio-adjusted design weights.

[123] The GEM macro, which was written in SAS/IML® software, was developed at RTI International (a trade name of Research Triangle Institute) for weighting procedures.

[124] In the 2007 survey, the software used for most imputation modeling was SUDAAN®. However, the logistic model for the old method of imputing health insurance variables used SAS® to maintain consistency with the practice of previous survey years. SAS software is a registered trademark of SAS Institute, Inc. SUDAAN is a registered trademark of Research Triangle Institute.

[125] The covariate age was centered within each age group to reduce the effects of multicollinearity, particularly with the squared and cubed age terms. For more information on "centering" and "multicollinearity," refer to Draper and Smith (1981).

10.3.5 Multivariate Imputation of Health Insurance and Private Health Insurance (Old Method)

The final imputed values for overall health insurance (using both the INSUR and INSUR3 definitions) and private health insurance were obtained using neighborhoods built upon a vector of predicted means. The vector had two elements: *P*(overall health insurance, as defined by INSUR3) and *P*(private health insurance | overall health insurance, as defined by INSUR3). For both overall and private health insurance, the imputation method used was the MPMN procedure, which is described in Appendix C. Similar to the response propensity models, the multivariate assignments were done separately within the same four age groups: 12 to 17, 18 to 25, 26 to 64, and 65 or older.

A respondent was eligible to be a donor for a given item nonrespondent if he or she had complete data across PINSUR, INSUR, and INSUR3 and was within the same age group. Logical constraints were placed on individuals who were missing one or two of the three indicators. Respondents who were missing either of the overall health insurance indicators, but did not have private health insurance, required donors who also did not have private health insurance.[126] If a respondent was missing only INSUR3, then INSUR must have been "no" because a "yes" value for INSUR would have necessarily meant that INSUR3 would have been "yes" and therefore nonmissing. Hence, donors also must have had a "no" value for INSUR. By the same token, if a respondent was missing only INSUR or was missing both PINSUR and INSUR, but not INSUR3, then INSUR3 must have been "yes" because a "no" value for INSUR3 would have necessarily meant that INSUR would have been "no" and therefore nonmissing. In this case, donors must also have had a "yes" value for INSUR3. Finally, respondents who indicated that they had health insurance, but were missing the private health insurance indicator, required donors who had some health insurance.[127] As a likeness constraint, potential donors were then further restricted to be the same age as the recipient. If no eligible donors were available who were the same age as the recipient, donors were sought who were within 5 years of the recipient. Finally, donors were required to have all applicable elements of the multivariate predictive mean vector "close to" (i.e., within the delta distance) the recipient's elements of the predictive mean vector. Because the imputation was multivariate, the set of deltas was also multivariate, where a different delta corresponded to each element of the predictive mean vector. Likeness constraints were loosened in the order listed above. Appendix G summarizes the patterns of missingness for overall and private health insurance, the logical constraints imposed on the set of donors, the frequency of occurrence of each missingness pattern, the likeness constraints, and the number of recipients with sufficient donors corresponding to each likeness constraint.

The full predictive mean vector contained elements for overall health insurance (as defined by INSUR3) and private health insurance (conditional on a "yes" response to the overall

[126] Technically, this was not a logical constraint because there was no restriction on whether the respondent did or did not have health insurance. However, because all respondents with private health insurance had health insurance and the recipient did not have private health insurance, the distribution would have been skewed in favor of a "yes" indicator if these respondents were allowed to be donors.

