Summary Documentation of Selected Activities from the National Household Survey on Drug Abuse

Machine Editing
Imputation
Sampling Weight Calibration
Small Area Estimation
Table Production
Disclosure

January 28, 2003

DEPARTMENT OF HEALTH AND HUMAN SERVICES
Substance Abuse and Mental Health Services Administration
Office of Applied Studies

Table of Contents

List of Exhibits

Introduction

This document summarizes the major project operations for the 2001 National Survey on Drug Abuse (NHSDA), in the following areas: Machine editing, imputation, sampling weight calibration, small area estimation, table production, and disclosure. Topics presented include an explanation of the process involved, a breakdown of the flow of operations, and information on the technical aspects of the task. Among the technical details provided is information on the computer software used, which includes SAS®, SUDAAN®, and software created at RTI, International.

Chapter One, Machine Editing, covers the different processes that are involved in editing the data, once the cleaned, raw data file of interview records becomes available. The task consists of nine steps: Identification of usable cases, initial age edits, data diagnostics on the usable cases, edits of the core demographic variables and the employment variables that are used to create the edited job status (JBSTATR), edits of the life drug variables, edits of date-dependent variables, edits of remaining core variables, edits of non-core self-administered variables, and edits of non-core interviewer-administered variables and field interview (FI) debriefing variables. Details on each step are provided, as well as names of representative programs. Exhibits are also included, which explain the abbreviations used in the edit programs.

Chapter Two describes the next task performed for the survey, Imputation. For the 2001 survey, missing values were imputed using a procedure developed specifically for NHSDA. This procedure, called Predictive Mean Neighborhoods (PMN), is a combination of model-based imputation and random nearest neighbor hot deck. The imputation task is performed sequentially in thirteen different stages: Preparation of files for weighting and editing teams; demographics; lifetime drug use; recency, 12-month, and 30-day frequency of drug use; age at first drug use; edits of household roster; household composition summary counts (from the household roster); pair relationships (from rosters of selected pairs); multiplicity counts (from rosters of selected pairs); household counts of pair types (from the household roster); binary income variables; finer categories of income; and health insurance. Two of the stages, the preparation of files for weighting and editing teams and the edits of household roster stages, do not involve PMN. However, the remainder of stages do involve PMN, which is implemented in the following steps at each stage: Setup of the dataset required for imputation, editing or definition of variable(s) requiring imputation (where applicable), creation of indicator variables to be used as covariates, fitting statistical models for a given response variable, and assignment of imputed values to nonrespondents using a univariate predicted mean obtained from the previous step. At this point, if the imputation is univariate, quality control checks are performed; otherwise, if it is multivariate, the previous step is repeated to fit the statistical model for the next variable in the sequence of variables in the multivariate set. Then the vector of predictive means is used to assign final imputed values to nonrespondents and final quality control checks are performed. For both univariate and multivariate cases, after these checks comes the final step: the creation of a dataset and frequencies for delivering the imputation-revised variables to the master file.

The next task for the survey, Sampling Weight Calibration, involves creating three complete sets of fully adjusted analysis weights, using RTI's recently developed Generalized Exponential Modeling (GEM). The three sets represent weights at the person level, household level and person-pair level. The person-level weights are the product of 14 weight components, whereas the household-level weights are the product of 14 components, and the person-pair level reflects 16 components. Chapter Three provides information on each of these components.

Small Area Estimation (SAE) is covered in Chapter Four. SAE has the goal of producing state-by-age-group prevalence estimates that are substantially more accurate than the direct survey estimates. The major steps in the task are the creation of predictor and outcome variables, selection of significant predictor variables, and production of state by age group small area estimates. The fourth chapter presents detailed information on these steps, including lists of the predictor and outcome variables, and provides technical information on the production of small area estimates.

The next chapter outlines Table Production for the 2001 NHSDA. For this task, a series of automated processes was developed in order to produce high quality tables while minimizing error and production time, and increasing quality control. These processes fall under three major steps for the task: Preparation of data and files for table production; calculation of estimates and generation of output data into ASCII files; and the production of tables. Chapter Five supplies details on these steps and identifies sample programs.

Chapter Six covers the Disclosure task, which results in the production of the public use file (PUF). The six major steps for this task are as follows: Initial data preparation, subsampling, substitution, calibration, disclosure treatment evaluation, and final confidentiality recodes. A summary is provided for each of these steps.

1. Machine Editing

This chapter lists the different processes that are involved in editing the NHSDA data, once a cleaned, raw data file of interview records becomes available. In order to ensure the most accurate information possible about drug use, the general aims of editing the data are to identify and address inconsistent data among related variables and to replace missing data with nonmissing values. This procedure uses data within a respondent's record to identify and address inconsistencies among related variables within a given module of the interview. As part of this procedure, variables also are identified that had been legitimately skipped because the condition(s) for asking the questions did not apply. Because logical editing of the data can be programmed to be executed by computers, the term "machine editing" also can be used to describe these procedures.

The flow of the processes for editing the data is described below. In each step of the operations, the logic for editing the data was programmed in the SAS® statistical software application (Version 8.2).

Step 1.1: Identify Usable Cases

This step identifies the cases that meet or exceed the minimum requirements for completeness of interview information. The requirements for cases to be considered usable are noted below.

1. The lifetime cigarette gate question CG01 must have been answered as "yes" or "no." This requirement is set so that lifetime use or nonuse would be fully defined for at least one substance. Consequently, data about lifetime use or nonuse of cigarettes can be used in subsequent statistical imputations for other drugs where lifetime use/nonuse is undefined.

2. At least nine of the following additional gates must have answers of "yes" or "no": (a) snuff, (b) chewing tobacco, (c) cigars, (d) alcohol, (e) marijuana, (f) cocaine (in any form), (g) heroin, (h) hallucinogens, (i) inhalants, (j) pain relievers, (k) tranquilizers, (l) stimulants, and (m) sedatives. Crack cocaine is not included in the usable case rule because the logic for asking about crack cocaine is dependent upon the respondent having answered the lifetime cocaine question as "yes." Although the CAI instrument also asks about pipe tobacco, this is not included in the usable case rule because there is only one other question about pipe tobacco in addition to the gate question.

For the multiple gate drugs (i.e., hallucinogens through the sedatives), respondents are considered to have provided usable data for that drug category if at least one lead lifetime question in the series is answered as "yes" or "no" (e.g., if at least one question in the series LS01a through LS01h is answered as "yes" or "no" for hallucinogens).

In addition, the interview includes follow-up probes for respondents who initially refused to answer a gate question. The interview includes follow-up probes for the following modules that were relevant to the usable case rule: cigarettes, snuff, chewing tobacco, cigars, alcohol, marijuana, cocaine, heroin, the specific hallucinogens LSD, PCP, and ecstasy, and the specific stimulant methamphetamine. Beginning in 2001, the interview also includes follow-up probes for any use of inhalants, pain relievers, tranquilizers, stimulants, and sedatives, if respondents initially refused to answer all questions about specific drugs in these modules. If respondents change their initial refusal to a response of "yes" or "no," in response to these follow-up probes, they are considered to have provided usable data to that drug's gate information.

The starting data set for identifying usable cases according to the criteria described above is the initial, cleaned raw data file for a given quarter (or combined raw data from multiple quarters). Cases are classified as USABLE = 1 if they meet the usable case criteria and as USABLE = 0 if they do not. Cases classified as USABLE = 0 include ineligible interview cases who were confirmed to be under the age of 12 or who verified that they are currently on active military duty.

Step 1.2: Initial Age Edits

This step creates a "best available" age (BESTAGE) for use in testing out edit programs prior to the availability of the final AGE variable from the imputation team. In this step, various flags are also created to aid the imputation team in determining whether further editing of BESTAGE is warranted to produce the final AGE for each case. This step is run on all cases from the raw data, including those that will not meet the usable case criteria.

Step 1.3: Run Data "Diagnostics" on the Usable Cases

This step examines usable cases to identify cases that have patterned responses in their data that would raise questions about the validity of the interview as a whole, or the validity of data in a given "core" drug module. Thus, cases that might otherwise have met the usable case criteria might be dropped completely or else they may be retained but with selected data being wiped out (i.e., logically assigned to be "bad data"). The usable cases from Step 1 along with BESTAGE from Step 2 are used as input data in this step.

At this point, the imputation team provides age and interview date information to start the 2001 "master" file of all final interview respondents.

Step 1.4: Edit Core Demographic Variables and Employment Variables that Are Used to Create the Edited Job Status (JBSTATR)

This step uses the final AGE variable to create the edited core demographic and key employment variables. These edits do not handle the Hispanic origin and race variables; these variables are created by the imputation team.

