

The 2007 NHIS Paradata File contains data about the NHIS data collection process. It may be used as a stand-alone data file or linked to the NHIS 2007 health data files.

The 2007 Paradata File and documentation can be found at:

[http://www.cdc.gov/nchs/about/major/nhis/nhis\\_2007\\_release.htm](http://www.cdc.gov/nchs/about/major/nhis/nhis_2007_release.htm)

**2007 National Health Interview Survey (NHIS)  
Public Use Data Release**

**NHIS Survey 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**

**June 2008**

**2007 National Health Interview Survey (NHIS)  
Public Use Data Release**

**Table of Contents**

The NCHS Web Page and NHIS Electronic Mail List .....	4
What’s New in 2007 .....	5
Introduction.....	6
Data Collection Procedures .....	7
Sample Design .....	7
Sample Reductions.....	8
Weighting Information.....	8
General Information about the Data .....	11
Information about the Data File Documentation .....	12
Information about the 2007 CAPI Questionnaire .....	15
Supplements, Supplement Co-sponsoring Agencies, and Question Locations .....	17
Household-Level File .....	19
Family-Level File .....	20
Person-Level File.....	24
I. Health Status and Limitation of Activity Section (FHS).....	24
II. Health Care Access and Utilization Section (FAU).....	31
III. Health Insurance Section (FHI) .....	31
IV. Socio-demographic Section (FSD) .....	34
V. Income and Assets Section (FIN) .....	37
Injury and Poisoning Episode File.....	43
I. Injury and Poisoning Episode File .....	44
II Verbatim Injury and Poisoning Episode File.....	48

Sample Child File .....	50
I. Child Conditions, Limitation of Activity and Health Status Section (CHS).....	50
II. Child Health Care Access and Utilization Section (CAU) .....	51
III. Child Mental Health Brief Section (CMB).....	52
IV. Child Mental Health Services (CMS).....	53
V. Child Influenza Immunization (CFI) .....	53
Child Complementary and Alternative Medicine File (CAL) .....	55
Sample Adult File .....	57
I. Adult Demographics Section (ASD) .....	57
II. Adult Conditions Section (ACN) .....	59
III. Adult Health Status and Limitation of Activity Section (AHS) .....	62
IV. Adult Health Behaviors Section (AHB) .....	65
V. Adult Health Care Access and Utilization Section (AAU).....	68
VI. Adult AIDS Section (ADS) .....	68
Adult Complementary and Alternative Medicine File (ALT) .....	70
Coverage Section .....	72
Guidelines for Citation of Data Source.....	73
References .....	74
Appendix I. Calculation of Response Rates for the 2007 NHIS .....	75
Appendix II. Race and Hispanic Origin.....	85
Appendix III. Variance Estimation and Other Analytic Issues .....	97
Appendix IV. A Preliminary Evaluation and Recommendations for use of the Mental Health Indicator (MHI) in the NHIS for Children Aged 2 to 3 years.....	109
Appendix V. The Short Strengths and Difficulties Questionnaire (SDQ).....	111
Appendix VI. Transition to the 2000 Census-based Weights.....	113
Appendix VII. Merging Data Files and Combining Years of Data .....	114
Appendix VIII. Core Changes/Additions/Deletions in 2007.....	120

## **The NCHS Web Page and NHIS Electronic Mail List**

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

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

The Web site features downloadable public use data and documentation for the 2007 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.

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 page, 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](mailto:nhislist@cdc.gov).

### **What's New in 2007**

- In order to achieve cost savings, the third-quarter NHIS sample was reduced by approximately 50%, cutting about 13% of the sample size of the 2007 NHIS sample overall. This cutback was in addition to the ongoing 13% reduction due to the new sample design instituted in 2006.
- In 2007 questions about high deductible insurance plans (HDHP), health savings accounts (HSA) or health reimbursement accounts (HRA), and flexible spending accounts (FSA) for medical expenses were added to monitor health care in the NHIS Health Insurance section (FHI).
- The new income questions in the 2007 NHIS reduced item non-response and improved data quality.
- The top income category in the grouped income recodes for 2007 was increased to \$100,000 and over.
- Cellular telephone questions were modified and moved from the Recontact Section to the Coverage Section.
- The Adult Complementary and Alternative Medicine Supplement was fielded in 2007. Cost data for complementary and alternative medicine use were collected for the first time.
- In 2007 the Child Complementary and Alternative Medicine Supplement was fielded for the first time.

## **2007 National Health Interview Survey (NHIS) Public Use Data Release**

### **Introduction**

The National Health Interview Survey (NHIS) 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 NHIS has been conducted continuously since its beginning in 1957. Public use microdata files are released on an annual basis.

The NHIS Core questionnaire items were revised every 10-15 years, with the last major revisions occurring in 1982 and in 1997. The NHIS that was fielded from 1982-1996 consisted of two parts: (1) a set of basic health and demographic items (known as the Core questionnaire) that remained stable from one survey year to the next, and (2) one or more sets of questions on current health topics that varied with each survey, referred to as Supplements. Despite periodic revisions to the Core questionnaire, Supplements played an increasingly important role in the survey as a means of enhancing topic coverage in the Core. Eventually, certain Supplements, such as “Family Resources” and “Health Insurance,” were incorporated in the NHIS Core on an annual basis.

The redesigned NHIS introduced in 1997 consists of a Basic Module or Core as well as variable Supplements. The Basic Module, which remains largely unchanged from year to year, consists of three components: the Family Core, the Sample Child Core, and the Sample Adult Core. The Family Core component collects information on everyone in the family, and its sample also serves as a sampling frame for additional integrated surveys, as needed. Information collected for all family members includes: household composition and socio-demographic characteristics, tracking information, information for matches to administrative data bases, and basic indicators of health status, activity limitations, injuries, health insurance coverage, and access to and utilization of health care services.

From each family in the NHIS, one sample child (if any children under age 18 are present) and one sample adult are randomly selected, and information on each is collected with the Sample Child Core and the Sample Adult Core questionnaires. Because some health issues are different for children and adults, these two questionnaires differ in some items, but both collect basic information on health status, health care services, and behavior. These sections of the survey yield the Sample Child and Sample Adult data files.

The Family Core yields several data files, including the Household-Level file, the Family-Level file, the Person-Level file, and two data files pertaining to injuries and poisonings. Because these files contain the same or comparable variables from one survey year to the next, they are suitable for trend analysis; moreover, multiple years of these data may be easily pooled to increase the sample size for analytic purposes.

## **Data Collection Procedures**

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 interviewers. Nationally, the NHIS uses about 600 interviewers, trained and directed by health survey supervisors in the 12 U.S. Census Bureau Regional Offices. The supervisors responsible for the NHIS are career Civil Service employees who are selected through an examination and testing process. Interviewers (also referred to as Field Representatives, or “FRs”) receive thorough training on an annual basis in basic interviewing procedures and in the concepts and procedures unique to the NHIS.

For the Family Core component of the Basic Module, all members of the household 18 years of age and over who are at home at the time of the interview are invited to participate and to respond for themselves. For children and those adults not at home during the interview, information is provided by a knowledgeable adult family member (18 years of age or over) residing in the household. Information for the Sample Child questionnaire is obtained from a knowledgeable adult residing in the household. For the Sample Adult questionnaire, one adult per family is randomly selected; this individual responds for him/herself to the questions in that section unless he/she is physically or mentally unable to do so, in which case (200-300 cases per year) a knowledgeable proxy is allowed to answer for the sample adult.

The NHIS is conducted using computer-assisted personal interviewing (CAPI). The CAPI data collection method employs computer software that presents the questionnaire on computer screens to each interviewer. The computer program guides the interviewer through the questionnaire, automatically routing the interviewer to appropriate questions based on answers to previous questions. Interviewers enter survey responses directly into the computer, and the CAPI program determines if the selected response is within an allowable range, checks it for consistency against other data collected during the interview, and saves the responses into a survey data file. On-line help facilities are available to aid interviewers in administering the CAPI questionnaire. This data collection technology reduces the time required for transferring, processing, and releasing data, and it ensures the accurate flow of the questionnaire.

## **Sample Design**

Traditionally, the sample for the NHIS is redesigned and redrawn about every ten years to better measure the changing U.S. population and to meet new survey objectives. A new sample design for the NHIS was implemented in 2006 and will be in effect for several years. The fundamental structure of the new 2006 NHIS sample design is very similar to the previous 1995-2005 NHIS sample design. State-level stratification is retained in the new NHIS sample design, which allows use of the NHIS for producing state estimates and for possible *future* dual-frame surveys at the state level. (Users should note that the current NHIS sample size is not sufficient to provide reliable state-level estimates for most states; however, for those states with larger populations, reliable state-level estimates can be produced. Since state identifiers are not

publicly released, use of that information can be made through the NCHS Research Data Center. Contact the NCHS Research Data Center for more information, or visit their Web page: <http://www.cdc.gov/nchs/r&d/rdc.htm> ) The new sample design reduced the NHIS sample size by about 13%.

Oversampling of the black and Hispanic populations was retained in the 2006 redesign to allow for more precise estimation of health characteristics in these growing minority populations. The new sample design also oversamples the Asian population. In addition, the sample adult selection process has been revised so that when black, Hispanic, or Asian persons aged 65 years or older are present, they have an increased chance of being selected as the sample adult. See Appendix III for more details.

NCHS survey integration and follow-back surveys continue to be facilitated by an area frame with independent address lists; while the area frame is based on Census 2000, the address lists are obtained in a separate listing activity, explicitly for the NHIS. Also, the NHIS sample continues to be divided into four individually representative panels to further facilitate integration with other NCHS surveys and to allow for sample size reductions while retaining representativeness. The NCHS report describing the 1995-2005 design, Series 2 - Number 130, provides much information that still applies to the new sample design. This publication is available on-line at: <http://www.cdc.gov/nchs/products/pubs/pubd/series/sr02/130-121/130-121.htm>. A new report providing a complete description of the 2006 NHIS sample design is being developed.

### **Sample Reductions in the 2007 National Health Interview Survey**

As in 2002, 2003, 2004, and 2006, the 2007 National Health Interview Survey (NHIS) was faced with a budget shortfall. As a result, NCHS and the Division of Health Interview Statistics (DHIS) decided to reduce the size of the 2007 NHIS sample. The goal of the 2007 sample cuts was strictly monetary savings. The NHIS sample was reduced by approximately 50% during July-September 2007. The 2007 sample reduction was implemented in the same way and during the same time of year as the 2006 sample reduction. Overall, about 13% of the households in the 2007 NHIS sample were deleted from interviewers' assignments. This cutback was in addition to the ongoing 13% reduction due to the new sample design that was implemented in 2006.

### **Weighting Information**

The sample is chosen in such a way that each person in the covered population has a known non-zero probability of selection. These probabilities of selection, along with adjustments for nonresponse and post-stratification, are reflected in the sample weights that are provided in the accompanying data files.

Since the NHIS uses a multistage sample designed to represent the civilian noninstitutionalized population of the United States, it is necessary to utilize the person's basic weight for proper analysis of person record data. In addition to the design and ratio adjustments included in the calculation of the Person-Level file's basic weights, the person weights are further modified by adjusting them to Census control totals for sex, age, and race/ethnicity populations (post-stratification). Beginning in 2003, NCHS made the transition to weights derived from the 2000 census-based population estimates. See Appendix VI for more detailed information.

Each file has one or more sets of weights based on the unit of analysis. Two sets of weights are provided on the Person-Level file:

Weight - Final Annual (WTFA) is based on design, ratio, non-response and post-stratification adjustments. This should be used in most analyses of the Family/Person data. National estimates of all person-level variables can be made using these weights.

Weight - Interim Annual (WTIA) does not include the post-stratification adjustment (age-sex-race/ethnicity adjustment to Census population control totals). It is required by some software packages for variance estimation for surveys with complex sample designs.