[127] Again, this technically was not a logical constraint. However, because all respondents who did not have health insurance also did not have private health insurance and the recipient had health insurance, the distribution would have been skewed in favor of a "no" indicator if these respondents were allowed to be donors.

health insurance [INSUR3] indicator). The portion of the full predictive mean vector used to determine the neighborhood for a particular item nonrespondent was dependent on the pattern of missingness for that item nonrespondent. If a respondent was missing INSUR, but not INSUR3, the predicted mean that was derived using INSUR3 was used. The portions of the full predictive mean vector used to create the MPMN for each missingness pattern, with accompanying adjustments, are provided in Appendix G. The Mahalanobis distance[128] was then calculated using only the portion of the predictive mean vector that was associated with the given missingness pattern. If no donors were available who had predicted means within a multivariate delta of the recipient's vector of predicted means, the neighborhood was abandoned, and the respondent with the closest Mahalanobis distance was selected as the donor. The procedure is described in detail in Appendix C.

10.4 Imputation-Revised Specific Health Insurance Variables (Constituent Variables Method, First Stage)

The constituent variables method of creating the final imputation-revised health insurance variables amounted to imputing missing values in each of the edited health insurance variables that, when combined together, constituted "overall health insurance." In the first stage of this method, four imputation-revised specific health insurance variables were created representing whether the respondent had health insurance from Medicaid or a State children's health insurance program (IRMCDCHP), Medicare (IRMEDICR), CHAMPUS (IRCHMPUS), or private health insurance (IRPRVHLT). Missing values in these variables were imputed in a multivariate imputation. These final variables were derived from the edited variables CAIDCHIP, MEDICARE, CHAMPUS, and PRVHLTIN, respectively. The second stage is described in Section 10.5.

10.4.1 Order of Modeling Health Insurance Variables (Constituent Variables Method, First Stage)

The first step in imputing the four specific health insurance variables was to determine the order in which the variables were to be modeled. A motivation for using a hierarchy in PMN for drug use variables is provided in Appendix C; this same rationale was used in developing the hierarchy for the health insurance variables. For a model predicting whether a respondent had a specific type of health insurance, other types of health insurance were useful covariates. Following a provisional imputation of missing health insurance values, the indicators earlier in the sequence were used as covariates for health insurance variables later in the sequence. Any imputed values in the health insurance variables were considered temporary at this point. This was because the final imputation was not done for health insurance variables until the modeling was completed for all four specific health insurance variables. The health insurance indicators were imputed in the following order: CAIDCHIP, MEDICARE, CHAMPUS, and PRVHLTIN.

[128] See Appendix C for a definition of Mahalanobis distance. A definition also can be found in Manly (1986).

10.4.2 Setup for Model Building (Constituent Variables Method, First Stage)

Once the hierarchy of health insurance variables was determined, the next step was to define respondents, nonrespondents, and the item response mechanism. For an individual to be considered an item respondent for the specific health insurance variables, he or she had to have complete data for the four edited specific health insurance variables. Imputation for CAIDCHIP, CHAMPUS, and private health insurance were conducted within the four age groups: 12 to 17, 18 to 25, 26 to 64, and 65 or older. Imputation for Medicare was conducted within the following three age groups: 12 to 17, 18 to 64, and 65 or older.[129]

Response propensity adjustments were then computed for each age group to make the item respondent weights representative of the entire sample. The covariates in the item response propensity model included centered age, centered age squared, gender, race/ethnicity, population density, percentage of housing in that segment that was owner-occupied, and a three-level income variable. There were also predictors that consisted of one-way interactions of centered age with race/ethnicity, centered age with gender, race/ethnicity with gender, centered age squared with race/ethnicity, and centered age squared with gender. For the three older age groups (18 to 25, 26 to 64, and 65 or older), the additional predictors of marital status, education level, and employment status also were considered in each model.

10.4.3 Sequential Model Building (Constituent Variables Method, First Stage)

Starting with CAIDCHIP, the probability that an individual was covered by a given type of health insurance was modeled for item respondents, within each age group, using the nonresponse-adjusted weights. For the models, the parameters were estimated using logistic regression in SUDAAN®.[130] The predictors included in all models were centered age, centered age squared, gender, race/ethnicity, population density, and percentage of housing in that segment that was owner-occupied. There were also predictors that consisted of one-way interactions used for the three younger age groups (12 to 17, 18 to 25, and 26 to 64): centered age with race/ethnicity, centered age with gender, race/ethnicity with gender, centered age squared with race/ethnicity, and centered age squared with gender. For the three older age groups (18 to 25, 26 to 64, and 65 or older), the additional predictors of marital status, education level, and employment status also were considered in each model. Additional predictors were specific to each model, depending upon the response variable of interest, and are described below.

