2006 National Health Interview Survey (NHIS) Paradata File
Public Use Data Release

NHIS Paradata File Description

Division of Health Interview Statistics
National Center for Health Statistics
Hyattsville, Maryland

Centers for Disease Control and Prevention
U.S. Department of Health and Human Services

January 2008

2006 National Health Interview Survey (NHIS) Paradata File
Public Use Data Release

Table of Contents

The NCHS Web Site and NHIS Electronic Mail List

Data users can obtain the latest information about the National Health Interview Survey, including the Paradata File, by periodically checking our Web site:

http://www.cdc.gov/nchs/nhis.htm

The Web site features links to downloadable public use data and documentation for the 2006 NHIS, as well as important information about any modifications or updates to the data and/or documentation. Published reports from previous years' surveys are also available, as are updates about future surveys and datasets.

For more information about the 2006 NHIS health data files, please see the Survey Description Document and supporting documentation for that release at the following site:

ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2006/srvydesc.pdf

Data users are encouraged to join the NHIS Listserv, an electronic mail list. The Listserv is made up of over 4,000 NHIS data users located around the world who receive e-news about NHIS surveys (e.g., new releases of data or modifications to existing data), publications, workshops, and conferences. To join, scroll down to "Related Sites" on the NHIS Web Site, and then click on "NHIS Listserv."

The Division of Health Interview Statistics also provides support services to data users. Users may contact us at 301-458-4901, or send e-mail to us at nhislist@cdc.gov.

2006 National Health Interview Survey (NHIS)
Public Use Paradata Release

Introduction

This document describes the first release of the 2006 National Health Interview Survey (NHIS) Paradata File, a complementary file to the 2006 NHIS public use health data files that were released in June 2007.

The National Health Interview Survey is a multi-purpose health survey conducted by the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC), and is the principal source of information on the health of the civilian, noninstitutionalized, household population of the United States. The U.S. Census Bureau, under a contractual agreement, is the data collection agent for the National Health Interview Survey. NHIS data are collected through a personal household interview by Census Field Representatives (FRs), with some telephone follow-up permitted to complete the case when necessary. The NHIS has been conducted continuously since its beginning in 1957. Public use microdata files are released on an annual basis.

The NHIS Paradata File does not contain health-related data, as does the NHIS regular public use data files, but rather contains paradata, which are data about the interview process. The data from the Paradata File are collected as part of the NHIS interview, using computer-assisted personal interviewing (CAPI). The paradata come from a number of sources:

- The Contact History Instrument (CHI), a supplement piece to the NHIS that collects data from the interviewer about each contact attempt. Data include strategies used for gaining participation and reasons for respondent reluctance.
- The Back section of the NHIS, where a series of questions are asked of the interviewer, including mode of interview (in-person visit vs. phone interview), the assessed cooperativeness of the respondent, and reasons for partial interviews/breakoffs.
- Date and time variables from each module of the instrument (Household, Family, Sample Child, Sample Adult). The date and time information is collected each time a module is started and completed.

The Paradata File is on a case (family) level, where one record represents one case. Unlike the NHIS public use health data release which contains information on fully complete and sufficiently complete interviewed cases only, the Paradata File also contains data on other types of cases, including cases that were ultimately refusals, insufficient partials, and other types of nonresponse. These are referred to as *Type A* cases. In addition, data are provided on cases that were deemed out of scope, such as families with Armed Forces-only adults, and cases that were screened out by race/ethnicity. The out-of-scope cases are referred to as *Type B* outcomes in this document. A more detailed explanation of the outcome codes on the Paradata File can be found in the section "General Information about the 2006 Data" later in this document.

The Paradata File is intended as both a stand-alone data file and one whose fully

complete and sufficiently complete cases can be linked with the NHIS regular public use health data file. More information about linking the regular health files and Paradata File can be found in Appendix I.

Sample Design

The sample design for the 2006 Paradata File is the same as that for the regular 2006 NHIS public use data files. This information can be found in the Survey Description Document accompanying the regular 2006 NHIS health data files release. This document is available at the following location:

ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2006/srvydesc.pdf

Weighting Information

The NHIS is designed to make inferences about the civilian noninstitutionalized population of the United States. This is accomplished by inflating the responses of each surveyed person or household by a national weight factor. The national weight permits calculation of approximately unbiased design-based estimates of U.S. target population totals.