Two sets of weights are also included on the Sample Child data file:

Sample Child Weight - Final Annual (WTFA\_SC) includes design, ratio, non-response and post-stratification adjustments for sample children. National estimates of all sample child variables can be made using these weights.

Sample Child Weight - Interim Annual (WTIA\_SC) does not include the post-stratification adjustment (age-sex-race/ethnicity adjustment to Census population control totals). It is required by some software packages for variance estimation for surveys with complex sample designs.

The Sample Adult data file contains two sets of weights:

Sample Adult Weight - Final Annual (WTFA\_SA) includes design, ratio, non-response and post-stratification adjustments for sample adults. National estimates of all adult sample variables can be made using these weights.

Sample Adult Weight - Interim Annual (WTIA\_SA) does not include the post-stratification adjustment (age-sex-race/ethnicity adjustment to Census population control totals). It is required by some software packages for variance estimation for surveys with complex sample designs.

In addition, two sets of weights are provided on the Household file:

Weight - Final Annual Household (WTFA\_HH) includes the probability of selection and non-response adjustments. This weight does not include a post-stratification adjustment to Census control totals for the number of civilian, non-institutionalized households in the U.S. because suitable control totals do not exist. Non-responding households have a zero weight in this field. WTFA\_HH is the appropriate weight to use when analyzing only responding households.

Weight - Interim Annual Household (WTIA\_HH) reflects the probability of household selection. It does not include non-response or post-stratification adjustments. WTIA\_HH is the appropriate weight to use when analyzing all households in the file, responding or nonresponding.

Lastly, the Family-Level weight is discussed in greater detail in the section of this document pertaining to the Family file.

**NOTE:** Analysts should be aware that 260 persons who were active duty members of the Armed Forces at time of interview are on the Person-Level file and will appear in the unweighted frequencies, despite the fact that NHIS covers only the civilian noninstitutionalized household population. These active duty members of the Armed Forces are included in that file because at least one other family member is a civilian eligible for the survey. The value of the final annual person weight (WTFA) for these military persons is zero, so they will not be counted when making national (i.e., weighted) estimates. Data for these Armed Forces members are included in all relevant files in order to aid any analyses pertaining to family structure or relationships. No active duty Armed Forces members were selected as sample adults.

### Recall Period and Weights

Some questions for particular events have recall periods referring to, for example, the “last 2 weeks” or the “last 3 months.” In general, annual estimates of events can be made using these types of variables. For example, for a variable that counts events experienced by a person within a two-week recall period, (variable) (26) (WTFA) = annual estimate; for a variable with a three-month recall period, (variable) (4) (WTFA) = annual estimate. This assumes that the average rate of occurrence is the same over the last year as over the last two weeks (or three months). Analysts are cautioned to check the accompanying file documentation and the questionnaire in order to ensure that annual estimates for these kinds of event variables are possible and have intrinsic meaning. Annual estimates of **events** should not be interpreted as annualized **person** experiences.

### Variance Estimation

The data collected in the NHIS are obtained through a complex sample design involving stratification, clustering, and multistage sampling. Because of this complex design and adjusted

sampling weights, the direct application of standard statistical analysis methods for estimation and hypothesis testing to unweighted data may yield misleading results. If data are not weighted, severely biased estimates may result. For this reason, as indicated previously, it is necessary to use the weights that are included in the accompanying data file for analyses.

Weighted data used in standard software packages may provide unbiased point estimates for commonly computed first-order statistics like means or regression coefficients, but the computed standard errors of the estimates may be too small. Also, standard packages may produce hypothesis test results (such as p values) that are incorrect. Hence, it is recommended that users of NHIS data utilize computer software that provides the capability of variance estimation and hypothesis testing for complex sample designs. NCHS uses the software package SUDAAN<sup>R</sup> (Research Triangle Institute 2004) with Taylor series linearization methods for NHIS variance estimation. Appendix III provides SUDAAN code and a description of its use to compute standard errors of means, percentages and totals with the 2007 NHIS database. Appendix III also provides example code for SPSS, Stata, R, SAS survey procedures, and VPLX.

Analyses of large NHIS domains usually produce reliable estimates, but analyses of small domains may yield unreliable estimates, as indicated by their large variances. The analyst should pay particular attention to the coefficient of variation (relative standard error) for estimates of means, proportions and totals. In addition, small sample sizes, or small numbers of primary sampling units containing targeted data, may be an indication of estimates lacking precision.

### **General Information about the 2007 Data**

The interviewed sample for 2007 consisted of 29,266 households, which yielded 75,764 persons in 29,915 families. The interviewed sample for the Sample Child component, by proxy response from a knowledgeable adult in the family, was 9,417 children under 18 years of age. Data were not collected on any infant who was born during the assignment week of the interview. The interviewed sample for the Sample Adult component, which required self-response to all questions unless the sample adult was physically or mentally unable to do so, was 23,393 persons 18 years of age and older. There were 287 cases where a knowledgeable proxy answered for the sample adult.

The total household response rate was approximately 87.1%: 8.0 percentage points of the noninterview rate (12.9%) were the result of respondent refusal and unacceptable partial interviews. The remaining 4.9 percentage points were primarily the result of failure to locate an eligible respondent at home after repeated contact attempts.

The conditional response rate for the Sample Child component was 88.4%, which was calculated by dividing the number of completed Sample Child interviews (9,417) by the total number of eligible sample children (10,658). The unconditional or final response rate for the Sample Child component was calculated by multiplying the conditional rate by the overall family response rate of 86.6%, yielding a rate of 76.5%.

The conditional response rate for the Sample Adult component was 78.3% of persons identified as sample adults. The final response rate for the Adult Sample Person component was calculated as (Overall Family Response Rate) (Sample Adult Response Rate), or (86.6%) (78.3%) = 67.8%.

Additional information about NHIS response rates can be found in Appendix I.

### **Information about the 2007 Data File Documentation**

As with previous data years, questionnaires, datasets, and related documentation for each data file is available on the NHIS Web site, <http://www.cdc.gov/nchs/nhis.htm>. The Web site provides the Survey Description Document; a Readme File containing a summary of data access instructions; Notices for Data Users; a log of release history and, if necessary, new notices about data problems or changes; Survey Questionnaires, Flashcards, the Field Representative Manual, and Survey Flowchart; Information on Co-Sponsors and Supplements; Race/Ethnicity Information; the Data Release with links to a page that contains the Family, Household, Person, Injury/Poisoning Episode, Injury/Poisoning Verbatim Episode, Sample Child, and Sample Adult Files; Imputed Income Files; and Summary Health Statistics reports (when available).

Each of the 2007 data release categories for Household, Family, Person, Injury/Poisoning Episode, Injury/Poisoning Verbatim Episode, Sample Child, and Sample Adult Files will include the following documents. A description of each type of document follows:

- Variable Summary Report
- Variable Layout Report
- Variable Frequency Report
- ASCII data
- Sample SAS statements
- Sample SPSS statements
- Sample Stata statements

The Variable Summary Report lists each variable, a brief description of the variable, the question number on which it was based, and 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, instrument variable name, universe, response values, and response value labels. Additional specific information is provided under “Sources,” “Recodes,” “Keywords,” and “Notes.” These terms are defined below:

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

Recodes - A *recode* is a variable derived from the reordering, collapsing, or verbatim coding of another variable, such as the family income recode (INCGRP) found in the Family File. Alternatively, a recode may be constructed from two or more variables, as

is the body mass index (BMI) variable included in the Sample Adult file. 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 respondent has worked at his/her job - 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. Some of the standardized time unit recodes may also be top-coded for confidentiality reasons.

**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. Currently, there are two generic notes that can appear in addition to specific information:

- 1) If the original questionnaire item was asked at the family level but resulted, after the editing process, in a person-level variable, this note is added: Family/person variable conversion
- 2) If other questions in the instrument ask about the same topic, or if similar questions appear in other sections of the instrument, this note is added: Refer to {variable name and section number} for a {family/person/child} level question on the same topic.

The universe refers to those respondents deemed eligible to answer a given question. For example, the universes for most Sample Adult variables are specified as `ASTATFLG = 1` and `(AGE GE '018' and AGE not IN ('997', '999'))`, followed by any other universe descriptors specific to the variable. `ASTATFLG = 1` refers to a variable on the Person file and indicates that the respondent was selected as a sample adult and answered at least the first three sections of the Sample Adult questionnaire (constituting a completed interview or an acceptable partial interview). Sample adults who are not eligible to answer a given question are considered to be not-in-universe. For example, a sample adult who reported that he did not have surgery in the past 12 months (`ASRGYR=2`) would not be eligible for a follow-up question (`ASRGNOYR`) about the number of times that he had surgery in the past 12 months. **It is important to note that for all data files, persons who are not-in-universe are no longer listed in the Variable Layout Report response categories as "Blank- Not-in-universe."** If a respondent discontinued the interview anytime after completing the first three sections of the Sample Adult component, his or her responses will appear as 8's (not ascertained) for the remaining variables in the Sample Adult file where the universes are applicable. **In addition, each year, there may be a few records (less than 10) where age is corrected due to data entry error. For the**

**records where age is corrected, neither the universes nor the variables affected will be changed; however, a new variable, AGE\_CHG, will indicate that a correction has been made on the record. Occasionally universe inconsistencies between variables may exist due to collection or processing errors.**

The universes for most Sample Child file variables are specified as CSTATFLG = 1 and (AGE LE '017' and AGE NE ' '), followed by any other universe descriptors specific to the variable. CSTATFLG = 1 refers to a variable on the Person file that indicates a selected sample child with a completed interview or an acceptable partial interview (completion of the CHS section, or about half the questions of the Sample Child Core). Again, responses from acceptable partial interviews have a code of 8, meaning "not ascertained," throughout the remaining, unanswered Sample Child sections where the universes are applicable.

The Variable Frequency Report provides the frequencies, percents, and the frequency missing (not-in-universe) for each variable. **For the 2005 data year and beyond, all response categories are shown in the Variable Frequency Report, including those response categories with a zero count in the data files. This allows users to see a complete list of response categories with frequencies for each variable without referring to additional documentation. In addition, the "frequency missing" label will be shown if a variable has not-in-universe cases or cases whose values fall out of range.** For example, if all sample adults are asked about a usual place for medical care when sick (AUSUALPL), then the "frequency missing" label is not shown.

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. Lastly, in some limited situations (primarily recodes), the "Refused," "Don't know," and "Not ascertained" categories are collapsed into a single category called "Unknown," which is typically designated with a "9" (with leading 9's to fill out the field, if necessary).

In addition, statistical noise at both the variable level and record level may have been added to allow for the protection of respondent confidentiality, and, at the same time, allow for release of files with as many variables as possible.

**It is also important to note that for the 2005 data year and beyond, some frequently used variables are repeated on various data files; therefore, merging of files may be required less often than for the 2004 data year files. However, each data file contains household, family, and person numbers that make merging the files possible, if needed. Appendix VII provides sample code for merging the files.**

## Information about the 2007 CAPI Questionnaire

The NHIS CAPI questionnaire, also referred to as the CAPI Reference Questionnaire or CRQ, is an integral part of the data documentation and should be consulted when analyzing data. Users desiring greater detail should also consult the 2007 NHIS Field Representative's (FR) Manual (both the questionnaire and FR Manual are available on the NHIS Web site, <http://www.cdc.gov/nchs/nhis.htm>). Every effort was made to insure that the variable names in the data are consistent with the question items in the instrument. In a few cases, this was not possible. Users should match the question number in the instrument to the variable number in the File Layout Report to resolve any discrepancies.

Because the questionnaire for the NHIS is administered by computer, the questionnaire exists as a long and complex computer program. While stringent quality control measures were applied, a few errors are known to have occurred in the program. Instrument problems were identified over the course of the year, and efforts were made to correct these errors. Some of these problems were resolved through correction of skip patterns, question wording changes, addition of questions, or other internal instrument corrections.