Predictors for the CAIDCHIP model included household size; a four-level family income variable;[131] binary indicators of whether the respondent's family in the household received income from public assistance, wages, interest, or social security; and for respondents aged 18 or

[129] The age groups 18 to 25 and 26 to 64 were combined for the Medicare variable because (1) only a small proportion of respondents in these age groups had Medicare, particularly for the 18-to-25 age group, and (2) a respondent of working age could have received Medicare only if he or she was not working because of disability. This was true regardless of whether the respondent was aged 18 to 25 or 26 to 64.

[130] SAS®-callable SUDAAN® was used to fit all binomial and polytomous logistic regression models. Details about the logistic regression model and additional references can be found in the *SUDAAN Language Manual Addendum, Release 9.0.3* (RTI International, 2007). SAS software is a registered trademark of SAS Institute, Inc. SUDAAN is a registered trademark of Research Triangle Institute.

[131] The four levels of the family income variable were less than $20,000; $20,000 to $49,999; $50,000 to $74,999; and $75,000 or more.

older, a binary indicator of whether the respondent had other family members in the household. The MEDICARE model included predictors for a binary indicator of whether the respondent was on social security for respondents aged 18 or older and a binary indicator of whether anyone in the respondent's family in the household received social security for respondents younger than 18. For CHAMPUS, predictors included a binary indicator of whether the respondent (or, if the respondent was younger than 18, the respondent's family in the household) received income from sources other than those given in the binary income questions (see Chapter 9 for details); a three-level income variable;[132] and for respondents aged 18 or older, an indicator of whether the respondent had ever been in the military service, designated by an imputation-revised version of the edited variable SERVICE.[133] The PRVHLTIN model included predictors for household size; a four-level family income variable (the same variable that was used in the CAIDCHIP model); binary indicators of whether the respondent's family in the household received income from public assistance, wages, interest, social security, or sources other than those given in the binary income questions; and for respondents aged 18 or older, a binary indicator of whether the respondent had other family members in the household.[134] The complete summary of the health insurance models can be found in Appendix F.

10.4.4 Computation of Predicted Means and Univariate Predictive Mean Neighborhoods (Constituent Variables Method, First Stage)

Following the modeling for the four specific health insurance variables corresponding to CAIDCHIP, MEDICARE, CHAMPUS, and PRVHLTIN, in the sequence listed in Section 10.4.1, missing values were replaced by provisional imputed values. This was necessary so that these variables could be used as covariates in subsequent models. Although no provisional imputed values were used to build the models, it was necessary to calculate predicted means for all respondents, including item nonrespondents, using the parameter estimates from the models. This sometimes required the use of the provisional values for the covariates. The predicted probabilities from these models were used to assign provisional values using the univariate predictive mean neighborhood (UPMN) imputation method as described in Appendix C.

10.4.5 Multivariate Imputation of Specific Health Insurance Variables (Constituent Variables Method, First Stage)

The final imputed values for CAIDCHIP, MEDICARE, CHAMPUS, and PRVHLTIN were obtained using neighborhoods built upon a vector of predicted means. For these four variables, the imputation method used was the PMN procedure, as described in Appendix C. Similar to the response propensity models, the multivariate assignments were done separately within the same four age groups: 12 to 17, 18 to 25, 26 to 64, and 65 or older. No logical constraints were applied to the health insurance variables, because no internal inconsistencies would have resulted from any type of donor. However, a number of likeness constraints were

[132] The three levels were less than $20,000; $20,000 to $49,999; and $50,000 or more.