Step 1.5: Edit Lifetime Drug Use (i.e., "gate" variables).

This step creates the edited lifetime drug use variables. Following this step, edited variables may have missing data that will subsequently be replaced with statistically imputed values. For hallucinogens, inhalants, pain relievers, tranquilizers, stimulants, and sedatives, information on "other" drugs that respondents specified using is taken into account to edit variables in these modules. In particular, data about over-the-counter (OTC) drugs that respondents specified in the Pain Relievers, Tranquilizers, Stimulants, and Sedatives modules are used to edit the lifetime data in these modules, because respondents were instructed not to report use of OTCs. However, data are not edited across modules. For example, if a tranquilizer is specified as "some other sedative," this indication of tranquilizer use in the Sedatives module is not used to edit data in the Tranquilizers section of the interview.

Step 1.6: Edit Date-Dependent Variables

This step uses imputed interview date information to edit data from cases whose interview dates were of questionable validity while their interviews were in progress. In particular, interview date information is used to set the reference period for questions about drug use in the past 12 months and past 30 days. Similarly, age-related questions (such as the age at which a respondent first used a drug) are related to the interview date and the respondent's date of birth. For these reasons, if interview dates stored by the CAI system were determined to be of questionable validity, this step sets all date-dependent variables in self-administered sections to bad data (self administered sections in 2001 were Tobacco through Youth Mental Health Service Utilization). Because date-dependent variables are set to bad data in this step (where applicable), how respondents originally answered the questions will not influence editing of variables in subsequent steps.

Step 1.7: Edit Remaining Core Variables

Once lifetime use/nonuse of a given drug has been established from Step 1.5 and after date-dependent data associated with questionable interview dates have been wiped out, the remaining variables in the Tobacco through Sedatives sections are edited. In particular, edited recency-of-use variables that establish when a respondent last used the drug of interest are created at this point. If inconsistencies are identified between the respondent's answer to a recency-of-use question and related data (e.g., if a respondent reported last using a drug more than 12 months ago but first using it at his/her current age), these inconsistencies are subsequently resolved through statistical imputation procedures.

At this point, the core edit programs are typically grouped according to drug modules with similar content. For example, respondents are asked the same basic questions in the marijuana, cocaine, crack cocaine, and heroin modules. Therefore, the edits for these drugs are handled together at this step in the editing; relevant programs that handle edits for these drugs contain the abbreviation "mch" (e.g., *recmch.sas*), where "mch" stands for "marijuana, cocaine, and heroin."

Step 1.8: Edit Noncore Self-Administered Variables

The contingent questioning approach in CAI allows respondents' answers to drug questions in core sections of the interview to determine whether respondents should be asked, or skipped out of, additional questions in noncore self-administered sections of the interview. In the Substance Dependence and Abuse section, for example, questions about dependence on, or abuse of, prescription pain relievers are relevant only for respondents who were nonmedical users of prescription pain relievers in the past 12 months. Thus, the edited pain reliever recency variable ANALREC from Step 1.7 above is a key variable for editing the prescription pain reliever variables in the Substance Dependence and Abuse section.

For the Parenting Experiences section, the logic for routing respondents into this section requires that (a) two people were selected for an interview at that selected dwelling unit (SDU); (b) a 12-17 year old was selected for an interview at that SDU; and (c) the (adult) respondent is the parent of the youth who also was selected for an interview. Consequently, the sample design variables pertaining to selection of pairs (PAIRSEL) and the pair combination at responded at that SDU (PAIRRESP) also are required to edit the Parenting Experiences data and associated Field Interviewer checkpoints that govern whether respondents are eligible for the Parenting Experiences questions.

Because the content of the noncore self-administered modules differs considerably across modules, the edit programs at this stage of the procedures are restricted to individual modules. In addition, edits for the Substance Treatment module have been further subdivided into edits of variables pertaining to receipt of treatment services (questions TX01 through TX07 and TX24 through TX44 in 2001) and variables pertaining to the perceived need for treatment services (questions TX08 through TX23 in 2001).

Step 1.9: Edit Noncore Interviewer-Administered Variables and Field Interview (FI) Debriefing Variables

Unlike the editing steps described above, editing of noncore interviewer-administered variables and FI Debriefing variables is not dependent on core drug data. Hence, these edits may technically be conducted after the final variables AGE and IRSEX become available, and after coding of any "OTHER, Specify" data has been completed for a given interviewer-administered section (e.g., other reasons specified for leaving school, from question QD24SP in 2001). Rather, these edits were labeled as "Step 1.9" to group them together based on the position of these variables toward the end of the interview.

As noted above, key noncore demographic variables pertaining to employment status are handled relatively early in the editing process (see Step 1.4). The noncore demographic section pertaining to employment and workplace issues contains additional variables that are not directly required for the job status variable JBSTATR. These additional employment variables are handled as part of these "Step 9" processes.

In addition, the Household Roster and Income portions of the noncore interviewer-administered section are not handled by the machine editing task but instead are handled totally by the imputation team. In particular, roster information from the household screening is used to edit the Household Roster variables.

2. Imputation

2.1 Introduction

Missing values are imputed using a new imputation procedure, which was developed specifically for the NSDUH when the instrument was changed from a paper-and-pencil (PAPI) format to a computer-assisted format (CAI) in 1999. The procedure, called Predictive Mean Neighborhoods (PMN), is a combination of model-assisted imputation and a random nearest neighbor hot deck, and is implemented for nearly all variables requiring imputation (a random imputation within bounds is utilized for birth date). A complete description of the PMN procedure, and its application to the person-pair data of the NSDUH, is given in Singh, Grau, and Folsom (2001).

Models incorporate nonresponse-adjusted sampling design weights, with a response propensity adjustment computed to make the item respondent weights representative of the entire sample within a given domain. The predictive means are used to define the neighborhoods, from which donors are randomly selected for the final assignment of imputed values. This assignment is either done one value at a time (UPMN) or using several response variables at once (MPMN).

Wherever necessary and feasible, additional restrictions are placed on the membership in the hot-deck neighborhoods. These constraints are 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 cannot be loosened. The latter, called "likeness constraints," can be loosened if insufficient donors are available to meet the restriction. If more than one likeness constraint is placed on a neighborhood, the restrictions are loosened in a priority order deemed appropriate for the response variable in question.

Because drug use, as well as variables related to income, insurance, and household composition, are highly correlated with age, and in order to facilitate easier implementation of the procedures, the model building and final assignments of imputed values for all drug, income, insurance, and household composition (roster-derived) variables are each done separately within distinct age groups. The drug variables are imputed within each of three age groups: 12 to 17 year olds, 18 to 25 year olds, and persons 26 years of age or older. The income, insurance, and household composition (roster-derived) variables are done within the following age groups: 12 to 17 year olds, 18 to 25 year olds, 26 to 64 year olds, and persons 65 years of age or older.

2.2 Thirteen Stages of Imputation Task

The imputation task is performed in thirteen stages, eleven of which involve an application of PMN. In general, stages earlier in the sequence need to be completed prior to subsequent stages. However, there are exceptions. For example, the roster edits only require imputed demographic variables. Whenever possible, SAS® is the software used to implement the procedures. SUDAAN®, which incorporates the sample design, is the preferred software for fitting models, rendering acceptable standard errors of the parameter estimates. The thirteen stages are listed below:

- Preparation of Files for Editing and Weighting Teams
- Demographics
- Lifetime Drug Use
- Recency of Drug Use; 12-month and 30-day Frequency of Drug Use
- Age at First Drug Use
- Edits of Household Roster
- Household Composition Summary Variables (from Household Roster)
- Pair Relationships (from Rosters of Selected Pairs)
- Multiplicity Counts (from Rosters of Selected Pairs)
- Household Counts (from Household Roster)
- Binary Income Variables
- Finer Categories of Income
- Health Insurance

2.3 Stages Not Involving PMN

In the "Preparation of Files" stage, household- and person-level files are prepared for the weighting team; these files include information from the screener and census data for each segment. Age, interview date, birth date, and sex are also created at this stage, for use by both the editing and the imputation teams; here missing values are replaced by randomly generated replacement values. At the "Edits of Household Roster" stage, the household roster is edited to remove nonsensical relationship, age, and sex indicators at the roster level. Missing values for roster members are not imputed. Neither of these stages involves an application of PMN.

2.4 Stages Involving PMN: Implementation Steps

What follows is a step-by-step overview of the tasks required for the PMN imputation method that is applied in each of the remaining 11 stages. In summary, the steps listed below describe the process that prepares files for imputation, performs the actual imputation, and creates the final deliverable data set. The number of programs required for each step varies according to the imputation stage listed above.