When errors were detected and diagnosed, and time permitted, the instrument was changed to correct for the errors. In 2007, instrument changes were kept to a minimum, so that there was basically one version of the NHIS in the field across all four quarters of the survey year. Analysts are encouraged to read the notes in the Variable Layout Report for important information pertaining to specific variables that may have changed across quarters.

### Questionnaire Sections

The 2007 NHIS contained the annual Basic Module, which is broken into various sections that group questions into broad and specific categories. Each section is designated by a section title and corresponding three-digit acronym (or section code); questionnaire items are numbered sequentially (but not consecutively) within their respective sections, with the section acronym making up part of the item number. Multiple-part questions have an extension added to their three-digit acronym. For example, the first item in the FHS section is identified as FHS.010\_00.000; note that FHS.010\_00.000 also has an associated variable name, PLAPLYLM. The following list details the various questionnaire sections, their acronyms and description titles.

**Table 1. 2007 NHIS Core Questionnaire Sections and Topics**

#### A. Household

Section No.	Section Code	Description
I	HHC	Household Composition

## B. Family Core

<b>Section No.</b>	<b>Section Code</b>	<b>Description</b>
<b>I</b>	FID	Family Identification and Verification
<b>II</b>	FHS	Health Status and Limitation of Activity
<b>III</b>	FIJ	Injury/Poisoning
<b>IV</b>	FAU	Health Care Access and Utilization
<b>V</b>	FHI	Health Insurance
<b>VI</b>	FSD	Socio-demographic
<b>VII</b>	FIN	Income and Assets

## C. Sample Child Core

<b>Section No.</b>	<b>Section Code</b>	<b>Description</b>
<b>I</b>	CID	Child Identification and Verification
<b>II</b>	CHS	Conditions, Limitation of Activity and Health Status
<b>III</b>	CAU	Health Care Access and Utilization

## D. Sample Adult Core

<b>Section No.</b>	<b>Section Code</b>	<b>Description</b>
<b>I</b>	AID	Adult Identification and Verification
<b>II</b>	ASD	Demographics
<b>III</b>	ACN	Conditions
<b>IV</b>	AHS	Health Status and Limitation of Activity
<b>V</b>	AHB	Health Behaviors
<b>VI</b>	AAU	Health Care Access and Utilization
<b>VII</b>	ADS	AIDS

## E. Recontact

Section No.	Section Code	Description
I	REC	Recontact Information and Follow-up

In addition to the three Core sections comprising the Basic Module, the 2007 NHIS contains several other data files: the Household- and Family-level files, the Injury/Poisoning Episode file, and the Injury/Poisoning Verbatim Episode file. The Household file is derived largely from the Household composition section of the Basic Module and describes characteristics of each household. The variables contained in the Family-level file are reconstructions of the person-level data from the Basic Module sections at the family level. The Injury/Poisoning files are derived from information obtained from the injury/poisoning questions in the Family Core section

### **Supplements, Supplement Co-Sponsoring Agencies, and Question Locations, 2007 NHIS**

The terms “supplement” and/or “supplementary questions” refer to any co-sponsored questions that are in the NHIS for a year (or more) at a time. Beginning in 1997, co-sponsored questions were referred to as a “topical module” or “periodic module,” but these terms proved to be neither mutually exclusive nor exhaustive of the possible types of supplements. Therefore, effective 2001, we use the terms “supplement” or “supplementary questions” to describe co-sponsored questions.

In 2007 NHIS supplementary questions on adult immunization, child immunization, child mental health, and child mental health services are found in the Core questionnaires. A supplement or one or more supplementary questions may be interwoven among core questions, or may be placed at the end of a core section. The existence of three extra digits (.xxx) at the end of the question number helps to identify supplementary questions in the Core questionnaires. The Adult Immunization, Child Immunization, Child Mental Health, and Child Mental Health Services Supplements are released in the same files as other Core data.

There are two additional supplements in 2007. The Adult Complementary and Alternative Medicine Supplement (ALT) and the Child Complementary and Alternative Medicine Supplement (CAL) are found in separate sections in the questionnaire and are released in two separate stand-alone data files.

A chart of all 2007 co-sponsored supplements and their question numbers is below. In addition, users can obtain information about co-sponsored supplements from 1997-2007 on our Web site: <http://www.cdc.gov/nchs/about/major/nhis/co-sponsors.htm>

**Supplement Co-Sponsoring Agencies and Question Location, 2007 NHIS**

<b>Topic</b>	<b>Co-sponsoring Agency</b>	<b>Title</b>	<b>Survey Section/Number</b>
Child Alternative Medicine	National Center for Complementary and Alternative Medicine (NCCAM) <sup>1</sup>	Child Complementary and Alternative Medicine Supplement	CHS.110_01.010-CHS.110_08.080; CHS.112_01.010-CHS.112_15.150; CHS.120_01.010-CHS.120_17.170; CHS.370_01.010-CHS.375_02.020; CAL.010-CAL.486
Adult Alternative Medicine	Office of Dietary Supplements (ODS) <sup>1</sup>	Adult Complementary and Alternative Medicine Supplement	ACN.031.01.010-ACN.031_06.060; ACN.121_01.010-ACN.128_00.040; ACN.297_00.010; ACN.372_00.010-ACN.372_00.040; ALT.010-ALT.974
Hearing	National Institute on Deafness and Other Communication Disorders (NIDCD) <sup>1</sup>	Hearing Supplement	ACN.400_00.010 – ACN.415_00.040
Mental Health	National Institute of Mental Health (NIMH) <sup>1</sup>	Child Mental Health	FHS.065; CAU.265; CMB.010– CMB.030
Mental Health Services	Center for Mental Health Services (CMHS) <sup>3</sup>	Child Mental Health Services	CMS.010 – CMS.080
Immunization	National Immunization Program (NIP) <sup>2</sup>	Adult /Child Immunization	AAU.310 - AAU.380; CFI.010 – CFI.025_02
<b>Agencies Providing General Support</b>			
n/a	Center for Mental Health Services (CMHS) <sup>3</sup>	Collection and Analysis of Mental Health Data using NHIS	
n/a	Agency for Healthcare Research and Quality (AHRQ)	Use of 2 NHIS sample panels to support the MEPS	
n/a	National Center for Health Statistics <sup>2</sup>	Cell phone usage	

<sup>1</sup> National Institutes of Health (NIH)

<sup>2</sup> Centers for Disease Control and Prevention (CDC)

<sup>3</sup> Substance Abuse & Mental Health Services Administration (SAMHSA)

## **2007 National Health Interview Survey Household-Level File**

The Household file is considered as the base file from which all other files are built. That is, the main sampling unit in the NHIS is the household, and each record on the Household file represents an eligible sampling unit. Each record on the Household file represents a unique household included in the NHIS sample or sampling frame. Each household has a unique unit number (HHX). This unique unit number is needed for merging data files.

Some of the variables found only in this file include: the nature/reason for “Type A” non-responses and number of responding and non-responding families and persons. (For information about Type A non-response, see Appendix I.) Variables in other NHIS data files that may be appropriately analyzed at the household level can be merged with this file for analysis.

The universe for the Household file is all eligible households, including both responding households and non-responding (Type A) households. The Household file contains information on 33,615 households: 29,266 households were interviewed, while 4,349 were not interviewed. The nature of non-interviews for Type A households, such as refusal or failure to locate an eligible respondent, is detailed in the variable NON\_INTV.

The total non-interview rate for the Household file was 12.9% of households. The response rate for the Household file is calculated as the number of responding households divided by the total number of eligible households (responding + non-responding households), or 87.1%.

## 2007 National Health Interview Survey Family-Level File

The Family-Level file contains variables that describe characteristics of the 29,915 families living in households that participated in the 2007 NHIS. A family is defined as 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, such as unmarried couples who are living together. Each record in the file represents a unique family. The universe for all variables in this file is limited to all responding families in those households participating in the 2007 survey; this is specified as FM = ALL in the Family-Level file Variable Layout. Note that multiple families may share one household. Users wishing to determine the number of responding and non-responding families in each household are referred to ACPT\_FAM and REJ\_FAM in the Household file or HHX and FMX in the Family file.

As Table 2 indicates, 98% of NHIS households consist of one family. All relationships in the household are recorded relative to a household reference person, who is generally the person who owns or rents the housing unit. Note that when there is only one family per household, all household and family relationships (as indicated by the Person file variables RRP and FRRP, respectively) will be identical.

**Table 2. Number of Families per Household, 2007 NHIS (unweighted counts)**

<u>Families per household</u>	<u>Frequency</u>	<u>Percent</u>
1	28,771	98.3
2	381	1.3
3	88	0.3
4	17	0.1
5	5	0.0
6	3	0.0
7	1	0.0

In the small number of instances where there is more than one unrelated family living in a single household, the various NHIS questionnaires (e.g., Family Core, Sample Adult Core, etc.) will then be administered separately to *each* family within the sampled household. Moreover, one household reference person is chosen for the housing unit *and* one family reference person is designated for each distinct family within the household. Each family in the household will thus have the same household reference person but a *different* family reference person, and all relationships in both the household and the family will be described relative to these two persons. Examples of multi-family households include several unrelated roommates sharing a house or apartment; a family with an unrelated lodger and his/her child; a family with a live-in housekeeper and his/her spouse; etc.

Family size may vary considerably. Table 3 shows a breakdown of the 29,915 families by number of family members.

**Table 3. Size of Family, 2007 NHIS (unweighted counts)**

<u>Number of Members</u>	<u>Frequency</u>	<u>Percent</u>
1	8,849	29.6
2	9,032	30.2
3	4,622	15.5
4	4,132	13.8
5	2,034	6.8
6	778	2.6
7	270	0.9
8	124	0.4
9	36	0.1
10	22	0.1
11	8	0.0
12	4	0.0
13	1	0.0
14	2	0.0
16	1	0.0

The first part of the Family file contains the technical variables that identify or describe the record type (all observations in this file have a record type value of “60”), the survey year, the household and family numbers, the interview month and year, characteristics of the family’s housing unit, geographic information associated with the housing unit, variables used for variance estimation, and a family-level weight variable.

The second part of the file consists of a series of recodes derived from five Family Core sections of the NHIS that collapse the individual level observations into information about their respective families.

Generally, the Family file consists of two types of recodes. The first is a simple “yes-no” measure that indicates whether *any* family member falls into a particular category or exhibits a particular characteristic. Every yes-no measure also has a corresponding counter that indicates the number of family members in that category or with that characteristic. Note that counters always consist of values from zero to 30; in addition, no frequencies will be shown if a family is not contained in the universe for a specific question. For example, FSALYN and FSALCT, two recodes from the Income and Assets section of the Family Core, are limited to families with at least one member aged 18 or older; families consisting solely of emancipated minor(s) are coded as blanks to indicate that they are out of the universe, and thus, are not shown. The Family file also contains some counters that lack corresponding yes-no indicators.

For example, FHSTATEX, FHSTATVG, FHSTATG, FHSTATFR, and FHSTATPR (all derived from PHSTAT, FHS.500) provide counts of the number of family members in excellent, very good, good, fair, and poor health, respectively. Counters were also constructed to indicate the number of working adults in the family, the number of adults in the family looking for work, the number of adults working full time, the number of children (under age 18) in the family, and the number of family members aged 65 and older.

Because most of the variables in the Family file are recodes of the person-level variables in the family core, the sum of the number of persons across all families in each family-level counter should be equal to the number of “yes” responses in its person-level source. Returning to our previous example, consider FSALCT: 11,973 families have one member receiving income from wages/salary, 8,420 families have two members (or  $2(8,420)=16,840$  persons) with wage/salary income, 1,364 families have three members (or  $3(1,364)=4,092$  persons), 384 families have four members (or  $4(384)=1,536$  persons), 67 families have five members (or  $5(67)=335$  persons), 12 families have six members (72 persons), 1 family has seven members (7 persons), and 1 family has eight members (8 persons) with wage/salary income in 2007. Thus, the sum of persons across the 22,222 families answering “yes” to FSALYN, the associated yes-no indicator, is 22,222 ( $11,973 + 8,420 + 1,364 + 384 + 67 + 12 + 1 + 1$ ), which is equal to the 34,863 “yes” responses to the person-level source variable, PSAL. Users are advised to check the Variable Layout Report for each Family file recode in order to determine its person-level source variable.