[133] The variable SERVICE generally had a very low level of missingness (0 missing values in the 2007 survey). Because covariates in these models were not supposed to have any missing values, the missing value in the SERVICE variable was randomly imputed as a "yes" if the random number was greater than the mean value of SERVICE across all the other respondents, and imputed as "no" otherwise.

[134] If the respondent did not have other family members in the household, the family income binary indicators listed as predictors were equivalent to the personal income binary indicators.

applied, depending upon the missingness pattern. The variables that were included as likeness constraints were highly correlated with the response variables but, in most cases, could not be included as predictors in the models because of the large number of missing values in the predictors. In general, any nonmissing values that the recipient had for CAIDCHIP, MEDICARE, CHAMPUS, or PRVHLTIN had to match between donor and recipient, though this constraint was often the first one that was loosened. In addition, the donor's predicted mean(s) for each variable that was missing was required to be within 5 percent of the recipient's predicted mean(s). This was usually the last constraint to be loosened. Finally, specific likeness constraints were associated with each of the four variables and are discussed briefly below. The order in which the constraints were loosened depended upon the missingness pattern, and these constraints are described in Appendix G. The portions of the full predictive mean vector used to create the multivariate neighborhoods for each missingness pattern, with accompanying adjustments, also are provided in Appendix G.

For the variable CAIDCHIP, the donor and recipient had to have the same status regarding whether or not a respondent's family had received any government public assistance. This was measured by the variable GOVTPROG, which is described in Chapter 9. For the variable MEDICARE, a respondent of working age (between 18 and 64) could have received Medicare only if he or she were not working because of disability. If MEDICARE was missing, a constraint was included that required donors and recipients to have the same status in this regard, using the appropriate level of the variable JBSTATR (respondent work situation in the past week). This constraint was never loosened. In addition, the donor and recipient had to have the same status regarding whether or not a respondent's family had received Social Security.

In the models for CHAMPUS, two variables were included as covariates that also were used as likeness constraints. An imputation-revised version of the variable SERVICE (whether the respondent had ever been in the military service) was used in the CHAMPUS model, whereas SERVICE was used directly as a likeness constraint. The other variable was a binary indicator of whether the respondent (or the respondent's family in the household, if the respondent was younger than 18) received income from sources other than those given in the binary income questions (see Chapter 9 for details). Neither likeness constraint was loosened in the 2007 survey for any of the age groups, making their inclusion in the models unnecessary.

In the model for PRVHLTIN, a four-level income variable was used as a covariate that also was used as a likeness constraint for the youngest three age groups. This likeness constraint was never loosened in the 2007 survey, making its inclusion in the models unnecessary for these three age groups. If it had been loosened, the donor and recipient would have been required to have the same value for a two-level income variable (less than $20,000 and $20,000 or more). For respondents aged 65 or older, this two-level income variable was used as an initial likeness constraint and was never loosened in the 2007 survey.

10.5 Imputation-Revised Recoded Variables for Any Other Health Insurance and Overall Health Insurance (Constituent Variables Method, Second Stage)

In the second stage of this method, a variable was created (IROTHHLT) that indicated whether respondents had any type of health insurance, even though they reported or were

imputed to have none of the four types of specific health insurance, as recorded by IRMCDCHP, IRMEDICR, IRCHMPUS, and IRPRVHLT. The final overall health insurance indicator was created by combining IRMCDCHP, IRMEDICR, IRCHMPUS, IRPRVHLT, and IROTHHLT.

10.5.1 Order of Modeling Health Insurance Variables (Constituent Variables Method, Second Stage)

Only one variable required imputation in the second stage. Therefore, an order of imputation was unnecessary.