A weight variable is included on the Paradata File (WTIA_PD). This weight reflects the probability of household selection and does not include nonresponse and post-stratification adjustments. It is the appropriate weight to use when analyzing all cases (responding, nonresponding, and out-of-scope) on the Paradata File (see below).

When using the Paradata File on a stand-alone basis, data users have the option of weighting or not weighting estimates. If the goal is to make population inferences, then the paradata weight should be used. For example, a data user may want to estimate the proportion of the household population that fails to participate in face-to-face interview surveys because of privacy concerns or time burdens. The paradata weight would be necessary to address this research question.

However, not every analysis involving the Paradata File will entail making population inferences. For example, a data user may want to describe the level of effort (number of contact attempts) expended by NHIS interviewers across a wide range of final case dispositions (complete interview, partial interview, noncontact, refusal, etc.). Use of WTIA_PD would not be necessary in this situation.

NOTE: If data from the Paradata File are used in support of analyses of the 2006 NHIS health data files, the corresponding weights provided on the health data file should be used. For example, a data user may be interested in exploring the determinants of health care access and utilization among sample adults and would like to control for mode of interview. The necessary mode measures located on the Paradata File would be merged with the Sample Adult File, and the final sample adult weight (WTFA_SA) would be used in the subsequent analysis.

Variance Estimation

   The data collected in the NHIS are obtained through a complex sample design involving stratification, clustering, and multistage sampling. While standard software packages may provide unbiased point estimates for commonly computed first-order statistics like means or regression coefficients, the computed estimated standard errors of the estimates may be too small if the software package does not account for the complex sample design. Also, standard packages may produce hypothesis test results (such as p values) that are incorrect. Therefore, it is recommended that users of the NHIS Paradata File utilize computer software that provides the capability of variance estimation and hypothesis testing for complex sample designs. NCHS uses the software package SUDAAN (Research Triangle Institute, 2004) with Taylor series linearization methods for NHIS variance estimation. SUDAAN code and a description of its use to compute standard errors of means, percentages and totals with the 2006 NHIS data files can be found at the following location:

ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2006/srvydesc.pdf

General Information about the 2006 Data

   The Paradata File contains 44,264 records. Of these records, 34,270 cases are considered "in scope," i.e., cases that are eligible for interview. A case with an outcome code of 201 is a fully completed interview, in which the household composition, family, sample adult, and sample child (if a child was present) modules were completed. A case with an outcome code of 203 is a sufficient partial interview, meaning that at least a sufficient portion of the family module was completed. Cases with an outcome code of 213, 215, 216, 217, 218, or 219 are in-scope cases that did not result in either a fully complete or sufficiently complete interview. These include refusal cases (218) and insufficient partial cases (215) where the family module was started, but not completed to a sufficient degree. These are referred to as *Type A* or *nonresponding* cases. In addition, 9,994 cases are considered "out of scope" and thus ineligible for interview (also called *Type B* cases). Type B cases are families comprised of military-only adults, families whose usual residence is elsewhere, and families screened out by race/ethnicity. Type B cases are included in the data file because, although they are not counted in the NHIS response rate, they represent households where some contact was made with person(s) at that address in order to determine eligibility for the interview. Table 1 shows the frequency distribution of interview outcome codes for the Paradata File.

Table 1: Frequency Distribution of Cases by Outcome Code	
Outcome Code	*Frequency*
In-scope cases	
Interview Cases	
201 – Fully Complete Case	24,323
203 - Sufficient Partial Case	5,847
Type A Cases (Nonresponding)	
213 - Language Problem	63
215 - Insufficient Partial Case	438
216 – No one home, repeated calls	891
217 – Temporarily absent, no follow-up	204
218 – Refusal Case	2,156
219 – Other Type A	348
Out-of-scope cases	
299 - **Type B**: Occupied entirely by Armed Forces adults, occupied entirely by persons with Usual Residence Elsewhere, screened out by household (race/ethnicity)[1]	9,994
TOTAL	44,264

Information about the 2006 Paradata File Documentation

Datasets and related documentation for the Paradata File are available on the NHIS Web site, **http://www.cdc.gov/nchs/nhis.htm**. The Web site provides the Paradata File Description Document and supporting documentation. Information about the general 2006 NHIS data release is also available here.

Like the regular NHIS release of health data, the Paradata File includes the supporting documentation listed below.