### Family Structure Variables

The 2007 NHIS Family file contains two variables describing family type and structure in both general and detailed terms. FM\_TYPE consists of just four categories, and represents an initial classification of families according to the numbers of adults and children that are present. In addition, FM\_STRP and FM\_STRCP categorize families according to familial relationships and, when children are present, parental marital status. FM\_STRP and FM\_STRCP differ in how they categorize unmarried parents with children. FM\_STRP includes all cohabiting couple families in the same category (FM\_STRP = 42), regardless of the adults’ relationships to the child(ren) in the family. FM\_STRCP is identical to FMSTRCT2, a recode on the 1998, 2001-2003 NHIS Family files, and distinguishes between families consisting of unmarried parents who are related biologically or by adoption to all children in the family (FM\_STRCP = 41), and families consisting of a parent, his or her child(ren), and his or her partner, who is unrelated to the child(ren) present in the family (FM\_STRCP = 43). In both recodes, families that could not be classified are coded “99.” Emancipated minors are treated as adults with respect to FM\_TYPE, FM\_STRP, and FM\_STRCP, despite the fact that they may be under 18 years of age.

## The Family File Weight

The ideal situation for creating weights for the Family file would be to use independent estimates of the number of families from a reliable source, such as the U.S. Census Bureau, to perform post-stratification adjustments in a manner similar to what is done for the NHIS Person file weight. Unfortunately, no suitable independent estimates exist.

Due to the lack of appropriate independent estimates, a variation of the “principal person” method is used to create the 2007 NHIS Family file weight (WTFA\_FAM). Our method is similar to that used in the Current Population Survey to create their household- and family-level weights. Briefly, a person-level ratio adjustment is used as a proxy for the NHIS family-level ratio adjustment. Use of the person weight with the *smallest* ratio adjustment within each family (that is, the smallest post-stratification factor between the interim and final person weights within the family) is believed to provide a more accurate estimate of the total number of U.S. families than either the use of other person weights in the family or the use of no ratio adjustments whatsoever.

Accordingly, the weight provided with the 2007 NHIS Family file, WTFA\_FAM, corresponds to the 2007 NHIS person weight for one of the persons in the family. As a result, the Family weight contains factors for selection probabilities at the household level, household nonresponse adjustment, and several ratio adjustment factors that are applied to all person weights.

## **2007 National Health Interview Survey Person-Level File**

The Person-level variables are derived from the six substantive sections making up the Family Core of the 2007 NHIS. The information in the Family Core questionnaire is collected for all household members. Any adult household members who are present at the time of the interview may take part; information regarding adults not participating in the interview, as well as about all household members under age 18, is provided by a knowledgeable adult member of the household. (If there is more than one family in the household, then these procedures are followed for each family in the household. See the Family-Level File for more information.) The six sections comprising the Family Core are discussed in greater detail below.

### **I. Health Status and Limitation of Activity Section (FHS)**

The Health Status and Limitation of Activity (FHS) section of the Family Core for the 2007 NHIS contains information addressing respondent-assessed disabilities, disability-associated conditions, and overall health status for all family members. Users should note that additional information on health status and disability is also included in other sections of the Sample Adult file, as well as in the Sample Child file.

#### **Limitation of Activity at the Person Level**

Information on activity limitations, including questions about work limitations; the need for personal assistance with personal care needs such as eating, bathing, dressing, and getting around inside the home; and the need for personal assistance with handling routine needs such as everyday household chores, doing necessary business, and shopping or running errands, is collected for each family member (with some exclusions for children and youth). If any limitations are identified, the respondent is asked to specify the health condition(s) causing the limitation(s) and indicate how long he or she has had each such condition.

Since cognitive impairment is increasingly recognized as a source of activity limitations among older adults, the FHS section includes an indicator that identifies family members who are limited because of difficulty remembering or periods of confusion. Other indicators in this section identify family members who have difficulty walking without any special equipment or limitations related to specific personal care needs. In addition, the section contains information about children who receive special education or early intervention services. Information regarding limitations in play activities is also collected for young children.

The 2007 FHS time variables and recodes, which indicate how long respondents have had the condition(s) causing their limitation(s), were processed using procedures similar to those used in 2002-2006. Substantively, the 2002-2007 variables and recodes are similar to those from previous years (1997-2001), but the detailed unknown categories that were included in the earlier data were collapsed into broader categories starting in 2002.

## Conditions

For each family member with a previously mentioned limitation, the respondent was asked about the condition or health problem associated with that limitation, as well as the length of time he/she has had the condition. Respondents were then handed one of two flash cards listing various condition categories. These categories are broad in scope, and vary according to age. Information about family members under age 18 was solicited for the following fixed condition categories listed on the first flash card: “vision/problem seeing,” “hearing problem,” “speech problem,” “asthma/breathing problem,” “birth defect,” “injury,” “mental retardation,” “other developmental problem (e.g., cerebral palsy),” “other mental, emotional, or behavioral problem,” “bone, joint, or muscle problem,” “epilepsy or seizures,” “learning disability,” “attention deficit/hyperactivity disorder,” and two instances of “other impairment problem” (if the family member was limited by a condition not listed in one of the fixed categories). Respondents could supply a verbatim response of up to 50 characters for one or both of the “other impairment problem” categories.

The fixed response categories in the instrument for adults age 18 or older were equally broad, and comprised the conditions listed on the second flash card: “vision/problem seeing,” “hearing problem,” “arthritis/rheumatism,” “back or neck problem,” “fractures, bone/joint injury,” “other injury,” “heart problem,” “stroke problem,” “hypertension/high blood pressure,” “diabetes,” “lung/breathing problem,” “cancer,” “birth defect,” “mental retardation,” “other developmental problem (e.g., cerebral palsy),” “senility,” “depression/anxiety/emotional problem,” and “weight problem.” Starting in 2001 and continuing in 2007, if an adult family member was limited by a condition not listed in one of these 18 fixed categories, the interviewer accessed a second screen containing 17 additional condition categories and two “other impairment problem” categories. These conditions were not read aloud to respondents, but if the respondent said a family member’s condition was limited by one of these 17 conditions, the interviewer recorded this information. If the family member was limited by a condition not listed on either the second flash card or on the interviewer’s computer screen, then the interviewer entered a verbatim response of up to 50 characters for one or both of the “other impairment problem” categories. Respondents could list any number of applicable conditions.

During data processing, the verbatim responses recorded by interviewers were reviewed to determine if any responses could be back-coded to one of the 13 fixed categories for respondents under age 18, or to one of the 18 fixed categories for adult respondents. If so, these “other” responses were assigned to the appropriate response categories (the first 13 for children, and the first 18 for adults). For adults, an additional 16 *ad hoc* categories were created during data processing to categorize responses that fell outside the fixed 18 condition categories included in the instrument: these *ad hoc* categories were assigned numbers 19\_ through 34\_. (Note: Due to a naming convention error in 2002 and 2003 these same *ad hoc* categories were assigned numbers 19 through 34 without an underscore.) In addition, responses in the 17 general categories seen only by the interviewer were also back-coded and categorized into 8 of the *ad hoc* categories. The resulting 34 output categories for adults and 13 output categories for children were based on the International Classification of Diseases, Ninth Revision, Clinical Modification (see Table 4, below; note that the ICD-9-CM codes shown in this table are *not* included on the data file).

Any verbatim conditions that could not be back-coded to one of the original categories or recoded to one of the *ad hoc* categories (for adult respondents) remained in the “other impairment problem” categories, and were renumbered “90” and, if necessary, “91” for both children and adults. The specific condition categories as well as the “other impairment problem” categories were subsequently transformed into variables indicating whether or not the condition was responsible for the respondent’s difficulty with any activity (a mention/not-mention format). Note that the verbatim responses associated with the “other impairment problem” categories are not included as a separate field on the public use file. In addition, because the 16 adult *ad hoc* categories were not included on the flash cards given to respondents during the course of the interview, it is possible that frequencies obtained for these conditions causing limitations will be underestimates. Therefore, these variables should be analyzed with care. Moreover, none of the FHS condition variables (the 13 child variables, LAHCC1 through LAHCC13, and the 34 adult variables, LAHCA1 through LAHCA34\_) should be used to estimate prevalence for the conditions they represent, because only those persons with a previously reported limitation were eligible for the condition questions that followed. Analysts who are interested in estimating the prevalence of particular conditions are referred to the Sample Adult and Child Cores.

## Recodes

The recode LA1AR is a summary measure that indicates household members reporting *any* limitation regarding one or more of the activities discussed during the course of the FHS section of the interview. In other words, respondents who answered “yes” to PLAPLYLM, PSPEDEIS, PLAADL, PLAIADL, PLAWKNOW, PLAWKLIM, PLAWALK, PLAREMEM, or PLIMANY are coded “1” for LA1AR. LA1AR includes three response levels: “1” for limited, “2” for not limited, and “3” for unknown if limited. (For comparability with previous years, level 3 may be collapsed into level 2.) LACHRONR is based on LA1AR but adds the additional criterion of whether at least one of the reported causal conditions is a chronic condition. Users can utilize the information contained in LA1AR to control for “unknown if limited” cases with respect to LACHRONR (that is, when LACHRONR = 0).

Also, a series of age-group-specific recodes (e.g., under 18 *versus* 18 and over) regarding conditions limiting activity and duration of limiting conditions have been created. Because the questions about limitation of activity in the redesigned NHIS are asked differently for different age groups, and because the questions are more general (in some cases) or more specific (in other cases) than in pre-1997 years, the degree to which a respondent is limited cannot be determined.

## Chronic Conditions

Each condition reported as a cause of an individual’s activity limitation has been classified as “chronic,” “not chronic,” or “unknown if chronic,” based on the nature of the condition and/or the duration of the condition. Conditions that are generally not cured once acquired (such as heart disease, diabetes, and birth defects in the original response categories, and amputee and “old age” in the *ad hoc* categories) are considered chronic, while conditions related to pregnancy are always considered not chronic. Additionally, other conditions must have been present for three months or longer to be considered chronic. Conditions are

considered chronic for children less than one year of age who have had a condition “since birth.” **Because the presence of a limitation determined whether persons were eligible for the condition questions and the chronicity recodes, we caution data users that these variables should *not* be used to produce estimates of prevalence rates of chronic conditions.**

**Table 4. FHS Categories with Approximate ICD-9-CM Ranges**

<b>A. Codes for Adults (ages 18 or more years)</b>	
<b><u>NHIS Category</u></b>	<b><u>ICD-9-CM Codes</u></b>
<b>1 - Vision or seeing problem</b>	360-379
<b>2 - Hearing problem</b>	387-389
<b>3 - Arthritis / rheumatism</b>	711-712, 714-716, 720.0, 721, 729.0
<b>4 - Back or neck problem</b>	722-724, 732.0, 737
<b>5 - Fractures, bone or joint injury</b> <i>Injury with specific mention of bone or joints</i>	800-848, 850-999
<b>6 - Other injury</b> <i>Injury without specific mention of bone or joints</i>	850-999
<b>7 - Heart problem</b>	410-417, 420-429, 745, 746, 785.0-785.3
<b>8 - Stroke problem</b>	430-438
<b>9 - Hypertension or high blood pressure</b>	401-405
<b>10 - Diabetes</b>	250
<b>11 - Lung or breathing problem</b>	460-461, 465-466, 470-471, 473, 477, 480-487, 490-496, 500-508, 510-519
<b>12 - Cancer</b>	140-208
<b>13 - Birth defect</b> <i>Excludes Down’s syndrome and microcephalus</i>	740-742.0, 742.2-744, 747-757.9, 758.1-759
<b>14 - Mental retardation</b> <i>Includes Down’s syndrome and microcephalus</i>	317-319, 742.1, 758.0
<b>15 - Other developmental problem</b> <i>Includes learning disabilities</i>	315, 343, 783.4
<b>16 - Senility (and other cognitive problems)</b>	290
<b>17 - Depression, anxiety or emotional problem</b> <i>Includes neurotic disorders, personality disorders, and other nonpsychotic mental disorders, excluding alcohol and drug related problems and developmental problems</i>	300-302, 306-314, 799.2
<b>18 - Weight problem</b> <i>Indicates a problem with being overweight or obese</i>	