10.5.2 Setup for Model Building (Constituent Variables Method, Second Stage)

Imputation for the any-other-health-insurance variable was conducted within the following age groups: 12 to 17, 18 to 25, and 26 or older.[135] For a respondent to be considered an item respondent for modeling the any-other-health-insurance variable, he or she first had to be part of the domain, which included respondents who had either a reported or imputed "no" value to all four imputation-revised specific health insurance variables (IRMCDCHP, IRMEDICR, IRCHMPUS, and IRPRVHLT). Among respondents who were part of the domain, item respondents had to have complete data for the variable ANYOTHER, as defined in Section 10.2.2. Response propensity adjustments were computed within each age group to make the item respondent weights representative of the entire domain. The covariates in the item response propensity model included centered age, centered age squared, gender, race/ethnicity, population density, percentage of housing in that segment that was owner-occupied, and a three-level income variable. There were also predictors that consisted of one-way interactions of centered age with race/ethnicity, centered age with gender, race/ethnicity with gender, centered age squared with race/ethnicity, and centered age squared with gender. For the two older age groups (18 to 25 and 26 or older), the additional predictors of marital status, education level, and employment status also were considered in each model.

10.5.3 Sequential Model Building (Constituent Variables Method, Second Stage)

The probability that an individual was covered by any other health insurance was modeled for item respondents within the domain defined in the previous section, within each age group, using the nonresponse-adjusted weights. The parameters were estimated using logistic regression in SUDAAN, with the same base set of predictors that were used for the specific health insurance variables. In particular, these included centered age, centered age squared, gender, race/ethnicity, population density, percentage of housing in that segment that was owner-occupied, and a three-level income variable. This base set also consisted of one-way interactions of centered age with race/ethnicity, centered age with gender, race/ethnicity with gender, centered age squared with race/ethnicity, and centered age squared with gender. For the two older age groups (18 to 25 and 26 or older), the additional predictors of marital status, education level, and employment status also were considered in each model. The following predictors were specific to the any-other-health-insurance model: household size, binary indicators of whether the respondent's family in the household received income from public assistance, wages, interest, social security, and for respondents 18 or older, a binary indicator of whether the respondent had

[135] Three age groups were used, instead of four, because of the small number of respondents who would have been included in the 65-or-older age group.

other family members in the household.[136] The complete summary of the health insurance models can be found in Appendix F.

10.5.4 Computation of Predicted Means and Univariate Predictive Mean Neighborhoods (Constituent Variables Method, Second Stage)

Following the modeling of the any-other-health-insurance variable, missing values were replaced by imputed values. In the usual way, predicted means were calculated for all respondents, including item nonrespondents, using the parameter estimates from the models. The predicted probabilities from these models were used to assign imputed values using the UPMN imputation method as described in Appendix C.

10.5.5 Assignment of Imputed Values (Constituent Variables Method, Second Stage)

Separate assignments of provisional values were performed within the three age groups. The imputed values from these assignments were considered final. The imputation-revised version of the any-other-health-insurance variable was called IROTHHLT.

10.6 Creation of the Final Overall Health Insurance Variable (Constituent Variables Method)

The final overall health insurance variable was created by combining IRMCDCHP, IRMEDICR, IRCHMPUS, IRPRVHLT, and IROTHHLT. If a respondent had a reported or imputed "yes" value for any of these five variables, the respondent was considered to have health insurance. Otherwise, he or she did not have health insurance. This was recorded using the variable IRINSUR4, which was distinguished from the overall health insurance variable that was created using the old method, IRINSUR3. Though IRINSUR4 was technically a recoded variable created from other variables, an imputation indicator was nevertheless created, called IIINSUR4. Specifically, IIINSUR4 was set to "3" if any of the five constituent health insurance variables were imputed, "2" if none of the five variables were imputed and at least one was logically assigned, and "1" otherwise.

[136] If the respondent did not have other family members in the household, the family income binary indicators listed as predictors were equivalent to the personal income binary indicators.

References

Aldworth, J., Liu, B., & Copello, E. (2007, April 16). *Simulated effect of new income questions on imputed income and health insurance variables in the 2008 NSDUH*. (Prepared for the Substance Abuse and Mental Health Services Administration, Office of Applied Studies, under Contract No. 283-2004-00022.) Research Triangle Park, NC: RTI International.