Step 2.4.1: Data Setup

Set up the data set required for imputation. This generally involves subsetting the master data file. It may also involve merging in other files such as the raw data file or an "Other-specify" data file.

Step 2.4.2: Editing or Definition of Variable Requiring Imputation (where applicable)

Before imputation can occur on a variable with missing values, the variable must be edited (if it corresponds directly to a question on the questionnaire) or defined (if it does not correspond directly to a question). This step is usually performed by the editing team. Otherwise, the variable is edited or defined by the imputation team.

Step 2.4.3: Create Indicator Variables and Adjust Respondent Weights

Create indicator variables for those variables that will be used as covariates in models in this, and subsequent, steps. A response propensity adjustment is calculated and applied to the analysis weight for the domain in question (i.e. for "recency", the domain is "all lifetime users") using the Generalized Exponential Modeling (GEM) software macro, which was developed at RTI.

Step 2.4.4: Fit a Statistical Model

For a given response, fit a statistical model. The type of model depends upon the distribution of the response variable. Models include least squares regression models with an appropriately transformed response, binary and multinomial logistic models, poisson regression models, and interval-censored failure time models. From these models, a predicted mean is obtained and saved in a dataset. The covariates used in these models were created in the previous step.

Step 2.4.5: Assign Imputed Values to Nonrespondents (Univariate)

Assign imputed values to nonrespondents using a univariate predicted mean obtained from the statistical model in the previous step. If a live donor is used, this donor is randomly selected from a neighborhood defined by the predicted mean. If the final imputation is univariate, this imputation is considered final. However, in a multivariate imputation, the imputation is provisional, so that a response variable earlier in a sequence of multiple variables can be used as a covariate for a response variable later in the sequence. For instance, a provisionally imputed value for cigarette recency can be used in the cigarette "30-day frequency of use" model. Use logical constraints to ensure that imputed values are consistent with preexisting nonmissing values. Use likeness constraints to ensure that donors and recipients are as alike as possible.

If the final imputation is univariate, create an indicator variable that distinguishes imputed from non-imputed values, and proceed to step 2.4.7. Otherwise, return to Step 2.4.4, to fit the statistical model for the next variable in the sequence of variables.

Step 2.4.6: Assign Imputed Values to Item Nonrespondents (Multivariate)

If the final imputation is multivariate, use the vector of predictive means to define a neighborhood, from which a donor is randomly selected. Use logical constraints to ensure that imputed values are consistent with preexisting nonmissing values. Use likeness constraints to ensure that donors and recipients are as alike as possible. Create an indicator variable that distinguishes imputed from non-imputed values.

Step 2.4.7: Perform Final Quality Control Checks

The specific quality control checks will depend upon the variable being imputed. In general, however, the final quality control checks are implemented with two goals: (1) ensure that the imputed values are consistent with pre-existing non-missing values, and (2) ensure that the imputation procedures worked correctly. Imputed values that are shown to be outside an acceptable range are flagged to determine the source of the inconsistency. A quality control check common to all variables is a comparison of the distribution of imputed values with the distribution of the variable requiring imputation among complete cases.

Step 2.4.8: Deliver the Imputation-Revised Variables to the Master File

Create a dataset and frequencies for delivering the imputation-revised variables to the master file. In some cases (income, health insurance, roster, and the pair work) edited variables are also delivered.

2.5 References

Singh, A.C., Grau, E.A., and Folsom, R.E. (2001). Predictive mean neighborhood imputation with application to person pair data of NHSDA drug use variables. *Proceedings of the Section on Survey Research Methods of the American Statistical Association.*

3. Sampling Weight Calibration

For the National Survey on Drug Use and Health, three sets of fully adjusted analysis weights are created using RTI's recently developed Generalized Exponential Modeling (GEM) methodology, which allows for unit-specific bounds on adjustment factors (and thus has a built-in control on extreme weights) and provides a unified approach to adjustments for nonresponse, poststratification, and extreme values. (The GEM software is written in SAS® Macro Language, version 8.2). In GEM, a final extreme value adjustment is performed after poststratification, if necessary, to control for possible extreme weights that may be present after poststratification. The extreme-value adjustment is basically a repeat poststratification that provides tighter control on extreme weights while preserving calibration controls. For additional information on GEM, see Folsom and Singh (2000). The three sets represent weights at the person level, household level and person-pair level. These three analysis weights share the same first nine weight components at the screening dwelling-unit level. In addition to these common components, each of the analysis weights has some additional level-specific weight components.

3.1 Person Level Weight

Person-level analysis weights are the product of 14 weight components, each representing either a probability of selection at a particular stage, or some form of nonresponse adjustment, poststratification adjustment or extreme-value adjustment.

Phase One (Screening Dwelling-Unit Level)

Step 1: Design-Based Weights

In this step, six weight components are created. They reflect the selection probability at two stages, one for selecting the segment and the other for selecting a dwelling unit from the segment; in addition, they reflect inflation factors for sub-segmentation, added dwelling units and percent released to the field interview regions. The six weight components are:

#1. Inverse of probability of selecting segment
#2. Quarter segment factor
#3. Subsegmentation inflation factor
#4. Inverse of probability of selecting screening dwelling unit
#5. Added screening dwelling unit factor
#6. Screening dwelling unit percent release factor

Step 2: Weight Adjustments

In this step, adjustments for nonresponse, poststratification, and extreme values at the screening dwelling-unit level are implemented. The nonresponse adjustment accounts for the failure to obtain screening interviews from eligible dwelling units. The poststratification adjustment adjusts the person-level counts obtained from the screener dwelling units to census controls. The extreme-value adjustment is a repeat poststratification. If it is not needed due to a low percentage of extreme weights (as in the case of 2001), a value of one is assigned to all responding screening dwelling units for weight component #9. The three weight components are:

#7. Screening dwelling unit nonresponse adjustment
#8. Screening dwelling unit poststratification adjustment
#9. Screening dwelling unit extreme-value adjustment

Phase Two (Person Level)

Step 1: Design-Based Weight

In this step, one weight component is created. It reflects the probability of selecting a person from the selected screening dwelling unit. The weight component is:

#10. Inverse of probability of selecting person from screening dwelling units

Step 2: Weight Adjustments

In this step, four adjustments are implemented. The first adjustment is the selected person poststratification, in which the input weights (product of #1-#10) are poststratified to the population controls based on the weights of all screened persons (product of #1-#9). This poststratification adjustment is somewhat innovative. It takes advantage of the two-phase nature of the design, since the screener data provide a large sample containing information about demographic variables for the screening dwelling units. The respondent nonresponse adjustment accounts for the failure to obtain respondents at the person level. The respondent poststratification forces the sample estimates, based on respondent person weights, to equal specified control totals obtained from the Census Bureau's population estimates of the civilian non-institutional population aged 12 and older. The extreme-value adjustment, as before, is a repeat poststratification and used only if needed. (For 2001, it was not needed since the percentage of extreme weights was not high after the poststratification adjustment, and the value of one is assigned to all responding persons for weight component #14.) The four weight components are:

#11. Selected person poststratification adjustment
#12. Respondent person nonresponse adjustment
#13. Respondent person poststratification adjustment
#14. Respondent person extreme-value adjustment

3.2 Household-Level Weight

Household-level analysis weights are the product of 14 weight components; among them, the first nine components at the screener dwelling-unit level are the same as for the person-level weights. The remaining five components, which are specific for the household-level weights, represent either a selection probability at a particular stage or some form of nonresponse adjustment, poststratification adjustment, or extreme-value adjustment. Thus, phase one is skipped for this section.

Phase Two (Household Level)

Step 1: Design-Based Weight

In this step, one weight component is created. It reflects the probability of selecting at least one person in the screened dwelling unit. The weight component is:

#10. Inverse of probability of selecting at least one person in the screened dwelling unit

Step 2: Weight Adjustments

In this step, four adjustments are implemented. The first adjustment is the selected household poststratification, in which the initial weights (product of #1-#10) are poststratified to the population controls based on the weights of all the completed screening dwelling units (product of #1-#9). The second adjustment, the respondent nonresponse adjustment, accounts for the failure to obtain a completed interview at the household level. The third, the respondent poststratification, forces the sample estimates based on respondent household weights to equal the population controls based on the poststratified screening dwelling-unit weights (product of #1-#9). The extreme-value adjustment is not performed unless the extreme-weight percentage is high after the poststratification adjustment; if it is performed, the value of one is assigned to all responding persons for weight component #14. The four weight components are:

#11. Selected household poststratification adjustment
#12. Respondent household nonresponse adjustment
#13. Respondent household poststratification adjustment
#14. Respondent household extreme-value adjustment

3.3 Person-Pair-Level Weight

The person-pair level analysis weights are the product of 16 weight components; among them the first nine components at the screener dwelling-unit level are the same as for the person-level weights. The remaining seven components that are specific for the person-pair level weights represent either a probability of selection at a particular stage or some form of nonresponse adjustment, poststratification adjustment, or extreme-value adjustment. Thus, phase one is skipped for this section.