**A. Codes for Adults (ages 18 or more years)**

<b><u>NHIS Category</u></b>	<b><u>ICD-9-CM Codes</u></b>
<b>19 - Missing limbs (any part) / amputee</b> <i>Indicates loss of a limb or digit</i>	
<b>20 - Other musculoskeletal system conditions</b> <i>Diseases of the musculoskeletal system and connective tissue not coded to 3, 4, 5</i>	710-739
<b>21 - Other circulatory system conditions</b> <i>Any diseases of the circulatory system not coded to 7, 8, 9</i>	390-459
<b>22 - Other endocrine system, etc. conditions</b> <i>Any Endocrine, Nutritional and Metabolic Diseases and Immunity Disorders not coded to 10 or 18</i>	240-279
<b>23 - Other Nervous system conditions</b> <i>Diseases of the nervous system and sense organs not coded to 1, 2, 15, 16</i>	320-389
<b>24 - Digestive system conditions</b>	520-579
<b>25 - Genitourinary system conditions</b>	580-629
<b>26 - Skin &amp; subcutaneous system conditions</b>	680-709
<b>27 - Blood &amp; blood-forming organ conditions</b>	280-289
<b>28 - Tumors &amp; cysts, benign &amp; unspecified</b> <i>Any mention of "tumor" without cancer, malignancy, etc.</i>	210-239
<b>29 - Alcohol &amp; drug related problems</b> <i>Any mention of "alcohol," "drugs" (or specific drug types), or substance abuse</i>	291-292, 303-305
<b>30 - Other mental conditions</b> <i>Any mental disorders not coded to 14 or 15 or 17</i>	293-299
<b>31 - After effects of surgery or other medical treatment</b> <i>Any mention of "surgery" or "operation" or other treatment as the causal condition; includes ongoing or recent treatment (1 year or less) or specific and sole mention of surgery/medical procedure as specific cause of limitation.</i>	
<b>32 - Old age</b> <i>Any mention of age as the only specified cause</i>	
<b>33 - Fatigue/Tiredness</b> <i>Any mention of tiredness, stiffness, or weakness without referring to any specific part of the body</i>	
<b>34 - Pregnancy related conditions</b> <i>Any mention of "pregnancy" or "childbirth"</i>	

**A. Codes for Adults (ages 18 or more years)**

**NHIS Category**

**ICD-9-CM Codes**

**90 - Others Not Elsewhere Classified**

*1st other-specify verbatim, not elsewhere classified*

**91 - Others Not Elsewhere Classified**

*2nd other-specify verbatim, not elsewhere classified*

**B. Codes for Children (ages under 18 years)**

**NHIS Category**

**ICD-9-CM Codes**

**1 - Vision or seeing problem**

360-379

**2 - Hearing problem**

387-389

**3 - Speech problem**

307.0, 307.9, 315.3, 784.3, 784.5

**4 - Asthma or breathing problem**

460-461, 465-466, 470-471, 473, 477, 480-487,  
490-496, 500-508, 510-519

**5 - Birth defect**

*Excludes Down's syndrome and microcephalus*

740-742.0, 742.2-757.9, 758.1-759

**6 - Injury**

800-999

**7 - Mental retardation**

*Includes Down's syndrome and microcephalus*

317-319, 742.1, 758.0

**8 - Other developmental problem**

343, 783.4

**9 - Other mental, emotional, or behavioral problem**

290-313, 799.2, V15.4

**10 - Bone, joint or muscle problem**

710-739

**11 - Epilepsy and seizures**

345, 779.0, 780.3

**12 - Learning disability**

315

**13 - Attention Deficit/Hyperactive Disorder (ADD/ADHD)**

314

**90 - Others Not Elsewhere Classified**

*1st other-specify verbatim that does not fit in any other category*

**91 - Others Not Elsewhere Classified**

*2nd other-specify verbatim that does not fit in any other Category*

Technical Notes

The data for five variables in the 2007 FHS section (PSPPEDEIS, PSPPEDEM, PLAIADL, PLAWKNOW, and PLAWKLIM) contain an inconsistency with the stated

universe for a single case. Users are advised to check the notes for each variable in the Variable Layout Report for guidance on this problem.

The condition variable LAHCA31\_ includes any causal condition that specifically mentioned “surgery” or “operation,” or otherwise indicates a medical treatment as the causal condition (either ongoing or occurring within the last year). The condition variable LAHCA33\_ includes any causal condition that specifically and solely mentioned “fatigue,” “weakness,” “lack of strength,” “tiredness,” “exhaustion,” etc. without reference to any particular part of the body. Lastly, the condition variable LAHCA34\_ includes any causal condition that specifically and solely mentioned “pregnancy,” “pregnant,” or “childbirth.”

## **II. Health Care Access and Utilization Section (FAU)**

The Health Care Access and Utilization (FAU) section of the Family Core of the 2007 NHIS has remained largely unchanged since 1997. The FAU section contains information addressing access to health care, utilization services, and health care contacts.

Since 1997, questions that ask about delay of health care because of worry about the cost, overnight hospital stays, home care, calls to health professionals, and office visits have been included in the survey; there is also an expanded list of health care professionals, and respondents were instructed to consider “care from ALL types of medical doctors, such as dermatologists, psychiatrists, ophthalmologists, and general practitioners,” as well as nurses, physical therapists, and chiropractors. Lastly, a question asking about 10 or more visits to doctors or other health care professionals in the last 12 months has been included.

### Technical Notes

A few large values were found for hospitalizations (HOSPNO) and hospital nights (HPNITE). In addition large numbers may exist for home care visits (PHCHMN2W), doctor visits (PHCDVN2W), and calls to health professionals (PHCPHN2W). Analysts should be aware that the above mentioned variables have not been edited for reasonableness.

Analysts are advised to read the notes in the Dataset Documentation for further information pertaining to any changes that may have occurred and to compare the 2007 documentation to documentation from the 2006 (and earlier) NHIS for any other changes that may have occurred over time to the variables in this section.

## **III. Health Insurance Section (FHI)**

The Health Insurance section of the 2007 NHIS Family Core has a full range of data items addressing health insurance. The flow of the questions pertaining to health insurance programs covered by this section is similar to the 1993-96 NHIS Health Insurance Supplements and the 1997-2006 NHIS Family Cores.

In 2007 three additional questions were added to the section. For persons with private health insurance, a new question regarding the annual deductible (HDHP) of each private health insurance plan was added. For 2007, a high deductible plan was defined as a health plan with an annual deductible of not less than \$1,100 for self-only coverage or \$2,200 for family coverage. For those plans considered to be high deductible health plans, a follow-up question (whether a health savings account or a health reimbursement account (HSAHRA) was used to pay for medical expenses) was asked. Lastly, a new question was added that asked if anyone in the family had a flexible spending account (FSA) for medical expenses.

The health insurance section (FHI) covers several different topic areas:

Type of health care coverage (Medicare, Medicaid, State Children's Health Insurance Program (SCHIP), MILITARY (TRICARE, VA, CHAMP-VA), State-sponsored health plan, Indian Health Service, other government programs, private insurance and single service plans);

Managed care arrangement and the need for referrals for those covered by Medicare, Medicaid, SCHIP, other State-sponsored health plans and other government programs;

Medicare managed care model types;

Enrollment in the Medicare Part D program;

Private insurance characteristics reported by the family respondent, including HMO, PPO, and POS status, high deductible health plan, health savings account or health reimbursement agreement for high deductible plans, source of coverage, existence of employer subsidies for premiums, amount paid by individual/family, managed care detail information, need for a referral, prescription drug benefit;

Private insurance plan types, including HMO model types coded from private plan names;

Types of single service plans;

Type of TRICARE coverage;

Periods of time without health insurance and reasons for no health insurance;

Out-of-pocket costs in the past year for medical expenses (excluding health insurance premiums);

Enrollment in a flexible spending account (FSA) for medical expenses.

Beginning in 2004, FHI data contain several modifications, as well as some new variables. The HIKIND list was shortened from 14 categories to 11. Private health insurance was combined into one category, HIKINDA, and military health coverage was combined into one category, HIKINDF. To increase the counts of single service coverage, SINCOV was added following the HIKIND question for persons who had not indicated earlier that they have a single service plan. A person who responded to either HIKINDJ or SINCOV received the single service detail questions. Response categories were changed in the PLNWRK question to get better precision as to how a private health plan was obtained. This detail is contained in PLNWRKN1 and PLNWRKN2. An additional question was added to the private plan detail to monitor the impact of the Medicare prescription drug benefit on private plan drug benefits. This information is contained in PRRXCOV1 and PRRXCOV2. Detailed information concerning the third and fourth plans for a

respondent is no longer available on the public use data file. Persons with three or more plans have a “yes” response to the PRPLPLUS variable. Detailed information on the third and fourth plan for a respondent is still available through the NCHS Research Data Center.

Details on type of military coverage are now contained in the variables MILSPC1, MILSPC2, MILSPC3, MILSPC4, and MILMAN. The wording on the MCCHOICE question was changed to address the new name for Medicare Plus Choice, which is Medicare Advantage. Follow-up questions were added regarding the State Children’s Health Insurance Program (SCHIP), State-sponsored and other public programs (OTHERPUB), and other government programs (OTHERGOV) to obtain managed care information for all types of public coverage.

Beginning with quarter 3, 2004, two new questions were added to reduce potential errors in reporting Medicare and Medicaid status. Persons 65 years and over not reporting Medicare coverage were asked explicitly about Medicare coverage in MCAREPRB. Persons under 65 with no reported coverage were asked explicitly about Medicaid coverage in MCAIDPRB. Respondents who were reclassified as covered by either of these additional questions received the appropriate follow-up questions.

#### Technical Notes

Analysts are strongly advised to use the recodes MEDICARE, MEDICAID, PRIVATE, SCHIP, IHS, MILITARY, OTHERPUB, OTHERGOV, and SINGLE for types of health care coverage because these recodes take into account the complicated editing process that takes place in the FHI section. The variables HILAST and HINOTYR, which reflect periods of noncoverage, cannot be used to estimate the rate of uninsurance. Users should derive such estimates from NOTCOV (if they do not count IHS as coverage) or, alternatively, the health insurance recodes (MEDICARE, MEDICAID, PRIVATE, SCHIP, IHS, MILITARY, OTHERPUB, and OTHERGOV). Using the most conservative estimate of the uninsured (which would exclude persons with IHS coverage only), a total of 594 persons did not receive the HILAST question during the course of the interview because they indicated that they had health care coverage. It was subsequently established during the course of editing that they lacked coverage (given the information that they provided about their insurance plan(s)). NHIS staff elected not to edit these people out of the universe for HINOTYR. In addition, a total of 759 respondents were not asked either the HILAST or the HINOTYR questions.