- Variable Summary Report
- Variable Layout Report
- ASCII Data
- Variable Frequency Report
- Sample SAS statement

---

[1] Prior to interviewing, one part of the NHIS sample is assigned to be "screened". In this part of the sample, the NHIS interview proceeds through the collection of the household roster. The interview then continues only if the household roster contains one or more black, Asian, or Hispanic persons. Otherwise, the interview terminates and the household is said to be "screened out". In the other part of the NHIS sample, full interviews occur at all households.

The *Variable Summary Report* lists each variable, a brief description of the variable, the question number on which it was based, and the variable location in the released ASCII file. For most variables, the *Variable Layout Report* provides the actual question that generated the data, questionnaire location information (if applicable), instrument variable name, universe, response values, and response value labels. Additional specific information is provided under "Recodes," "Sources," "Keywords," and "Notes." These terms are defined below:

> *Recodes* - A recode is a variable derived from the reordering, collapsing, or verbatim coding of another variable, such as the recode UNABLE1R found in the Paradata File. Alternatively, a recode may be constructed from two or more variables, such as the point in the interview period when the household composition section was started (STRTPNT). If a particular variable was used in making recode variables, then those recode variables are listed as a cross reference. Users will note that a number of standardized variables appear in the dataset. A *standardized variable* is a particular type of recode based on time unit information obtained during the course of the interview. When respondents are asked any questions pertaining to time - for example, how long the family has been without phone service - the answer is typically obtained in two parts. The respondent provides the number of time units, followed by the type of time unit. During the course of data editing, this information is standardized into a single appropriate time unit. Examples of this on the Paradata File include HOWLNGWK, RH1LNGDY, and RH2LNGDY.

> *Sources* - If the variable in question is a recode, then all variables that were used to make this recode are listed.

> The Paradata File contains a number of recodes related to the date and time variables, including the start point of the household composition section in the interview period and the time of day when the household composition section was started. For confidentiality reasons, many of the original continuous measures (i.e., number of days) have been recoded into ordinal, or ranked measures such as "Early," "Middle," or "Late". Further information about these specific recodes can be found in the "Measures of Time" section of the document.

> *Keywords* - Keywords are descriptive words or phrases relevant to the topic of the variable; these can be used for word searches.

> *Notes* - Notes provide information that analysts need to know about a particular variable, such as assumptions, limitations, caveats, differences between instrument versions, or other important information. Analysts are encouraged to read the notes for variables of interest.

> The Variable Layout Report specifies the universe for each variable in the Paradata File. The universe refers to those respondents eligible to be asked a given question. For example, the universe for many variables on the Paradata file is "All cases." This means that all cases should have a valid value for the variable.

The Variable Frequency Report provides the frequencies, percents, and the frequency missing (not-in-universe) for each variable. Within the NHIS, the same codes are used across all files to designate "Refused" and "Don't know" responses: refusals are coded as "7" (with leading 9's to the length of the field, as in 7, 97, 997, etc.), while "don't know" responses are "9" (again, with leading 9's to the length of the field, such as 9, 99, 999, etc.). A code of "8" is used to indicate "Not ascertained" responses, which typically occur when an in-the-universe respondent had a blank field or the field contained an impossible code. There are few variables on the Paradata File that contain values of "Refused" or "Don't know" because most variables are self-reported by the interviewer, and thus a value should be assigned to all cases. One item to note is that there are 988 cases that do not have a CHI record, because the interviewer did not record information for any of their contact attempts. This, however, represents a small proportion of total CHI entries.

A few of the variables on the Paradata File are also found on the regular NHIS health data files. In particular, the items on telephone usage and outages are found on both files. They are included on the Paradata File because insufficient partial cases (outcome codes 215) and other Type A and Type B cases that are not part of the regular NHIS health data files may contain this information.

Case-Level File

The Paradata File is a case-level file. A case is defined as one record consisting of an individual or a group of two or more related persons who are living together in the same occupied housing unit (i.e., household) in the sample. In some instances, unrelated persons sharing the same household may also be considered as one family (hence one case), such as unmarried couples who are living together. Each record in the file represents a unique family. Although multiple families may share one household, each individual family represents a separate case. What follows are descriptions of conceptually grouped variables included on the Paradata File.