Phase Two (Person-Pair Level)

Step 1: Design-Based Weight

In this step, one weight component is created. It reflects the selection probability of a person-pair in the screened dwelling unit. The weight component is:

#10. Inverse of probability of selecting a person-pair in the screened dwelling unit

Step 2: Weight Adjustments

In this step, six adjustments are implemented. Due to the variability introduced by the selection probability of a person-pair, the proportion of extreme weights is generally high. The built-in control for extreme values in the generalized exponential modeling is not sufficient; therefore, before the selected person-pair poststratification adjustment, two extreme-value adjustments are added. One is winsorization and the other is an adjustment using GEM. The third adjustment is the selected person-pair poststratification, in which the initial weights (product of #1-#12) are poststratified to the population controls based on the weights of all screener person-pairs (product of #1-#9). The respondent non-response adjustment accounts for the failure to obtain complete response at the person-pair level. The respondent poststratification forces the sample estimates based on respondent household weights to equal the population controls based on the poststratified screening dwelling-unit weights (product of #1-#9) from all screener person-pairs. The extreme-value adjustment for the respondent person-pairs is necessary due to the high proportion of extreme weights after the poststratification. The six weight components are:

#11. Selected person-pair extreme-weight trimming
#12. Selected person-pair extreme-value adjustment
#13. Selected household poststratification adjustment
#14. Respondent person-pair nonresponse adjustment
#15. Respondent person-pair poststratification adjustment
#16. Respondent person-pair extreme-value adjustment

3.4 References

Folsom, R.E. Jr., and Singh A.C. (2000). A generalized exponential model for sampling weight calibration for a unified approach to nonresponse, poststratification and extreme weight adjustments. *Proceedings of the Section on Survey Research Methods of the American Statistical Association,* 598-603.

4. Small Area Estimation (SAE)

The goal of the NHSDA Small Area Estimation task is to produce state by age group prevalence estimates that are substantially more accurate than the direct survey estimates. The task involves the following steps:

- Creation of Predictor and Outcome Variables
- Selection of Significant Predictor Variables
- Estimation of Survey Weighted Hierarchical Bayes Model Parameters
- Production of State by Age Group Small Area Estimates

These steps are described in the following paragraphs; in addition, they are represented schematically in Exhibit 4.1, which presents a flowchart of the small area estimation process.

Step 4.1: Create Predictor and Outcome Variables

Compile Predictor Variables from Various Sources

The continuous predictor variables are compiled on one of the following three levels: census block group, tract, or county. In addition to these variables, there are some indicator variables that are also used in the National Survey on Drug Use and Health (NHSDA) - Small Area Estimation (SAE) modeling. Details about the definitions and sources of all independent variables are given in the paragraphs that follow.

There are four person-level, thirteen block group-level, forty-four tract-level, forty-two county-level, and six state-level predictor variables used in NHSDA - SAE modeling. These predictor variables are obtained from numerous sources. The complete list of predictors along with their sources is given in Exhibit 4.1. Every year, the predictor variables are updated whenever possible. Information about the data sources is given below.

Claritas, Inc. All of the block group level variables are created using data from Claritas, Inc. The data is obtained on CDs in the form of ASCII flat files. The CDs contain data for 1999 and projections for 2004. The data is read using SAS® and linear interpolation was used to estimate data for 2000 and subsequent years. These block group-level variables are demographic variables that describe a block group's percentage of each race/ethnicity group, age group, and gender group; e.g., percentage of 0-18 in a block group. Their tract and county-level versions are also included in SAE modeling. The two other tract level variables (PASIAN, PINDIAN) are also created from the data on the CDs.

Census Bureau. Most of the tract-level variables are created using 1990 Census data. These are socioeconomic variables such as the percentage of families below the poverty level in a tract. The most current county-level food stamp participation rate data is obtained from Mr. William Bell of the US Census Bureau. The 1998 food stamp participation rate data was used in 2001 NHSDA SAE modeling. Some changes were made to some of the county codes to match county codes found in other data sources.

National Center for Health Statistics. For the 2000, 2001, and 2000-2001 pooled SAE modeling the mortality data from1993-1998 was used. The data based on ICD-9 death rates was obtained from the National Center for Health Statistics.

Area Resource File (ARF). For the 2000 and 1999-2000 pooled SAE modeling, the 2000 release of data from Bureau of Health Professions, Office of Research and Planning was used. For 2000-2001 pooled SAE modeling, the 2001 release of ARF data was used.

Uniform Crime Reports (UCR). UCR arrest totals that are available for download from: http://fisher.lib.Virginia.EDU/crime are used. For the 2000, 2001, and 2000-2001 SAE modeling, the 1998 data was used. For some counties the 1998 data was not available; in those cases, the most current available data was used.

Uniform Facility Data Set (UFDS) For the 1999, 2000, and 1999-2000 pooled SAE modeling analysis, the 1997 and 1998 UFDS data on drug and alcohol treatment rates was used. The data was obtained from Synectics for Management Decisions, Inc. For 2000 and 2000-2001 pooled SAE modeling, the 2000 UFDS (now called NSSATS) data was used.

National Survey on Drug Use and Health. On the person level, there are four variables created from the NHSDA sample to indicate the four levels of race/ethnicity and two levels of gender.

Exhibit 4.1 List of Predictor Variables Used in SAE Modeling

Variable Prefix*	Continuous Variable	Label	Source	Level
1. bp18	bpct18	% 0-18 in Block group	Claritas	Block Group
2. bp1924	bpct1924	% 19-24 in Block group	Claritas	Block Group
3. bp2534	bpct2534	% 25-34 in Block group	Claritas	Block Group
4. bp3544	bpct3544	% 35-44 in Block group	Claritas	Block Group
5. bp4554	bpct4554	% 45-54 in Block group	Claritas	Block Group
6. bp5564	bpct5564	% 55-64 in Block group	Claritas	Block Group
7. bp65	bpct65	% 65 and older in Block group	Claritas	Block Group
8. bpblk	bpctblk	% Blacks in Block group	Claritas	Block Group
9. bphis	bpcthis	% Hispanics in Block group	Claritas	Block Group
10. bpmal	bpctmale	% Males in Block group	Claritas	Block Group
11. bpfem	bpctfem	% Females in Block group	Claritas	Block Group
12. bpoth	bpctoth	% Other Race in Block group	Claritas	Block Group
13. bpwht	bpctwht	% Whites in Block group	Claritas	Tract
15. tp1924	tpct1924	% 19-24 in Tract	Claritas	Tract
16. tp2534	tpct2534	% 25-34 in Tract	Claritas	Tract
17. tp3544	tpct3544	% 35-44 in Tract	Claritas	Tract
18. tp4554	tpct4554	% 45-54 in Tract	Claritas	Tract
19. tp5564	tpct5564	% 55-64 in Tract	Claritas	Tract
20. tp65	tpct65	% 65 and older in Tract	Claritas	Tract
21. tpblk	tpctblk	% Blacks in Tract	Claritas	Tract
22. tphis	tpcthis	% Hispanics in Tract	Claritas	Tract
23. tpmal	tpctmale	% Males in Tract	Claritas	Tract
24. tpfem	tpctfem	% Females in Tract	Claritas	Tract
25. tpoth	tpctoth	% Other Race in Tract	Claritas	Tract
26. tpwht	tpctwht	% Whites in Tract	Claritas	Tract
27. hsdrop	hsdrop9	% High school dropouts	1990 Census	Tract
28. p40hu	p40hu	% Housing Units built 1940-1949	1990 Census	Tract
29.p64dis	p64dis	% Persons 16-64 with a work disability	1990 Census	Tract
30. pcuban**	pcuban	% Hispanics: Cuban	1990 Census	Tract
31. pflab	pflab	% Females >=16 years old, in labor force	1990 Census	Tract
32. pfnev	pfnev	% Females never married	1990 Census	Tract
33. pfnot	pfnot	% Females separated/divorced/widowed	1990 Census	Tract
34. phh1p	phh1p	% One person households	1990 Census	Tract
35. phhf18	phhf18	% Female headhsehld, no spouse, child<18	1990 Census	Tract
36. pindia	pindian	% American Indian, Eskimo, Aleut in tract	Claritas	Tract
37. pmlab	pmlab	% Males >=16 years old, in labor force	1990 Census	Tract