It was determined that some respondents indicated plans (in response to the questions HIPNAM1, HIPNAM2, HIPNAM3, and HIPNAM4) that were not private health insurance plans, or were single service plans that were excluded from the private health insurance coverage category. These respondents were reassigned to the appropriate response category with the enrollment recodes for MEDICARE, MEDICAID, SCHIP, IHS, MILITARY, OTHERPUB, OTHERGOV, and SINGLE. Similarly, in looking at the verbatim responses to the questions STNAME1, STNAME2 or STNAME3 that asks respondents for the name of their SCHIP, state sponsored or other government coverage respectively, it was found that some respondents indicated plans and names of programs that were clearly private health insurance, Medicare, Medicaid, military coverage, Indian Health Service, single service plans, or no coverage at all. Persons with these forms of coverage were reassigned to the appropriate enrollment recodes for

MEDICARE, MEDICAID, PRIVATE, IHS, MILITARY, and SINGLE. Respondents who answered “other state sponsored” or “other government coverage” who were subsequently determined through the STNAME2 or STNAME3 fields to be covered by the State Children’s Health Insurance Program were assigned to the SCHIP recode. In looking at the verbatim responses to the question MCHMO\_NA that asks respondents for the name of their Medicare managed care plan, it was found that some respondents indicated plans or programs that were clearly private health insurance, Medicaid, military coverage, Indian Health Service, single service plans, or no coverage at all. Persons with these forms of coverage were reassigned to the appropriate enrollment recodes for MEDICAID, PRIVATE, IHS, MILITARY, and SINGLE. Likewise, in looking at the verbatim responses to the questions MACHMD\_1 and MACHMD\_2 that ask respondents for the name of their Medicaid managed care plan, it was found that some of these respondents indicated plans or programs that were clearly private health insurance, Medicare, SCHIP, military coverage, Indian Health Service, single service plans, or no coverage at all. These respondents were also reassigned to the appropriate enrollment recodes for MEDICARE, CHIP, PRIVATE, IHS, MILITARY, and SINGLE. In addition, due to reengineering changes, there are no longer “99” response codes for the variables HITYPE1 and HITYPE2.

In addition, some respondents offering an “other” response to the survey item (HISTOPOT) that inquired about the reason(s) their coverage stopped subsequently indicated in their verbatim responses that they did in fact have health insurance. These persons were reassigned to the appropriate response category with the enrollment recodes for MEDICARE, MEDICAID, SCHIP, PRIVATE, IHS, MILITARY, OTHERPUB, and OTHERGOV. Analysts are therefore strongly advised to use the recodes MEDICARE, MEDICAID, PRIVATE, SCHIP, IHS, MILITARY, OTHERPUB, OTHERGOV, and SINGLE for types of health care coverage, because these take into account the above-mentioned back edits. In contrast, the data contained in HIKINDA-HIKINDK and MCAREPRB, MCAIDPRB, and SINCOV were not back-edited and reflect the respondents’ original replies. In addition, a recode (NOTCOV) is included in the data file that reflects the definition of noncoverage as used in *Health, United States, 2007* (in which persons with *only* Indian Health Service coverage are considered uninsured).

#### **IV. Socio-demographic Section (FSD)**

The Socio-demographic (FSD) section of the Family Core in the 2007 NHIS collects information on place of birth, citizenship status, and educational attainment for all family members, regardless of age. In addition, family members 18 years of age or older are asked if they were working last week, and if not, their main reason for not working. Additional questions inquired about the number of hours they worked during the previous week, whether their employer provided health insurance, and, if they worked during the previous calendar year, how many months they worked and an estimate of their earnings from wages. Analysts may also refer to the Adult Core Socio-demographic section (ASD) for additional occupational and employment data regarding those individuals selected as sample adults.

DOINGLWP and WHYNOWKP are the FSD equivalents of DOINGLWA and WHYNOWKA in the ASD section of the Sample Adult data file. For the majority of respondents,

DOINGLWP and DOINGLWA will have identical values (and, likewise, WHYNOWKP and WHYNOWKA). However, it is nevertheless possible that DOINGLWP and DOINGLWA (and WHYNOWKP and WHYNOWKA) may have inconsistent values across the Person and Sample Adult data files. Users wishing to reconcile any discrepant values are advised to use the values of DOINGLWA and WHYNOWKA (rather than DOINGLWP and WHYNOWKP, respectively), since the information obtained from the family respondent during the FSD portion of the interview (and reflected in DOINGLWP and WHYNOWKP) was subsequently confirmed or corrected by the sample adult during his or her interview (as reflected in DOINGLWA and WHYNOWKA). Additionally, both DOINGLWP and WHYNOWKP are substantively equivalent to previous years' versions of these variables (i.e., DOINGLW1 and WHYNOWK1).

The 2007 FSD section contains a variable called PLBORN, which is based on a question in the instrument that asked whether the respondent was born in the United States. If respondents replied affirmatively, they were asked the state in which they were born (PLBORN1). If respondents said they were not born in the U.S., they were asked the country in which they were born (PLBORN2). PLBORN1 and PLBORN2 are not included on the public use file for confidentiality reasons. However, the 2007 NHIS includes two public use recodes, GEOBRTH and REGIONBR, that are based on this restricted birthplace information (as well as the variable, PLBORN). GEOBRTH indicates geographic place of birth, and has three categories: born in one of the 50 United States or the District of Columbia; born in a U.S. territory; or not born in the U.S. or a U.S. territory. In order to make GEOBRTH comparable to previous recodes (for carrying out analyses on multiple years of NHIS data), users should collapse those respondents in the last two categories of GEOBRTH into a single category. This will result in a recode that is comparable to USBRTH\_P from the 2000-2001 NHIS or USBORN\_P from the 1997-1999 NHIS. The second recode, REGIONBR, categorizes all respondents into one of 12 categories depending on their country of origin. The CIA on-line World Factbook was used to place countries into the regional categories shown below (for more information about the Factbook, users should refer to <http://www.cia.gov/cia/publications/factbook/index.html>). Note that respondents born in Canada were included in the "Elsewhere" category of REGIONBR in order to satisfy NCHS confidentiality requirements. Users are cautioned that neither GEOBRTH nor REGIONBR indicate legal status or citizenship.

<b>Category</b>	<b>Countries/regions included</b>
United States	The 50 United States and the District of Columbia
Mexico, Central America, Caribbean Islands	Mexico, all countries in Central America and the Caribbean Island area, including Puerto Rico
South America	All countries on the South American continent
Europe	Albania, Austria, Azores Islands, Belgium, Bosnia, Bulgaria, Corsica, Crete, Croatia, Czechoslovakia, Denmark, Finland, France, Germany, Great Britain, Greece, Herzegovina, Holland, Hungary, Iceland, Ireland, Italy, Liechtenstein, Luxembourg, Macedonia, Majorca, Malta, Monaco, Montenegro, Netherlands, Norway, Poland, Portugal, Prussia,

	Romania, Scotland, Serbia, Sicily, Slovakia, Spain, Sweden, Switzerland, Yugoslavia
Russia (and former USSR areas)	Russia, Lithuania, Latvia, Ukraine, and all places formerly a part of the USSR
Africa	All countries on the African continent, plus the Canary Islands, Comoros, Madagascar, Madeira Islands
Middle East	Aden, Arab Palestine, Arabia, Armenia, Bahrain, Cyprus, Gaza Strip, Iran, Iraq, Israel, Jordan, Kuwait, Syria, Lebanon, "Middle East," Oman, Palestine, Persia, Qatar, Saudi Arabia, Syria, Turkey, United Arab Emirates, West Bank, Yemen
Indian Subcontinent	Afghanistan, Bangladesh, Bhutan, British Indian Ocean Territory, Ceylon, East Pakistan, India, Maldives, Nepal, Pakistan, Sri Lanka, Tibet, West Pakistan
Asia	Asia, Asia Minor, China, Japan, Mongolia, North Korea, South Korea
SE Asia	Borneo, Brunei, Burma, Cambodia, Christmas Island, Hong Kong, Indonesia, Laos, Malaysia, Myanmar, Philippines, Singapore, Taiwan, Thailand, Vietnam
Elsewhere	Guam, Bermuda, Canada, Greenland, Oceania, as well as "At sea," "High seas," "International waters," "North America"
Unknown	Places that could not be classified in the above categories.

Users seeking more detailed information on respondents' place of birth may gain limited, supervised access through the NCHS Research Data Center. For more information, refer to the Research Data Center Web page: <http://www.cdc.gov/nchs/r&d/rdc.htm> .

Respondents who were not born in one of the 50 United States or the District of Columbia were asked the year in which they came to the United States to stay. Respondents who could not recall or refused to answer were subsequently asked to estimate the number of years they had been in the United States since they came to stay. This information was combined to create a recode that indicates how long these respondents have been living in the United States (YRSINUS). The 2007 data also contain a citizenship recode (CITIZENP) that distinguishes between U.S. citizens and non-citizens.

#### Technical Notes

The data for eight variables in the 2007 FSD section (USYR, USLONG, and YRSINUS, HEADSTV1, PMILTRY, DOINGLWP, WHYNOWKP, and WRKLYR1) contain an inconsistency with the stated universe for a single case. Users are advised to check the notes for each variable in the Variable Layout Report for guidance on this problem.

A small number of respondents received a “1” or “2” on CITIZENP despite the fact that they answered “don’t know” or “refused” on PLBORN. (This inconsistency also occurred in the 2004 and 2005 NHIS data.) Analysts may want to reconcile these cases before using CITIZENP in their analyses.

## **V. Income and Assets Section (FIN)**

The Income and Assets (FIN) section of the Family Core contains information regarding a variety of income sources, as well as estimates of total combined family income and home tenure status. Respondents are asked whether anyone in the family received income from a variety of sources; if so, the respondent is then asked to name the member(s) receiving income from that source. The section also includes questions about the family’s total income from all sources in 2006, and their home tenure status. The basic universe for most questions is “all families;” however, note that universes for several questions (most importantly, PSAL, PSEINC, and PWIC) are further limited with respect to age (of family members).

### Sources of Income

The first two questions in the section ask about income from wages and salary, and from self-employment (business or farm) for family members 18 years of age and older. Subsequent questions are not limited to adult family members. Respondents were asked about income from Social Security or Railroad Retirement (including that which was received as a disability benefit); other pensions; Supplemental Security Income (SSI); Welfare/Temporary Assistance to Needy Families (TANF); other kinds of government assistance (e.g., job training or placement, transportation assistance, or child care); interest from checking or savings accounts, Individual Retirement Accounts (IRAs) or certificates of deposit, money market funds, treasury notes, bonds, or any other accounts; dividends from stocks, mutual funds, and/or net rental income from property, royalties, estates or trusts; child support payments; and other income sources (the question specifically mentions alimony, contributions from family or friends, Veteran’s Administration (VA) payments, Worker’s Compensation, and Unemployment Compensation as possible sources of “other” income). Respondents are told at the start of the Income and Assets section that all questions are seeking information about possible income sources in the previous calendar year (2006).

### Income Amounts (1997-2006)

In years prior to 1997, the NHIS obtained information about the amount of income received from each financial source, but that was dropped in the redesigned NHIS (1997 and beyond) in favor of a single overall estimate of combined family income. Unlike previous NHIS instruments, the 1997-2006 instrument contained three questions to identify the family’s combined income from all sources during the previous calendar year, including a question (FIN.250) that allowed the respondent to supply a specific dollar amount (up to \$999,995). Any family income responses greater than \$999,995 were entered as \$999,996. Respondents who did not know or refused to give a dollar amount to this question were then asked if their total

combined family income for the previous year was \$20,000 or more, OR less than \$20,000 (FIN.260). If the respondent answered this question, he/she was then given one of two flash cards and asked to indicate which income group listed on the card best represented the family's combined income during the previous calendar year (FIN.270). One flash card listed incomes that were \$20,000 or more, and the other flash card listed incomes that were less than \$20,000.