Measures of Time

A series of time measures can be found on the Paradata File. The first set of time measures convey information on the start dates of main sections of the survey instrument and the completion date of the case. For confidentiality purposes, actual dates are not provided. Instead, original date variables have been recoded into categorical measures that anchor the case within the NHIS 17-day interview period.[2] For example, STRTPNT indicates that the household composition section of the survey instrument was started in the early, middle, or late portion of the 17-day interview period. Similarly, ENDPNT indicates that the case was completed in the early, middle, or late portion of the 17-day interview period. Among several possible uses, measures of this sort allow researchers to explore whether the quality of data (e.g., item nonresponse) collected varies by when in the interview period the case was started and completed.

---

[2]When an interviewer receives a new assignment of sample cases (issued each Monday for 50 weeks of the year), he/she has 17 days (interview period) to complete the assignment. An interview period begins on a Monday and ends on a Wednesday.

The following measures correspond to the starting points, within the 17-day interview period, of main sections of the survey instrument. Valid values are present on these measures for cases where the specific section or module of the survey instrument was started.

- STRTPNT indicates that the household composition section of the instrument was started early in the interview period, in the middle of the interview period, or late in the interview period.

- FMSTRPNT indicates that the family module was started early in the interview period, in the middle of the interview period, or late in the interview period.

- SCSTRPNT indicates that the sample child module, if there was a sample child, was started early in the interview period, in the middle of the interview period, or late in the interview period.

- SASTRPNT indicates that the sample adult module was started early in the interview period, in the middle of the interview period, or late in the interview period.

The following measure identifies the point in the 17-day interview period when the case was completed and a final disposition (or outcome) code was assigned.

- ENDPNT indicates that the case was completed and a final disposition (or outcome) code was assigned early in the interview period, in the middle of the interview period, or late in the interview period.

Four additional items capture the time of day when sections of the survey instrument were started. The measures were recoded into three broad time periods of morning (including overnight), afternoon, and evening. Valid values are present on these measures for cases where the specific section or module of the survey instrument was started.

The following is a description of the time of day measures:

- HHC_TOD indicates that the household composition section was started in the morning, afternoon, or evening.

- FAM_TOD indicates that the family module was started in the morning, afternoon, or evening.

- SC_TOD indicates that the sample child module, if there was a sample child, was started in the morning, afternoon, or evening.

- SA_TOD indicates that the sample adult module was started in the morning, afternoon, or evening.

Measures of Contactability

One of the conceptual groupings of measures on the Paradata File are those describing contactability, or the effort involved in making contact with a sample household. As with all measures on the file, this particular group may be utilized in many types of analyses, though conceptually they are particularly amenable to exploring survey contact and, conversely, noncontact processes.

The bulk of "noncontact" measures on the Paradata File are case-level summaries of CHI-based, interviewer-recorded explanations of noncontact outcomes. Recall that for each contact attempt that an interviewer makes on a sample household, she/he must report one of the following three outcomes: contact with a sample unit member, contact with a non-sample unit member, and noncontact. There are two sets of CHI-based noncontact measures based on mode of attempt.

- In-person noncontact measures: NCTPR01R through NCTPR14R:

  These dichotomous measures (0=never entered by interviewer across all in-person noncontacts, 1=entered by interviewer at 1 or more in-person noncontacts) describe noncontact outcomes in personal visit contact attempts. Examples include "no one home—appointment broken" and "unable to reach/locked gate/buzzer entry." The interviewer could record up to 14 different reasons or explanations.

- Telephone noncontact measures: NCTEL01R through NCTEL07R:

  These dichotomous measures (0=never entered by interviewer across all telephone noncontacts, 1=entered by interviewer at 1 or more telephone noncontacts) describe noncontact outcomes in telephone contact attempts. Examples include "busy signal" and "disconnected." The interviewer could record up to 7 different reasons or explanations.

Five additional measures are included under this conceptual grouping: CTSTAT1, CTSTAT2, CTSTAT3, UNABLE1R and VISITCNT.

- CTSTAT1, CTSTAT2, and CTSTAT3 provide the number of contact attempts that resulted in contact with sample unit members, the number of contact attempts that resulted in contact with non-sample unit members, and the number of contact attempts that resulted in noncontact.

- UNABLE1R provides the number of times contact was made with a sample unit member but the interviewer was unable to conduct or complete an interview because the "eligible person was not available." This measure captures situations of sub-level noncontact - contact was made at the household level, but the household member needed to complete a particular section of the interview was not available. This largely pertains to the sample adult, followed by the sample child respondent.