Chen, P., Dai, L., Gordek, H., Laufenberg, J., Liu, B., Sathe, N., & Westlake, M. (2009). Person-level sampling weight calibration [2007]. In *2007 National Survey on Drug Use and Health: Methodological resource book* (Section 12, prepared for the Substance Abuse and Mental Health Services Administration, Office of Applied Studies, under Contract No. 283-2004-00022, Phase I, Deliverable No. 39; RTI/0209009.374.002). Research Triangle Park, NC: RTI International.

Chromy, J. R. (1979). Sequential sample selection methods. In *Proceedings of the 1979 American Statistical Association, Survey Research Methods Section, Washington, DC* (pp. 401-406). Washington, DC: American Statistical Association. [Available as a PDF at http://www.amstat.org/sections/srms/proceedings/]

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/sections/srms/proceedings/]

Cox, B. G., & Cohen, S. B. (1985). *Methodological issues for health care surveys*. New York: Marcel Dekker, Inc.

Cox, D. R., & Snell, E. J. (1989). *The analysis of binary data* (2nd ed.). (Monographs on Statistics and Applied Probability Series). Boca Raton, FL: CRC Press.

Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association, 87*(418), 376-382.

Draper, N. R., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York: John Wiley & Sons.

Folsom, R. E., & Singh, A. C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. In *Proceedings of the 2000 Joint Statistical Meetings, American Statistical Association, Survey Research Methods Section, Indianapolis, IN* (pp. 598-603). Alexandria, VA: American Statistical Association. [Available as a PDF at http://www.amstat.org/sections/srms/proceedings/]

Folsom, R. E., & Witt, M. B. (1994). Testing a new attrition nonresponse adjustment method for SIPP. In *Proceedings of the 1994 Joint Statistical Meetings, American Statistical Association, Social Statistics Section, Toronto, Ontario, Canada* (pp. 428-433). Alexandria, VA: American Statistical Association.

Iannacchione, V. (1982). Weighted sequential hot deck imputation macros. In *Proceedings of the Seventh Annual SAS Users Group International Conference* (pp. 759-763). Cary, NC: SAS Corporation.

Kroutil, L. A. (2004). Procedures for editing interviewer-administered data in the 2002 NSDUH computer-assisted interview. In *2002 National Survey on Drug Use and Health: Methodological resource book* (prepared for the Substance Abuse and Mental Health Services Administration, Office of Applied Studies, under Contract No. 283-98-9008, Deliverable No. 28, RTI/07190). Research Triangle Park, NC: RTI International.

Kroutil, L. A., & Handley, W. (2008). 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, Office of Applied Studies, under Contract No. 283-2004-00022, Deliverable No. 39, RTI/0209009.373). Research Triangle Park, NC: RTI International.

Kroutil, L. A., Handley, W., Felts, B. J., Bradshaw, M. R., & Chien, C. (2008). 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, Office of Applied Studies, under Contract No. 283-2004-00022, Deliverable No. 39, RTI/0209009.373). Research Triangle Park, NC: RTI International.

Kroutil, L. A., & Chien, C. (2008). Procedures for editing interviewer-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, Office of Applied Studies, under Contract No. 283-2004-00022, Deliverable No. 39, RTI/0209009.373). Research Triangle Park, NC: RTI International.

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.