In the 2004 Survey Description Document, data analysts were made aware of an unanticipated issue in 2004 related to the collection of exact amount income data (FIN.250). Specifically, a much larger than expected proportion of respondents reported a family income in the last calendar year of "\$2." In 2004, 2,133 persons (2.25%) had a response of "\$2" to the exact amount of family income question (FIN.250). By comparison, in 2002 (the most recent data year without sample cuts), 136 persons (0.15%) had a response of "\$2" to the exact amount of family income question (FIN.250). In an attempt to reduce the amount of these types of responses, an edit, which would trigger on very high or very low income amounts, was added to the survey instrument. This edit asked the interviewer to verify if the entered information was correct; the interviewer was instructed not to ask the respondent to verify the amount. This change was implemented in the 2005 NHIS starting in Quarter 2. The number of "\$2" responses decreased from 214 in Quarter 1 of 2005 to 59 in Quarter 2, 44 in Quarter 3, and 41 in Quarter 4. In 2004, all of the "\$2" responses to the exact amount of family income (FIN.250) were assigned the value of "not ascertained" and were subject to income imputation. The same procedure was applied to "\$2" responses for 2005: they were assigned the value of "not ascertained" and were subject to income imputation. However, in 2006, any "\$2" responses to total family income were retained and thus not subject to income imputation. This action was chosen since the edit which checks for very high and very low income amounts was in place for the entire time period covered by the 2006 NHIS.

Additionally, a more detailed indicator of poverty status was created by utilizing published information from the U.S. Census Bureau regarding 2006 poverty thresholds (see *Income, Poverty, and Health Insurance Coverage in the United States: 2006*). A ratio of the 2006 income value reported by respondents to the poverty threshold for the same year was constructed, given information on the family's overall size as well as the number of children (aged 17 and under) present in the family. The resulting ratio was subsequently ordered into a poverty gradient consisting of 14 categories (RAT\_CAT). Users should note that the universe for this variable is considered to be all families, because the initial income question was asked of all families. However, the income-to-poverty ratios and resulting RAT\_CAT values could not be calculated in two situations: for families who simply did not supply adequate income information (e.g., those who would only indicate that their income was above or below \$20,000, as well as those who declined to give any income information), and for families where the number of children aged 17 or under equaled the overall number of family members (these observations are coded "99" and "96," respectively, for RAT\_CAT). Analysts should also note that the distribution of income-related recodes INCGRP and RAT\_CAT may differ slightly from 2005 to 2006 because of the different treatment of the "\$2" family income responses.

Starting in 2004, INCGRP, RAT\_CAT, and HOUSEOWN were moved from the Person file to the Family file, replacing the 1997-2003 Family file variables FINCGRP, FRAT\_CAT, and FHOUSE, respectively. Analysts should also note that a second income recode (AB\_BL20K),

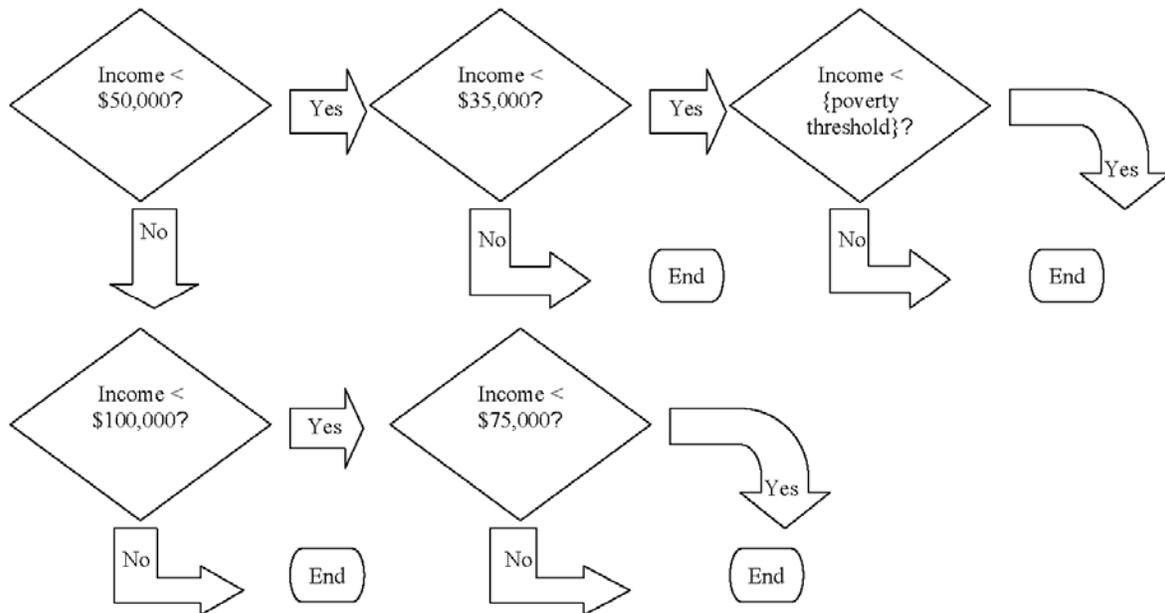
which was included on the 1997-2003 Person file, was deleted from the NHIS public use files starting in 2004 because it could be created from INCGRP and was redundant. In addition, prior to 2004, FGAH was found on the Person file but has since been moved to the Family file.

#### Income Amounts (2007)

Starting with the 2007 NHIS, the income amount follow-up questions that had been in place since 1997 were replaced with a series of unfolding bracket questions. This decision was based on a) the relatively poor performance of the 1997-2006 versions of the follow-up income amount questions, and b) the results of a 2006 pilot test that compared unfolding bracket follow-up questions to the income amount follow-up questions used since 1997.

As mentioned previously, a detailed indicator of poverty status (RAT\_CAT) has been available on NHIS data files since 1997. This variable requires that income be reported in at least interval form. The income amount follow-up question, FIN.270, which collected income information in interval form (1997-2006), had very low response rates: only 15-20% of respondents provided a useable response. Because of the relatively low response rate for the income interval follow-up question, new income follow-up questions utilizing an unfolding bracket methodology were field-tested during the second quarter of the 2006 NHIS. The unfolding bracket method asked a series of closed-ended income range questions (e.g., “is it less than \$50,000?”) if the respondent did not provide an answer to the exact income amount question. The closed-ended income range questions were constructed so that each successive question establishes a smaller range for the amount of the family’s income. A figure demonstrating the flow of the tested income follow-up questions is shown below:

**FLOW DIAGRAM OF FIELD-TESTED INCOME FOLLOW-UP QUESTIONS  
(UNFOLDING BRACKETS), NHIS 2006, 2<sup>nd</sup> QUARTER**



The unfolding bracket income follow-up questions had much higher response rates than the income follow-up questions that had been used 1997-2006 (FIN.260, FIN.270), and as a result, the poverty measure RAT\_CAT had fewer responses with unknown values. The percentage of unknowns for RAT\_CAT during the second quarter of the 2006 NHIS was 17.3% (unweighted) compared with 30.6% (unweighted) based on the original FIN.260 and FIN.270 income follow-up questions. Because of these positive results, the unfolding bracket income follow-up questions were implemented in the 2007 NHIS, and questions FIN.260 and FIN.270 were removed.

Because of these new income follow-up questions, new income variables have been added to the 2007 NHIS. First, grouped income recodes INCGRP2 and INCGRP3 have replaced the 1997-2006 income recode INCGRP. INCGRP3 provides a bridge to the 1997-2006 variable INCGRP. The new income recodes INCGRP2 and INCGRP3 have a higher income top-code than in prior years: in the 2007 NHIS the income top-code was increased to \$100,000 and over. In addition to the new grouped income recodes, two poverty status recodes have been added to the 2007 NHIS. Poverty ratio recodes RAT\_CAT2 and RAT\_CAT3 have replaced the 1997-2006 poverty ratio recode RAT\_CAT. While the poverty ratio recodes in the 2007 NHIS are similar to the 1997-2006 version, the 2007 versions have additional categories reflecting added poverty ratio detail. The additional categories for RAT\_CAT2 allow for data users to construct a three-category poverty ratio variable, and RAT\_CAT3's additional categories allow data users to construct a four-category poverty ratio variable, using the additional information provided by the 2007 NHIS income follow-up questions.

### Home Ownership

Respondents were also asked whether the family's house or apartment was owned or being bought, rented, or occupied by some other arrangement (FIN.280). If the family was renting the current residence, a follow-up question (FGAH or FIN.282) asked whether the family was paying lower rent due to governmental rental assistance.

### Program Participation

Respondents were asked in the final part of the FIN section if any family members were authorized to receive food stamps in 2006, and if so, which members. In addition, respondents were asked whether any family member(s) had *ever* applied for Supplemental Security (SSI) or Social Security disability benefits (even if the claim(s) had been denied). Lastly, if one or more family members had received food stamps or Temporary Assistance to Needy Families (TANF), the respondent was asked, in two separate questions, for how many months during the last calendar year food stamps and/or TANF were provided.

Finally, the NHIS contains three person-level variables relating to the Women, Infants, and Children (WIC) program. The first of these variables, ELIGPWIC, indicates if the person was in a family with at least one WIC age-eligible person (children 0-5 years of age or females 12-55 years of age). If there is at least one WIC age-eligible person in the family, the family respondent is asked if anyone in the family received WIC benefits in the previous calendar year (PWIC). An

additional variable, WIC\_FLAG, is also included in the Person file. WIC\_FLAG indicates if persons who received WIC benefits were age-eligible for the WIC program.

#### Technical Notes

As previously mentioned, the majority of the questions in the FIN section are structured to ask first whether any family member received the applicable income source and, if yes, then to determine which family members received the income source. This format also applies to other items in the section such as TANF, food stamps, and WIC benefits. As mentioned in the 2003 Survey Description Document, the 1997–2003 NHIS only allowed six persons per family to be indicated as receiving the income and/or program source. However, this problem was corrected for 2004 (and beyond). Analysts interested in using these program participation variables for 1997–2003 should refer to the 2003 Survey Description Document for guidance. Further, since qualification for these programs is usually based on a family’s economic circumstances, these program participation variables may have limited analytic value at the person level. Therefore, analysts may find more utility in using the corresponding variables from the Family file.

## **2007 National Health Interview Survey Injury and Poisoning Episode Files**

The Family Core portion of the 2007 survey included questions about medically consulted injuries and poisonings that occurred for any member of the family within a three-month reference period. All injury and poisoning information was provided by the family respondent. Two data files containing injury and poisoning information were created from these data: the Injury/Poisoning Episode file and the Verbatim Injury/Poisoning Episode file.

The inclusion criteria used beginning in 2004 were also used in 2007. In 1997-2003, the Injury/Poisoning Episode file and the Verbatim Injury/Poisoning Episode file contained episodes that were reported to occur within 104 days or four months of the interview and episodes where the date of the injury or poisoning was not reported. Beginning in 2004, the decision was made to retain all injury/poisoning episodes that reportedly occurred during the three months (91 days) prior to the date the injury/poisoning questions were asked based on responses to family level questions FIJ.010\_01.000 to FIJ.028\_00.000 (listed below), regardless of whether or not the date of the injury or poisoning episode subsequently reported by the family respondent in the family level questions was outside the 91 day reference period. Flags have been created to indicate which episodes may thus have occurred outside the 91 day reference period (ETFLG and BEIFLG).