- VISITCNT provides the number of personal visits made by the interviewer to the sample unit. This is an interviewer-reported measure captured in the Back section of the main survey instrument. It is not part of CHI. Each time interviewers make personal visits to a sample unit, they are asked to indicate the number of times they have made a personal visit to that unit so far. The item retains the entry from the last personal visit, allowing the interviewer to add 1 to the previous value. NOTE: Unlike the number of attempts measure taken from CHI, VISITCNT does not include telephone attempts.

Points to Remember When Working with These Measures:

- As noted previously, CHI entries were not made for 988 (or 2.23%) cases.

- The CHI-based measures presented above were created by summing a particular entry across all contact attempts and then top-coding values greater than one. For example, if two personal visit contact attempts for a household resulted in noncontact and the interviewer entered "on vacation, away from home" for both, the summary CHI measure (NCTPR09R) for this reason would have a value of 1. This would indicate that "on vacation, away from home" was entered at one or more personal visit contact attempts that resulted in noncontact. The measures are not an indication of what occurred at any one attempt unless the case was resolved in a single contact attempt.

- These measures are based on interviewer self-reports - a corresponding amount of recall error is expected, especially when CHI entries were not made at the time of the contact attempt.

Contact Strategies Attempted

As part of CHI, interviewers are instructed to record all strategies used on every contact attempt. Twenty-three possible strategies, including "none" and "other" categories, are available. The following is a list of each possible strategy and its description:

- Advance letter given (STRAT01R): Interviewers are instructed to ask the household member if he/she received the advance letter (mailed from the regional office several weeks before the interview period). If the household member did not receive the letter, the interviewer will provide the household member with a copy of the letter. The letter can also be left in situations of noncontact.

- Scheduled appointment (STRAT02R): If the respondent is unavailable at the time of contact, the interviewer may choose to schedule an appointment some other time during the interview week.

- Left note/appointment card (STRAT03R): If no contact is made at the time the interviewer visits the household, he/she may leave a personalized note or an appointment card.

- Left promotional packet/informational brochure (STRAT04R): If the householder is not available at the time of visit, the interviewer may give the householder informational materials to review before an interview is conducted. These materials may include brochures about how the NHIS data are used, a ruler, magnet, or other information related to the survey.

- Called household (STRAT05R): The interviewer may choose to call the household to make contact. This is usually done after a personal visit has been attempted.

- Left message on answering machine (STRAT06R): Interviewers may choose to leave a message on an answering machine for the householder.

- Will request No One Home Letter (STRAT07R): The interviewer may request the regional office staff to send a letter to a household for which repeated visits have yielded no response.

- Will request Refusal Letter (STRAT08R): The interviewer may request the regional office staff to send a letter to a household that has refused.

- Will request Better Understanding Letter (STRAT09R): The interviewer may request the regional office staff to send a letter to a household that has requested more information about the survey.

- Called contact person (STRAT10R): If the interviewer has difficulty contacting a respondent, he/she may choose to contact a friend or relative in order to help to locate the respondent.

- Stake-out (STRAT11R): The interviewer may wait near the residence in the hopes that the interviewer will see the householder arrive.

- Checked with neighbors (STRAT12R): The interviewer may try to learn more about the best time to reach the householder by asking neighbors about the householder's schedule.

- Contacted other family members (STRAT13R): If the interviewer has difficulty contacting a respondent, he/she may choose to contact a relative of the respondent in order to help to locate the respondent.

- Contact property manager (STRAT14R): The interviewer may need to contact the property manager if the community where the household is located has limited access.

- Visited county assessor/post office/permit office (STRAT15R): The interviewer may utilize the county assessor, post office, or permit office in finding an address that is difficult to locate.

- On-line tracking database (STRAT16R): The interviewer may utilize a tool with Census-based information about the sample address.

- Sought help from SFR/RO (STRAT17R): If the interviewer is having difficulty contacting the household or gaining cooperation from a householder, he/she may contact a senior interviewer and/or the regional office staff for advice or assistance.

- Reassignment (STRAT18R): The interviewer may ask that the case be reassigned if the householder speaks a different language, is located far away, or is difficult to contact or persuade participation.

- Used MAF or ALMI (STRAT21R): The interviewer used a Census-based mapping system to locate the household.

- None (STRAT22R): No strategies were used on this contact attempt.

- Other-specify (STRAT23R): Some other strategy not listed was used on the contact attempt.