38. pmnev	pmnev	% Males never married	1990 Census	Tract
39. pmnot	pmnot	% Males separated/divorced/widowed	1990 Census	Tract
40. poldhu	poldhu	% Housing Units built 1939 or earlier	1990 Census	Tract
41. poprm	poprm	Average persons per room	1990 Census	Tract
42. ppover	ppover	% Families below poverty level-tract	1990 Census	Tract
43. ppubas	ppubass	% Households w/ public assistance income	1990 Census	Tract
44. prent	prented	% Housing units rented	1990 Census	Tract
45. psch12	psch12	% 9-12 years and no high school diploma	1990 Census	Tract
46. psch8	psch8	% 0-8 years of school	1990 Census	Tract
47. pschas	pschas	% Associates degree	1990 Census	Tract
48. pschsc	pschsc	% Some college and no degree	1990 Census	Tract
49. pschco	pschco	% Bachelors, Grad, Professional Degree	1990 Census	Tract
50. rh43a	rh43a	Median rents for rental units	1990 Census	Tract
51. rh61a	rh61a	Median Value Owner occ Housing Units	1990 Census	Tract
52. rp80a	rp80a	Median household income	1990 Census	Tract
53. pasian	pasian	% Asian, Pacific Islander in tract	Claritas	Tract
54. adhra0	adhra0	Alcohol death rate, direct cause	ICD-9	County
55. adhra1	adhra1	Alcohol death rate, indirect cause	ICD-9	County
56. arate	arate	Alcohol treatment rate	UFDS	County
57. brate	brate	Alcohol & Drug treatment rate	UFDS	County
58. cdhra0	cdhra0	Cigarettes death rate, direct cause	ICD-9	County
59. cdhra1	cdhra1	Cigarettes death rate, indirect cause	ICD-9	County
60. cp18	cpct18	% 0-18 in county	Claritas	County
61. cp1924	cpct1924	% 19-24 in county	Claritas	County
62. cp2534	cpct2534	% 25-34 in county	Claritas	County
63. cp3544	cpct3544	% 35-44 in county	Claritas	County
64. cp4554	cpct4554	% 45-54 in county	Claritas	County
65. cp5564	cpct5564	% 55-64 in county	Claritas	County
66. cp65	cpct65	% 65 and older in county	Claritas	County
67. cpblk	cpctblk	% Blacks in county	Claritas	County
68. cphis	cpcthis	% Hispanics in county	Claritas	County
69. cpmal	cpctmale	% Males in county	Claritas	County
70. cpfem	cpctfem	% Females in county	Claritas	County
71. cpoth	cpctoth	% Other race in county	Claritas	County
72. cpwht	cpctwht	% Whites in county	Claritas	County
73. ddhra0	ddhra0	Drugs death rate, direct cause	ICD-9	County
74. ddhra1	ddhra1	Drugs death rate, indirect cause	ICD-9	County
75. drate	drate	Drug treatment rate	UFDS	County
76. drgpos	drgposrt	Drug Possession rate	UCR	County
77.drgsal	drgsalrt	Drug Sale rate	UCR	County
78. drgvio	drgviort	Drug Violation Rate	UCR	County
79. fabpov	fabpov	Families below poverty level-cnty	ARF	County
80. fspart	fspart	Food stamp participation rate	Census Bureau	County
81. mjpos	mjposrt	Marijuana possession rate	UCR	County
82. mjsal	mjsalrt	Marijuana sale/manufacture rate	UCR	County
83. ocpos	ocposrt	Opium cocaine possession rate	UCR	County
84. opcoc	opcocrt	Opium cocaine sale/manufacture rate	UCR	County
85. otdrps	otdrpsrt	Other drug possession	UCR	County
86. othdrg**	othdrgrt	Other: Dangerous non narcotics	UCR	County
87. sercr	sercrrt	Serious crime rate	UCR	County
88. unemp	unemp	Unemployment rate, county	ARF	County
89. viocr	viocrrt	Violent Crime rate	UCR	County
90. income	income	Per capita income (in 1000s)	ARF	County

* The categorical variable names created from the deciles and the polynomial coefficient names can be formed by appending “ca” and “p1”, “p2”, and “p3” to the variable prefixes.

** There are two additional variables formed from the prefixes pcuban and othdrg with appended ‘0’s. These indicator variables denote areas that have no occurrence of Cubans in a tract, and no arrest for other dangerous non-narcotics, respectively. The variables pcubanca and othdrgca do not have all ten categories, and pcubanca only has two polynomials.

Categorical/Indicator Variables:

Variable		Label	Source	Level
1.	race1ind	=1 if Hispanic, =0 otherwise	Sample	Person
2.	race2ind	=1 if non-Hispanic Black, =0 otherwise	Sample	Person
3.	race3ind	=1 if non-Hispanic Other, =0 otherwise	Sample	Person
4.	male	=1 if male, =0 if female	Sample	Person
6.	reg2	=1 if Midwest region, =0 otherwise	1990 Census	State
7.	reg3	=1 if South region, =0 otherwise	1990 Census	State
8.	pd1	=1 if msa with 1 million + people, =0 otherwise	1990 Census	County
9.	pd2	=1 if msa with less than 1 million people =0 otherwise	1990 Census	Count
10.	pd3	=1 if non-msa urban, =0 otherwise	1990 Census	Tract
11.	uclass9	Underclass indicator	Urban Institute	Tract
12.	pcuban0	=1 if no Cubans in tract, =0 otherwise	1990 Census	Tract
13.	purbp	=1 if urban area, =0 if rural area	1990 Census	Tract
14.	othdrg0	=1 if no arrests for dangerous non-narcotics, =0 otherwise	UCR	County
15.	catage***	=1 if age 12-17, =2 if age 18-25, =3 if age 26-34, =4 if age 35+	Sample	Person

*** This variable is used for grouping. It is not an explanatory variable.

Variables used only for SMI (serious mental illness) in conjunction with other variables:

Variable Prefix*	Continuous Variable	Label	Source	Level
Dui	duirt	Driving under influence arrest rate	UCR	County
Sui	suirate	Avg suicide rate(1996-1998, per 10000)	ARF	County

* The categorical variable names created from the deciles and the polynomial coefficient names can be formed by appending "ca" and "p1", "p2", and "p3" to the variable prefixes.

Variables Used only for TXNOSPEC (treatment gap) in conjunction with other variables:

Continuous Variable	Label	Source	Level
Ami	Total SAPT Block Grant Application Index	SAMHSA	State
Ci_01_03	2001-2003 Cost of Services Factor Index	SAMHSA	State
Ttr	Total Taxable Resources Per Capita Index for 1998	SAMHSA	State

Create Outcome Variables

Typically state-by-age-group level SAE estimates are produced for 18-20 binary outcome variables. Every year a few outcome variables will be replaced by others that might have been introduced as "new" variables in the survey year. Most of the outcome variables can be picked up from the Analysis master file, and some need to be created using algorithms suggested and verified with SAMHSA. In 1999, a single year sample file was produced for SAE modeling; whereas in 2000, a pooled analysis file using the 1999 and 2000 NHSDA sample was used to increase the precision of state estimates. In 2000 a new variable called "Treatment Gap" was introduced and in 2001 the "Serious Mental Illness" outcome was introduced. For the 2001 SAE exercise, a pooled analysis file combining the 2000 and 2001 NHSDA data was created.

Exhibit 4.2 List of Outcome Variables for 2000-2001 NHSDA-SAE Modeling

ABODALC	= past year alcohol dependence or abuse
ABODILAL	= past year dependence or abuse of any illicit drug or alcohol
ABODILL	= past year any illicit drug dependence or abuse
ALCMON	= past month use of alcohol
BENGAL	= past month 'binge' alcohol use
CIGMON	= past month use of cigarettes
COCYR	= past year use of cocaine
DEPNDALC	= past year alcohol dependence
DEPNDILL	= past year any illicit drug dependence
MJUSENV	= never used marijuana**
MRJ24	= used marijuana in the past 24 months**
MRJMON	= past month marijuana use
IEMMON	= past month use of any illicit drug except marijuana
RISKALC	= perceptions of great risk of having 5 or more drinks of an alcoholic beverage once or twice a week
RISKCIG	= perceptions of great risk of smoking one or more packs of cigarettes per day
RISKMJ	= perceptions of great risk of smoking marijuana once a month
SUMMON	= past month use of any illicit drug
TOBMON	= past month use of any tobacco product
TXNOSPEC	= treatment gap
SMI	= serious mental illness (only using 2001 sample)

**These two variables are used in to produce estimates for Avg. annual incidence of marijuana which is defined as

Avg. annual incidence=0.5*Mrj24/(0.5*Mrj24 + Mjusenv).