Morton, K. B., Martin, P. C., Hirsch, E. L., & Chromy, J. R. (2008, January). Sample design report [2007]. In *2007 National Survey on Drug Use and Health: Methodological resource book* (Section 2, prepared for the Substance Abuse and Mental Health Services Administration, Office of Applied Studies, under Contract No. 283-2004-00022, Phase III, Deliverable No. 8, RTI/0209009.330.004). Research Triangle Park, NC: RTI International. [To be available as a PDF at http://www.oas.samhsa.gov/nsduh/methods.cfm#2k7]

Office of Applied Studies. (2001). *Development of computer-assisted interviewing procedures for the National Household Survey on Drug Abuse*. (DHHS Publication No. SMA 01-3514, Methodology Series M-3). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://www.oas.samhsa.gov/nsduh/methods.cfm#Reports]

Office of Applied Studies. (2008). *Results from the 2007 National Survey on Drug Use and Health: National findings* (DHHS Publication No. SMA 08-4343, NSDUH Series H-34). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://oas.samhsa.gov/p0000016.htm]

Office of Management and Budget. (1997). Revisions to the standards for the classification of federal data on race and ethnicity. *Federal Register, 62*(210), 58781-58790. [Available at http://www.whitehouse.gov/omb/fedreg/1997standards.html]

Penne, M. A., Lessler, J. T., Bieler, G., & Caspar, R. (1998). Effects of experimental audio computer-assisted self-interviewing (ACASI) procedures on reported drug use in the NHSDA: Results from the 1997 CAI field experiment. In *Proceedings of the 1998 Joint Statistical Meetings, American Statistical Association, Social Statistics Section, Dallas, TX* (pp. 744-749). Alexandria, VA: American Statistical Association. [Available as a PDF at http://www.amstat.org/sections/srms/proceedings/]

RTI International. (2006, November). *2007 National Survey on Drug Use and Health: CAI specs for programming, English version*. (Report No. RTI 0209009, prepared under Contract No. 283-2004-00022, Deliverable No. 2). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies. [Available as a PDF at http://www.oas.samhsa.gov/nsduh/methods.cfm#2k7]

RTI International. (2007). *SUDAAN® Language manual addendum, Release 9.0.3*. Research Triangle Park, NC: RTI International. [Available at http://www.rti.org/SUDAAN/pdf_files/SUDAAN_Language_Manual_Addendum_903.pdf]

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics, 4*(1), 87-94.

Ruppenkamp, J., Frechtel, P., Aldworth, W. J., Carpenter, L., Clarke, A., Davis, T., Kroutil, L. A, & Martin, P. C. (2007). Methamphetamine analysis report [2006]. In *2006 National Survey on Drug Use and Health: Methodological resource book* (Section 15, prepared for the Substance Abuse and Mental Health Services Administration, Office of Applied Studies, under Contract No. 283-2004-00022, Deliverable No. 39, RTI/0209009). Research Triangle Park, NC: RTI International.

SAS Institute. (1999). *SAS/STAT user's guide: Version 8.* Cary, NC: SAS Institute.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data* (No. 72, Monographs on Statistics and Applied Probability). Boca Raton, FL: Chapman and Hall/CRC.

Shiffman, S., Hickcox, M., Gnys, M., Paty, J. A., & Kassel, J. D. (1995, March). *The Nicotine Dependence Syndrome Scale: Development of a new measure.* Poster presented at the annual meeting of the Society for Research on Nicotine and Tobacco, San Diego, CA.

Shiffman, S., Waters, A. J., & Hickcox, M. (2003). The Nicotine Dependence Syndrome Scale: A multi-dimensional measure of nicotine dependence. Unpublished manuscript.

Singh, A. C., & Mohl, C. A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology, 22*, 107-115.

Westlake, M., Barnett-Walker, K., Chen, P., Gordek, H., & Laufenberg, J. (2009). Questionnaire dwelling unit-level and person pair-level sampling weight calibration [2007]. In *2007 National Survey on Drug Use and Health: Methodological resource book* (Section 13, prepared for the Substance Abuse and Mental Health Services Administration, Office of Applied Studies, under Contract No. 283-2004-00022, Deliverable No. 39, RTI/0209009.375.002). Research Triangle Park, NC: RTI International.

Williams, R. L., & Chromy, J. R. (1980). SAS sample selection MACROS. In *Proceedings of the Fifth International SAS Users Group International Conference* (pp. 382-396). Cary, NC: SAS Corporation.