Cooperation

Another conceptual grouping of variables found in the Paradata File is related to respondent cooperation. The more the respondent cooperates with the question and survey processes, the greater the chance for a quality interview. Variables such as COOPFAM and RESPOND are interview-assessed measures of cooperativeness and the likelihood that the respondent would participate in another survey. Several measures of cooperation also come from the Contact History Instrument.

- Language concerns: LANG1R – LANG5R

  These dichotomous coded measures (0=never entered, 1=entered at 1 or more contacts) describe contact outcomes in contact attempts, both in person and over the telephone. Some examples include "No household member able to translate" and "Unable to find translator."

- Respondent reluctance: RELUC01R – RELUC23R

  These dichotomous coded measures (0=never entered, 1=entered at 1 or more contacts) represent contact outcomes in contact attempts, both in person and over the telephone. Some examples include "Too busy" and "Interview takes too much time."

- Unable to complete interview: UNABLE2R-UNABLE8R

  These dichotomous coded measures (0=never entered, 1=entered at 1 or more contacts) represent contact outcomes in contact attempts, both in person and over the telephone. Some examples include "Inconvenient time" and "Health problem." These variables could also be conceptually grouped with the partial/break-off variables from the file.

Partials/Breakoffs

A series of questions to ascertain reasons for partially completed interviews (interviews that are not entirely complete) appear in the Back section of the instrument. PARWHY is a variable that differentiates between the interview being stopped in progress (break-off) and being unable to reach a particular respondent or obtain some information on follow-up (partial, other than break-off). For cases marked as a break-off in PARWHY, BRKWHER is a variable that determines the main section of the survey in which the break-off occurred. BRKRES allows interviewers to select one *main* reason that the interview stopped in progress from a list of ten answer categories. Similarly, NCOMRES allows interviewers to record the *main* reason that the interview is not complete.

Mode Measures

The NHIS is primarily a personal visit survey. However, there are limited circumstances in which the interviewers are permitted to conduct a section by telephone. Once an initial personal contact has been made, telephone contacts may be used to complete partial interviews, or to collect other missing parts of the interview. There are two sources of mode data on this file: data at the household level from the Back section and data from the summary Contact History File.

Back Section Variables

At the completion of each interview, interviewers are asked a series of questions about the administration of the interview. Interviewers record a "yes" or "no" answer to each of the following question(s): "Were any of the following sections conducted primarily by telephone? Household Composition? Family? Sample Adult? Sample Child?" No precise meaning of the word "primarily" is given to the interviewers; that determination is left up to the interviewers.

INTMODE indicates whether any sections of the interview were conducted by telephone.

CHI Variables

For each contact attempt, the interviewer records whether the attempt was made by telephone or by personal visit.

- MODE_P indicates the number of contact attempts on a case that were made by personal visit. To ascertain the proportion of contact attempts made by personal visit, divide MODE_P by the total number of contact attempts (TOTCOUNT).

- MODE_T indicates the number of contact attempts on a case that were made by telephone. To ascertain the proportion of contact attempts made by telephone, divide MODE_T by the total number of contact attempts (TOTCOUNT).

Case-Level Information

There is also a set of variables that give general case-level information about the record. For each case, the variable SRVY_YR (survey year) will have a value of "2006." This variable will become more useful in future years, when data may be pooled for trend analysis. Interview quarter (INTV_QRT) and assignment week (ASSIGNWK) indicate the general time when the case was in the field. There are 4 quarters and 13 assignment weeks in each quarter, although cases are not worked in weeks 1-2 of the first quarter due to interviewer training. As mentioned previously, each interview period is 17 days in length.

The variable OUTCOME1 represents the disposition of the case. If the Paradata File is linked to the health data files, it is important to remember that there is not a 1:1 match between the two. The Paradata File represents field data before cleaning. There are some 201 and 203 outcomes (fully complete and sufficiently complete interviews) on the Paradata File that were dropped from the final 2006 public use health files because the record was not deemed adequate for data release (for example, there may have been too many "Don't know" or "Refused" responses). As mentioned before, the Paradata file also contains in-scope but not sufficiently complete cases (refusals, insufficient partial cases, etc.) and out-of-scope cases (usual residence elsewhere, screened out by race or ethnicity, etc).