Also, note that RISKMF, RISKCIG and RISKALC have missing values. For these variables, sample weights are adjusted to match appropriate population totals.

Link Predictors to the NHSDA Sample

The county-level predictors are merged to the NHSDA sample using county codes and the tract level predictor variables are merged to the sample using "majority tract." A majority tract on the sample is defined as the census tract containing the largest number of dwelling units in that area segment. The NHSDA sample does not have block group identifiers, it has segment identifiers. A segment is defined as a small cluster of census blocks containing dwelling units. A segment may contain dwelling units from several block groups. Therefore, a single block group-level Claritas variable cannot be directly linked to the NHSDA sample segments. Instead, a weighted average of the block group variables is computed and this weighted average links to each segment and merges the averages to the sample file. Here the block group weight is the ratio of the number of segment dwelling units in that block group to the total number of dwelling units in the segment.

Create Deciles and Orthogonal Polynomial Coefficients

When all possible predictor variables have been updated and linked to the NHSDA sample, the predictor variables are categorized into sample deciles. These categorical variables are used to create linear, quadratic and cubic orthogonal polynomials. In 1999, the sample deciles were created using PROC UNIVARIATE in SAS®. For the 2000 and 2001 NHSDA SAE modeling, the 1999 decile's cutoffs were used to categorize the updated predictor variables. In some cases, due to many ties (especially in the first category where there could be a lot of zeros), the categorized predictor variables did not have a uniform distribution. Such variables required special dummy variables indicating a high occurrence of tied values in the first category. Typical examples of this are the variables PCUBAN and OTHDRGRT which had a high occurrence of zero values so variables PCUBAN0 and OTHDRG0 were formed to indicate tracts/counties with these zero values.

Create the Universe of Predictor Variables

There are approximately 226,000 block groups in the United States. The collection of all predictor variables defined at the block group level is henceforth called a Universe file. The categorized predictor variables and corresponding orthogonal polynomials are created using the same deciles that were calculated on the sample. In addition to the predictor variables, 32 additional population count variables are created. These variables are the population counts in each of 32 cells (four race/ethnicity categories, two gender levels and four age-group categories). The 32-cell population counts are created using Claritas population projections.

The categories white, black, and other in the Claritas four-way table were not the same as the categories white-non-Hispanic, black-non-Hispanic, and other-non-Hispanic which were defined on the NHSDA sample. This caused the Hispanics to be double counted. Using the 1990 Census, the Claritas projections were ratio adjusted to remove the Hispanics from those race groups.

In addition to this block group-level adjustment to the Claritas population projections, the state-level 32-cell population counts aggregated from the universe file were scaled so that these scaled population counts matched the corresponding NHSDA sample weights totals from each state. However, in a few cases, some of the state level cells were empty in the NHSDA sample, whereas corresponding cell counts on the universe file had nonzero population counts. When this happened, the original count was not adjusted on the universe file. Therefore, the state-level population totals from the universe file did not exactly match the corresponding weight totals from the sample; but they were very close to each other.

Combine Predictor and Outcome Variables

Once all the predictor variables and outcome variables are created they are merged together to create the sample analysis files. The four age group specific sample analysis files are also created for selection of significant fixed effect predictors for SAE modeling. Step 4.2 describes the variable-selection methodology for SAE modeling.

Step 4.2: Select Significant Predictor Variables

To select fixed effect predictor variables for the SAE modeling, a multi-step variable selection methodology is adopted by combining outputs from different variable selection software such as SAS®, SUDAAN® and the Chi-squared Automatic Interaction Detection (CHAID) algorithm. The following paragraphs describe this methodology.

CHAID. To detect interactions between independent variables and select main effects, an Exhaustive CHAID algorithm is used, which is found in the software package AnswerTree® 2.1 by SPSS®. For every outcome variable four age-group-specific CHAID trees are created. All the categorized predictor variables are entered into the CHAID algorithm. The level $\alpha_{split} = .03$ is used, after also growing and examining the trees for $\alpha_{split} = .05$ and $\alpha_{split} = .01$. If the p-value is less than or equal to α_{split}, then it splits the node based on the set of categories of the predictor variable. If the p-value is greater than α_{split}, then it does not split the node and that node is terminal. This tree-growing process continues until one of the stopping rules is met, in which case, the node will not be split and it becomes a terminal node. The stopping rules are given in the AnswerTree® manual. The maximum depth of the tree is set to 80; the minimum number of cases in the parent node is set to 1,000 and the minimum number of cases in a child node is set to 300. All of the decilized predictor variables are input as ordinal variables, and the region, population density, and urban/rural indicator, and underclass indicator independent variables are input as nominal variables. Then the CHAID trees are generated, and the terminal nodes obtained from such trees are of interest. It is possible for a terminal node to be a pure node. That is, all the cases of the dependent variable in that node are zero or all the cases are one. On the rare occasions when pure nodes are encountered, they are appropriately merged with other nodes.

After completing the trees for all of the outcome variables, the next step is to import the AnswerTree® decision rules into our SAS® procedures to create (0, 1) indicator variables indicating terminal nodes in CHAID trees. The significant main effects variables are also extracted from these CHAID trees. The indicator variables indicating terminal nodes; linear, quadratic and cubic orthogonal polynomials corresponding to the significant main effects from the CHAID trees; and interactions of linear orthogonal polynomials with race and gender are used in the subsequent steps of the variable selection methodology.

SAS Stepwise Logistic Regression. A SAS® Stepwise Logistic regression model is also developed independently of CHAID, using all the available predictor variables, for each age-group and for every outcome variable. Note that for continuous predictor variables, only their first order orthogonal polynomial is used in the stepwise models. The predictor variables with a SLE=5% and a SLS=3% are allowed to stay in the model. The significant predictor variables are then extracted from the SAS® logistic regression models. If a linear orthogonal polynomial is selected, then the corresponding quadratic and cubic orthogonal polynomials are also extracted, and the interaction of the linear orthogonal polynomial with race, gender and region are also extracted. Then this list of predictors is combined with the one created using CHAID. Any duplicates are deleted from the list.

SAS®/SUDAAN® Logistic Regression. Next, the significant predictor variables selected from the previous steps are entered into a SAS® stepwise logistic model at the one percent significance level. The 1%-significant variables are then entered into a SUDAAN® logistic regression model. All predictor variables that are still significant at 1% are used as fixed effects in the survey-weighted hierarchical Bayes SAE models.

For outcome variables modeled in 2000, the starting predictor set was the final predictor set used in the 1999 analyses. This set was further reduced by modeling the 2000 data using SUDAAN® selection at the 1% level of significance. This is the final set of predictors used in all models after 2000. Note that race and gender are forced in all the models. In the past, region was forced as well, but region, and interactions of region with race and gender, were removed from all models after 2000.

Step 4.3: Produce State by Age Group Small Area Estimates

The following paragraphs briefly describe the methodology used for estimating mixed logistic regression model parameters and production of SAEs for 50 states and the District of Columbia.

Estimate Model Parameters Using Survey Weighted Hierarchical Bayes Methodology

Mixed logistic models are fitted using a survey weighted hierarchical Bayes (SWHB) methodology (see Folsom, Shah, and Vaish, 1999). The estimation of model parameters is not straightforward. A series of iterative steps is employed to generate posterior sample values of the desired fixed and random effect parameters from their underlying joint posterior distribution. PROC GIBBS software is used for this purpose. This software was developed by RTI specifically for fitting SAE models to NHSDA data. It uses a Markov Chain Monte Carlo algorithm to generate samples from the posterior distribution of the fixed effects, random effects, and the associated variance-covariance matrices. PROC GIBBS generally takes about 4-16 hours (on a one GHz PC with at least 512 MB RAM) to generate 10,000 replicate samples for each of the models vector of parameters. Every eighth replicate is selected, yielding a total of 1,250 independent samples from the joint posterior distribution of the parameters. The selected set of 1,250 parameter vectors is then used to produce predicted prevalence for every block group on the universe file. These block group-level predicted values are formed and aggregated to the state level using an RTI-developed procedure called PROC GSTAT.