Guidelines for Citation of Data Source

With the goal of mutual benefit, the National Center for Health Statistics (NCHS) requests that recipients of data files cooperate in certain actions related to their use. Any published material derived from the 2006 Paradata File should acknowledge NCHS as the original source. The suggested citation to appear at the bottom of all tables and graphs is as follows:

Data Source: National Center for Health Statistics (2006).

In a bibliography, the suggested citation should read:

National Center for Health Statistics (2008). *Paradata Data File Documentation, National Health Interview Survey, 2006 (machine readable data file and documentation).* National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, Maryland.

The published material should also include a disclaimer that credits any analyses, interpretations, or conclusions reached to the author (recipient of the data file) and not to NCHS, which is responsible only for the initial data. Users who wish to publish a technical description of the data should make a reasonable effort to insure that the description is consistent with that published by NCHS. Information on how to cite electronic media is available at:
http://www.cdc.gov/nchs/howto/citelec.htm

References

Delwiche, LD and SJ Slaughter (1998), *The Little SAS Book: A Primer* (2nd edition), SAS Institute: Cary, NC.

National Center for Health Statistics (2006). *Survey Description Document, National Health Interview Survey, 2006* .

Research Triangle Institute. (2004). SUDAAN *Language Manual; Release 9.0*, Research Triangle Institute, Research Triangle Park, NC.

Appendix I- Linking the 2006 Paradata File to 2006 Public Use Health Data Files

The variables HHX and FMX can be used to link the 2006 Paradata File with the 2006 NHIS health data files. HHX is a household identification number for each record in both files. FMX represents the family number of the case within the household. For approximately 98% of 201 or 203 cases (complete or sufficiently complete interviews), there is a single family in the household, thus the FMX value will be "01". For those cases with more than one family in the household, HHX will be the same, but the FMX will have values of "01", "02", etc.

Nearly all of the 201 or 203 outcomes on the Paradata File can be linked with the health data files. As mentioned previously, the Paradata File represents the field data before cleaning; thus, there are some cases, including 201 and 203 outcomes, that are dropped before the creation of the public use health data files. These non-matched 201 and 203 outcomes, along with the Type A and Type B outcomes, will be excluded when the files are linked.

For example, the 2006 Family Health File has 29,868 records. Merging on the variables HHX and FMX, 29,868 of the 44,264 records from the Paradata File will be pulled in to complete the merge. The final size for the merged file in this case should be 29,868 records. If making population inferences using this linked file, the final annual weight on the health data file (WTFA_FAM) should be used.

An example of SAS code to merge the 2006 Paradata File with the 2006 NHIS Family File is below:

/* Merge the Paradata File and the Family file. */

 /* Create a Paradata dataset with selected variables and sorted by HHX and FMX.*/

DATA PARADATA (KEEP=HHX FMX MODE_P); /* PARADATA is a SAS dataset; the KEEP statement retains only the listed variables for processing. */

SET NHIS.PARADATA; /*The SET statement reads data from the 2006 Paradata File. */
PROC SORT DATA=PARADATA; /* Sort by HHX, and FMX, the household and family identifier. */
BY HHX FMX; RUN;

/* Create a Family dataset with selected variables and sorted by HHX and FMX. */

DATA FM (KEEP=HHX FMX INCGRP RAT_CAT WTFA_FAM); /* FM is a SAS dataset; the KEEP statement retains only the listed variables for processing. */
SET NHIS.FAMILYXX; /*The SET statement reads data from the 2006 Family file. */
PROC SORT DATA=FM; /* Sort by HHX and FMX the household and family identifier. */
BY HHX FMX; RUN;

DATA PARAFM; /* New combined dataset called **PARAFM** */
MERGE FM (IN=FROMFM) PARADATA ; /* Merge the newly created FM and PARADATA files, using an IN statement.*/

BY HHX FMX; IF FROMFM = 1; /* The combined dataset PARAFM will contain only those records that are in the Family file; the Paradata file's MODE variable will be appended to these records. */
PROC SORT DATA= PARAFM; /* Sort by HHX and FMX, the household and family identifiers. */
BY HHX FMX; RUN;

In the code above, the IN statement creates a temporary SAS variable (called FROMFM) that has a value of 1 if the dataset associated with the IN statement contributed to the current observation, or a value of 0 if it did not. The subsequent statement "IF FROMFM = 1" tells SAS to retain only those observations from the Family file (called FM). For more information on IN statements in SAS, consult Delwiche and Slaughter (1998).