To validate the convergence of the variance component chains, W1 and W2 chains, that are produced by PROC GIBBS, the Raftery and Lewis test available in Convergence Diagnostic Software (called CODA) is used. For this purpose, the log of the determinant of W1 and W2 is used to summarize these 4X4 matrices. The CODA software is written using SPLUS® and is available for download from the internet. For the Raftery and Lewis test, the default parameters are set as: quintile = 0.025, accuracy = +/- 0.0075 and the desired probability = 0.85. If the Raftery and Lewis calculation of the required chain length is less than 1,250, then convergence is confirmed. After validating the convergence of the W1 and W2 chains, the next step is to produce SAEs for the 50 states and the District of Columbia.

Produce Small Area Estimates

The universe file contains variables that are defined at the Census block group, tract, county, and state levels. There are approximately 226,000 data values in the universe file corresponding to each of the block groups in the United States. For each of the block groups, the universe file also contains population projections for each of 32 demographic cells (4 age groups x 2 gender groups x 4 race groups). Due to the huge size of the universe file, creating actual person-level records for the 32 demographic cells is not possible. Instead, the aggregation software, GSTAT, creates 32 virtual persons in the computer's memory corresponding to the 32 demographic cells in each of the block groups.

Separate vectors of fixed predictors are identified for each of four age groups (12-17, 18-25, 26-34, 35+). There are two types of random effects used in the model; namely, random effects for states, and for groups of three field interviewer regions[1]. The state-level random effects (η) and FI region group random effects (υ) are assumed to be four variate normal random vectors; that is, assume that the four age group specific random effects in η and υ are correlated with general variance-covariance matrices W_1 and W_2 respectively. It is also assumed that the random vectors η and υ are independently distributed. The paragraphs below describe the method used to form the predicted values on the Universe file and obtain the prevalence estimates (p_{sar}) corresponding to state-s, age group-a and replicate-r.

Let F_{sa} denote the number of FI region groups in state-s containing population members from age group-a. Similarly, let B_{saf} denote the number of block groups in state-s, age group-a and FI region group-f. For age group-a, let $x_{a1}, x_{a2}, \ldots, x_{aq}$ denote the fixed predictors and $\beta_{r1}, \beta_{r2}, \ldots, \beta_{rq}$ the associated parameter estimates from the r th replicate. Now, let η_{sar} and υ_{sfar} denote the state-s and FI region group-f level random effect estimates for age group-a from the r-th replicate. Let n_{safbij} denote the population projection for block group-b of FI region group-f of state-s for age group-a, gender-$i(i=1,2)$, and race-$j(j=1,2,3,4)$. The prevalence estimate, $p_{sarfbij}$, for the virtual person (ij) in age group-a in block group-b of FI region group-f of state-s for the r-th replicate is given by $p_{sarfbij} = (1+\exp(-\lambda_{sarfbij}))^{-1}$ where $\lambda_{sarfbij} = (\sum_{k=1}^{q} x_{ak}(sfbij)\beta_{rk} + \eta_{sar} + \upsilon_{sfar})$ with $x_{ak}(sfbij)$ denoting the value of the k-th fixed predictor for the virtual person (ij) in age group-a in block group-b of FI region group-f of state-s. Hence we have

$$p_{sar} = \frac{\sum_{f=1}^{F_{sa}} \sum_{b=1}^{B_{saf}} \sum_{j=1}^{4} \sum_{i=1}^{2} p_{sarfbij}\, n_{safbij}}{\sum_{f=1}^{F_{sa}} \sum_{b=1}^{B_{saf}} \sum_{j=1}^{4} \sum_{i=1}^{2} n_{safbij}}.$$

Now to find the SWHB prevalence estimates for state-s and age group-a calculate the average of p_{sar} over the 1250 replicates. The corresponding 95% prediction or credible intervals are obtained in the following manner.

Let $l_{sar} = \log(\frac{p_{sar}}{1-p_{sar}})$, $\bar{l}_{sa} = \frac{1}{1250}\sum_{r=1}^{1250} l_{sar}$ and $s_{sa}^2 = \frac{1}{1250}\sum_{r=1}^{1250}(l_{sar} - \bar{l}_{sa})^2$ then the lower bound L and the upper bound U of the 95% prediction interval is given by

$L = (1+\exp(-(\bar{l}_{sa} - 1.96\, s_{sa})))^{-1}$ and $U = (1+\exp(-(\bar{l}_{sa} + 1.96\, s_{sa})))^{-1}$.

[1] The states are stratified into Field Interviewer (FI) regions. The FI regions are comprised of contiguous Census tracts.

CODA is used to validate the convergence of prevalence chains for a few selected states. After successful validation of convergence of the MCMC chains, then next step is to produce state by age group level SAE tables in Excel®.

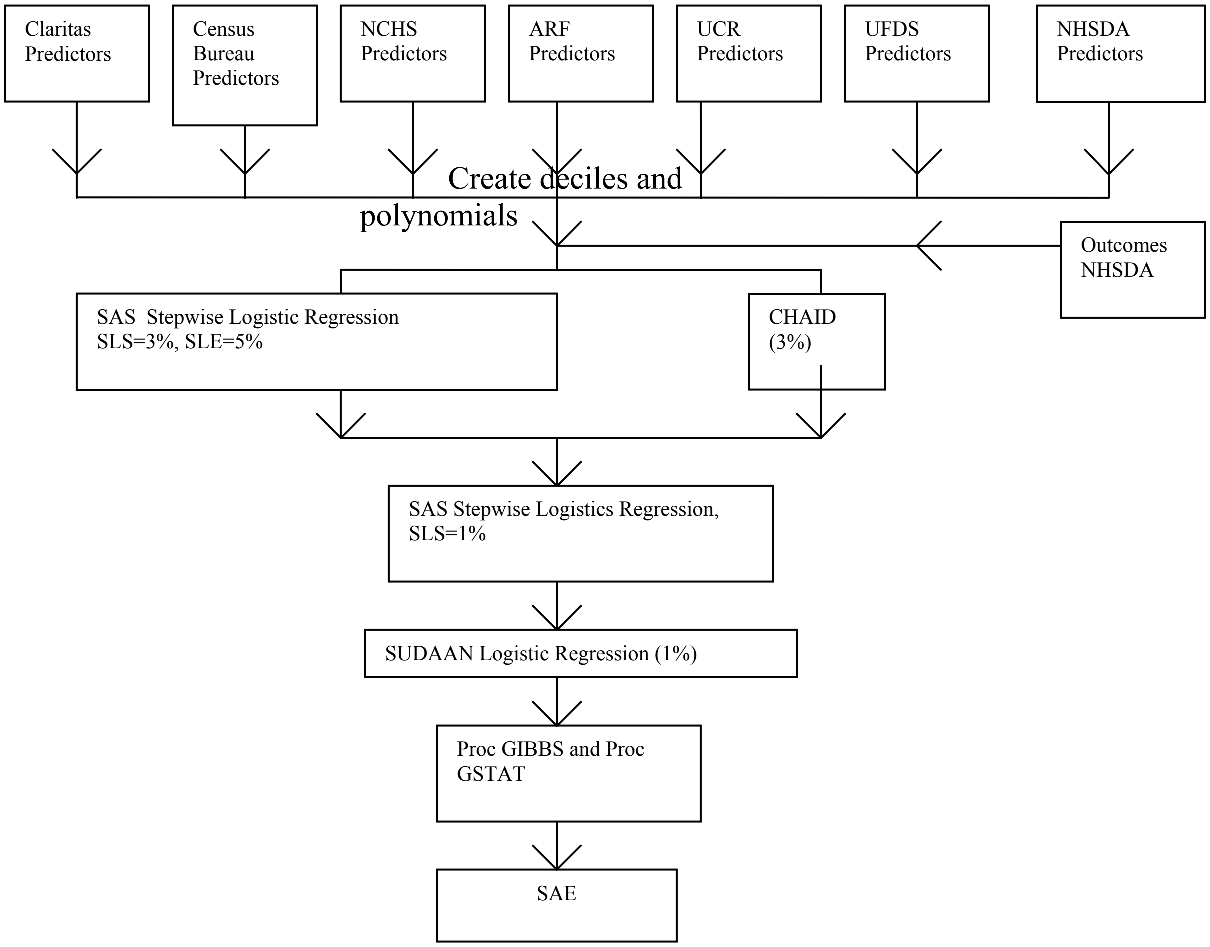
Produce SAEs for the Combined 2000 and 2001 NHSDA

The SAE measures of annual change were not precise enough to declare significant the size of annual changes that were observed. The SAE expert panel's[2] consensus was that the NHSDA should not be used to measure change between 1999 and 2000 or 2000 and 2001. Instead, the panel indicated that the SAMHSA would be better served by providing improved estimates of the prevalence levels based on combining two years of survey data. Therefore, combined 1999-2000 and 2000-2001 SAEs were produced. The following paragraphs briefly describe the methodology used for producing the combined 2000-2001 SAEs.

To produce the combined 2000-2001 SAEs, SWHB mixed logistic SAE models were fitted to the pooled 2000 and 2001 data. The fixed predictors used in these models were the same as those used in the 2000 SAE models. The pooled model parameters were then used to produce two year specific state level SAEs for each of 1,250 MCMC cycles, using the 2000 and 2001 universe files separately. The combined 2000 and 2001 SAEs for every MCMC cycle were produced by taking the weighted average of the two estimates using appropriate population counts. The final combined state level SAEs were produced by taking the average of weighted SAEs over all 1,250 MCMC cycles.

[2] The panel included William Bell of the U.S. Bureau of the Census; Partha Lahiri of the University of Nebraska; Balgobin Nandram of Worcester Polytechnic Institute and the National Center for Health Statistics; Wesley Schaible, formerly Associate Commissioner for Research and Evaluation at the Bureau of Labor Statistics; J.N.K. Rao of Carleton University; and Alan Zaslavsky of Harvard University. Other attendees involved in the development or discussion were Ralph Folsom, Judith Lessler, Avinash Singh, and Akhil Vaish of RTI and Joe Gfroerer and Doug Wright of SAMHSA.

Exhibit 4.3 Flowchart of the Small Area Estimation Process


Claritas Predictors
Census Bureau Predictors
NCHS Predictors
ARF Predictors
UCR Predictors
UFDS Predictors
NHSDA Predictors
Create deciles and polynomials
Outcomes NHSDA
SAS Stepwise Logistic Regression SLS=3%, SLE=5%
CHAID (3%)
SAS Stepwise Logistics Regression, SLS=1%
SUDAAN Logistic Regression (1%)
Proc GIBBS and Proc GSTAT
SAE

References

Folsom , R. E., Shah, B., & Vaish, A. (1999). Substance abuse in states: A methodological report on model based estimates from the 1994-1996 National Household Surveys on Drug Abuse. *Proceedings of the Section on Survey Research Methods of the American Statistical Association,* 371-375.

5. Table Production

For the production of NHSDA estimates, a series of automated processes was developed in order to produce high-end, publishable tables while minimizing error, minimizing production time, and increasing quality control. These processes employ the SAS® Macro Language available in SAS® v8.02 software, the SUDAAN® v8.0.1 software to properly account for the design of the NHSDA survey during statistical computation, and the macro facilities available in Corel© WordPerfect® 9 software to automate the production of tables.

Step 5.1: Prepare Data and Files for Table Production

- Create Analysis Datasets in SAS®
  - Analysis variables are created and output into SAS® datasets
  - Created variables follow coding specifications/restrictions pertaining to the analysis as well as those related to functionality in SAS® and SUDAAN®
  - Adjust analysis data according to type of analysis requested (single year, multi-year trend analysis, combined multi-year analysis.) This involves the creation of adjusted weights, changes in naming conventions, etc.
- Create Files Necessary for Final Table Production (See Step 5.3.)
  - Develop table shells in WordPerfect®
  - Create external files containing table title, note, and header information in WordPerfect®.

Step 5.2: Calculate Estimates and Generate Output Data into ASCII files

- SAS® macro programs were developed to automate the generation of required estimates and corresponding tabular information. These table-generation macros perform the following functions:
  - Compute weighted estimates using SUDAAN®
  - Compute variance estimates using SAS® and SUDAAN®
  - Compute significance tests using SUDAAN®
  - Implement NHSDA Suppression Rules to identify estimates of low precision
  - Compute 95% Confidence Limits (using SAS® statistical functions) for estimated totals and percents
  - Format output data for table presentation (inclusion of significance indicators, suppression indicators, title, note and footnote insertion indicators)
  - Output data into ASCII data files

- SAS® programs were developed to call the table-generation macros. Each program follows the same general template and allows the user to control the calculation and presentation of data that is output. Current programs produce from one to approximately 30 tables each. The tables can consist of up to six parts. User-defined specifications include:

    - Inclusion of analysis variables, domains, and design specifications for analysis
    - Control calculation of significance tests, confidence limits, suppression indicators
    - Exclusion/inclusion of data pertinent to given tables
    - Control of location and naming conventions of output data
    - Control of ordering and sequencing of output data
    - Control of numbering and inclusion of titles, notes, footnotes for tables

- Other automated table-generation programs were developed using SAS® and SUDAAN® for specific analyses including:
    - The evaluation of trends in initiation of drug use. For this analysis, programs call one SAS® macro for the computation of incidence rates for past years using multi-year data using SAS® and SUDAAN® and another SAS® macro for formatting output files to be used in table production.

    - The production of population counts among various domains. Similar to the main NHSDA automated table generation, SAS® programs call a SAS® macro that automates the computation of population numbers and format output data to be used in table production.

Step 5.3: Produce Tables

- Word-processing macros were developed to automate production of final, publishable tables. This process includes:
    - Proper placement of calculated estimates from ASCII data into existing table shells
    - Insertion of titles, notes, and headers from additional external references files into existing table shells

6. Disclosure

RTI developed a new tool for the disclosure treatment of the 2001 NHSDA. The resulting product was the 2001 NHSDA public use file (PUF) at the national level. The steps involved are described below.

Step 6.1: Initial Data Preparation (stratification for disclosure risk/recoding and suppression/substitution partners)

The identifying variables on the initial data file are specified. All obvious identifying variables such as detailed geographical information are suppressed. The remaining identifying variables are subjected to global recoding in order to reduce the number of risky records (e.g., cells with one, two, or three respondents). Disclosure risk strata are formed for subsampling and substitution. A substitution partner for each record is then found via a distance function, using identifying variables. The program to find substitution partners, written as a SAS® macro, takes more than ten hours for one run to complete on a Pentium III with 392 MB RAM and 800 MHZ speed.

Step 6.2: Subsampling

A disclosure loss incorporating the probability of subsampling of a record is assigned to each risk stratum. On this basis, a total disclosure loss function is defined. The within-stratum subsampling probabilities are determined such that the total disclosure loss is minimized subject to a set of variance constraints. The SAS® program used for this step uses PROC NLP to minimize the loss function subject to constraints, and then uses PROC SURVEYSELECT for subsample selection. The program typically runs pretty fast, but iterations are required using different tuning parameters in the loss function to achieve a desirable pattern in selection probabilities.

Step 6.3: Substitution

Optimal substitution is performed on the selected subsample. As in subsampling, a disclosure loss function is defined that incorporates the probability of a subsampled record being selected for substitution within a particular stratum. The within-stratum selection for substitution probabilities are determined such that the disclosure loss is minimized, given a set of mean squared error constraints. In addition to substituting the identifiers, variables related to the identifying variables are also substituted in order to maintain consistency among the variables. Moreover, to avoid possible risk of disclosure by a member of a pair, the pseudo-psu identifiers (VESTR and VEREP) for a small proportion (5%) of pairs that survived after subsampling and substitution are substituted from the corresponding partners. This step also uses PROC NLP of SAS®, but generally requires more effort and interventions to find a convergent and reasonable solution of the optimization problem.

Step 6.4: Calibration

The sample weights are calibrated so that the estimates of key outcome variables for various sociodemographic domains based on the PUF subsample reproduce the corresponding estimates for the original full data set. This is accomplished by using RTI's calibration tool, GEM, such that unit-specific bounds (in particular, for extreme weights) can be applied. The weight calibration report on the SAMHSA website contains a description of GEM, www.samhsa.gov/oas/NHSDA/2kmrb/00SamplingWeightW.pdf. This SAS® program for this step is not very time-consuming.

Step 6.5: Disclosure Treatment Evaluation

"Before" and "after" comparisons are made for some drug use prevalence measures for certain sociodemographic domains. Before and after comparisons are also made for the associated standard errors of the prevalence measures.

Step 6.6: Final Confidentiality Recodes

Although a core set of identifying variables are used to define risk strata, and confidentiality of records are protected with respect to these variables via subsampling and substitution, it is possible that some intruder might have information about some other variables which are not part of the core subset. To guard against such risk, the PUF is checked at the final step to further identify variables that have outlying values in the subsample, which might pose a confidentiality threat. These variables are top or bottom coded accordingly. The PUF is also checked to identify variables that have rare responses in the subsample. The variables are appropriately recoded to reduce disclosure risk. This step is done concurrently with other file preparation tasks to reduce the time necessary for the production of the PUF.