Family level injury/poisoning questions FIJ.010\_01.000 to FIJ.028\_00.000:

“DURING THE PAST THREE MONTHS, that is since [fill 1: date (91 days before today's date)], [fill 2: did you/did you or anyone in your family] have an injury where any part of [fill 3: your/the] body was hurt, for example, with a [fill 4: (random set of examples) cut or wound, broken bone, sprain or burn?]”;

“DURING THE PAST THREE MONTHS, how many different times [fill 1: were you/was ALIAS] injured?”;

“Did [fill 1: you /ALIAS] talk to or see a medical professional about [fill 2: any of these injuries/this injury/your injury or injuries/his injury or injuries/her injury or injuries]?”;

“Of [fill 1: the ^TFINJ3M/all the] times that [fill 2: you were/ALIAS was] injured, how many of those times was the injury serious enough that a medical professional was consulted?”;

“DURING THE PAST THREE MONTHS, that is since [fill 1: date (91 days before today's date)], [fill 2: were you/ were you or anyone in your family] poisoned by swallowing or breathing in a harmful substance such as bleach, carbon monoxide, or too many pills or drugs? Do not include food poisoning, sun poisoning, or poison ivy rashes.”;

“DURING THE PAST THREE MONTHS, how many different times [fill 1: were you/was ALIAS] poisoned? Do not include food poisoning, sun poisoning, or poison ivy rashes.”;

“Did [fill 1: you /ALIAS] talk to or see a medical professional about [fill 2: any of these poisonings/this poisoning/your poisoning or poisonings/his poisoning or poisonings/her poisoning or poisonings]?”;

“Of [fill 1: the ^TFPOI3M/all the] times that [fill 2: you were/ALIAS was] poisoned, how many of those times was the poisoning serious enough that a medical professional was consulted?”

## **I. Injury/Poisoning Episode File**

The Injury/Poisoning Episode file is an episode-based file: each medically consulted (e.g., call to a poison control center; use of an emergency vehicle or emergency room; visit to a doctor’s office or other health clinic; phone call to a doctor, nurse, or other health care professional) injury and poisoning episode reportedly occurred during the three months prior to the date the injury/poisoning questions were asked based on responses to family level questions FIJ.010\_01.000 to FIJ.028\_00.000, and resulted in one or more conditions. An injury episode refers to the traumatic event in which the person was injured one or more times from an external cause (e.g., a fall, a motor vehicle traffic accident). An injury condition is the acute condition or the physical harm caused by the traumatic event. Likewise, a poisoning episode refers to the event resulting from ingestion of or contact with harmful substances, as well as overdoses or wrong use of any drug or medication, while a poisoning condition is the acute condition or the physical harm caused by the event. A person may record up to a total of ten injury and/or poisoning episodes and will be represented in this file as many times as he/she had unique injury and/or poisoning episodes. Each episode must have at least one injury condition or poisoning classified according to the nature-of-injury codes 800-909.2, 909.4, 909.9, 910-994.9, 995.5-995.59, and 995.80-995.85 in the Ninth Revision of the International Classification of Diseases (ICD-9-CM) and one external cause of injury code of E800-E848, E850-E869.9, E880-E929.9, or E950-E999. Other health conditions that were reported as occurring with the injury or poisoning, even if they are not classified according to the above mentioned nature-of-injury codes (e.g., mononeuritis of unspecified site (355.9), other symptoms referable to back (724.8)), are also included in the Injury/Poisoning Episode file.

The Injury/Poisoning Episode file contains information about the external cause and nature of the injury or poisoning episode, what the person was doing at the time of the injury or poisoning episode, the date and place of occurrence, the elapsed time between the date of the injury or poisoning episode and the date the injury/poisoning questions were asked, where the person received medical advice, treatment, or follow-up care, whether the person was hospitalized, whether the person missed any days from work or school due to the injury or poisoning, ICD-9-CM diagnostic codes, and ICD-9-CM external cause codes. ICD-9-CM diagnostic and external cause codes were assigned for all injury and poisoning episodes based on information about how the injury or poisoning happened, the body part injured or poisoned, and the type of injury or poisoning, along with responses to questions about specific types of injury or poisoning episodes, and activity.

During the 2007 data editing process, 122 injury and poisoning episodes were removed out of an initial total of 1,944. These included episodes with no information regarding cause, date and place of occurrence, duplicate episodes, etc. In addition, episodes were removed if they

consisted *solely* of health conditions that could not be classified according to the nature-of-injury codes and external cause of injury codes listed above.

As in previous years, respondents reported episodes that they considered poisonings (e.g., food poisoning and allergic reactions) but that are not considered poisonings based on the ICD-9-CM. These types of episodes were included in the 1997-2003 data files. Beginning in 2004 and continuing in 2007, episodes that are not considered poisonings based on ICD-9-CM are no longer included in the Injury/Poisoning data files.

This file only contains information about injury and poisoning episodes. Other person-level information can be obtained by linking the Injury/Poisoning Episode file to other 2007 NHIS data files (Person, Sample Adult, and Sample Child) using the household serial number (HHX), family serial number (FMX), and person number (FPX). When using a linked Injury/Poisoning Episode file and Sample Adult file, analysis should be limited to those episodes for persons included in the Sample Adult file, and the Sample Adult weight should be applied. When using a linked Injury/Poisoning Episode file and Sample Child file, analysis should be limited to those episodes for persons included in the Sample Child file, and the Sample Child weight should be applied. See Appendix VII for additional information about merging data files.

#### Recall Period and Weights

Questions in the Injury/Poisoning section of the 2007 NHIS have a recall period of the “last 3 months.” However, as the time between the injury/poisoning episode and the date the injury/poisoning questions were asked increases, the annualized number of injuries/poisonings reported decreases. For most analyses of the injury/poisoning data (e.g., estimates for all types of injury/poisoning episodes and estimates for less severe injuries/poisonings), limiting data to episodes with a reported five weeks or fewer between the injury/poisoning episode and the date the injury/poisoning questions were asked is recommended because analyses showed that respondents tend to forget less serious injuries (Warner, et al., 2005). For analysis of injury/poisoning episodes resulting in more serious outcomes (e.g., estimates for fractures and hospitalizations) that are unlikely to be forgotten, the data should not be limited to the five-week period. The longer period of time between the injury/poisoning episode and the date the injury/poisoning questions were asked will increase the number of episodes reported and therefore increase the size of the sample and provide richer detail and greater stability in the estimate. We do not suggest calculating two estimates, one for serious and one for non-serious injuries/poisonings and combining the two estimates.

Analysts may wish to use the recommended five-week reference period to maintain consistency with other studies using the five-week reference period with NHIS injury/poisoning data. However, because the number of days since the injury/poisoning occurred is now provided for each episode on the public use data file, analysts can choose the time period that is the most appropriate for their analysis.

To calculate an annual estimate of the number of injuries and poisonings, the weighted number of episodes reported during a time period is multiplied by the number of time periods in a year. For instance, to estimate the number of injury or poisoning episodes occurring annually

using episodes with three months or less elapsing between the injury/poisoning and the date the injury/poisoning questions were asked, each three-month weighted count should be multiplied by 4 (i.e., by  $52/13=4$ ). If data are limited to episodes with five weeks or less between the injury/poisoning and the date the injury/poisoning questions were asked, each five-week weighted count should be multiplied by 10.4 (i.e., by  $52/5=10.4$ ).

Analysts are cautioned against estimating the number of different *people* injured or poisoned annually using the current NHIS questions. Estimating the number of persons injured using the annualizing method described in the above paragraph (i.e., multiplying the estimate by the number of time periods in a year) assumes that the same individuals experienced injuries at the same rate over the year. Analysts are cautioned to check the Dataset Documentation and the specific item in the questionnaire in order to insure that annual estimates for these kinds of injury or poisoning episodes have intrinsic meaning.

### Variance Estimation

This file does not contain the design variables used in variance estimation. To obtain the design information, the Injury/Poisoning Episode file must be linked to the Person file, the Sample Adult file or the Sample Child file.

### Technical Notes and Imputation Information

Two variables on the Injury/Poisoning Episode file, ICAUS and ECAUS, describe the external cause of the episode. ICAUS is the actual item found in the questionnaire. For each unique episode, the interviewer selected the category of ICAUS that he/she felt best described the episode based on the respondent's description of how the injury or poisoning happened (IPHOW). ECAUS is a recoded variable that describes the cause of the episode using categories based on ICD-9-CM external cause codes. The category into which an episode was placed was based entirely on the first ICD-9-CM external cause code listed for that episode. Appendix I in the Injury/Poisoning Episode Dataset Documentation contains a list of the ICD-9-CM external cause codes found in each category.

In 2006, the variable IPDATENO was changed from a two digit field to a three digit field. This change stemmed from revisions made to the error messages attached to the date fields in the CAPI instrument. The goal was to use consistent criteria for triggering soft and hard edits and consistent messages attached to those edits, regardless of which date path you took in the CAPI instrument. In 2007, the variable IPDATENO remained a three digit field.

Analysts are cautioned regarding their use of the variable RPCKDMR, which indicates the elapsed time between the date of the injury or poisoning episode and the date the injury/poisoning questions were asked. This variable is based on only the month, day, and year of the injury or poisoning episode provided by the respondent and the actual day the respondent was asked the injury/poisoning questions. No information from additional date questions that are currently in the survey were used in the creation of this variable. When possible, the elapsed time between the date of the injury or poisoning episode and the date the injury/poisoning questions were asked is given in days. The time between the date of the injury or poisoning

episode and the date the injury/poisoning questions were asked is only given in months when the day of the injury or poisoning episode was not reported. In previous years, the calculation of this variable was based on the last date when the interview was opened for examination or input of data, not necessarily on the date when the injury/poisoning questions were asked, which could be different. This could happen if the interviewer was unable to complete the interview in one visit and had to return at a later date, so the injury and poisoning questions may have been completed earlier than indicated by the date of the interview recorded by the CAPI instrument. If this occurred, the actual time between the date of the injury or poisoning episode and the date the injury/poisoning questions were asked would be less than the elapsed time indicated by the variable RPCKDMR. Beginning in 2004, the actual date when the injury and poisoning questions were completed was recorded and used in the calculation of this variable.

Beginning in 2004 and continuing in 2007, imputation was implemented for episodes that did not have a valid month, day, and year of occurrence. Imputation was done so that it would be possible to calculate a specific elapsed time in days between the date of the injury/poisoning episode and the date the injury/poisoning questions were asked for all episodes in the Injury/Poisoning Episode file and the Verbatim Injury/Poisoning Episode file. Since all episodes in the files now have a specific elapsed time (RPD) between the date of the injury/poisoning episode and the date the injury/poisoning questions were asked, analysts will be able to calculate estimates based on the time period of their choice.

The variable RPD indicates the elapsed time in days between the date of the injury or poisoning episode and the date the injury/poisoning questions were asked. This variable is based on all date information that was given by the respondent, and when date information was missing, imputed information was used in the creation of this variable. For some injury and poisoning episodes, the respondent was only able to provide the month and year of occurrence; or a time period within the month (beginning, middle, or end) and year of occurrence; or the number of days, weeks, or months ago. For cases in which a month but no time period during the month was provided, a day was imputed between 1 and the last day of the month. For cases in which the month of the injury/poisoning episode and the time period within the month was provided, the day of the month within that time period was imputed. If the episode was reported as occurring during the beginning of the month, a day of 1-10 was imputed; for cases in the middle of the month, a day of 11-20 was imputed; and for cases at the end of the month, a day of 21 to either 28, 29, 30, or 31, depending on the month, was imputed. In other instances, the respondent was only able to provide a time period (i.e., number of days, weeks, or months) between the date the injury/poisoning occurred and the date the injury/poisoning questions were asked. For responses given in days ago, the corresponding value of RPD was calculated. For responses given in weeks ago or months ago, RPD was imputed from within, respectively, the interval  $7(\# \text{ weeks ago}) \pm 3$  or the interval  $30(\# \text{ months ago}) \pm 15$ .

An elapsed time interval, with lower and upper bounds BIETD and EIETD, respectively, indicates the amount of uncertainty in the injury/poisoning episode date information that was provided by the respondent. If the specific day, month, and year of the episode were provided or could be deduced from information provided by the respondent, then  $BIETD = EIETD = RPD$ . Otherwise, BIETD and EIETD indicate the lowest and highest values of the elapsed time between the episode and the date the injury/poisoning questions were asked that were consistent

with the reported episode date information, and RPD was imputed to be within that interval. In a few cases where insufficient information was provided to determine an elapsed time interval, values of BIETD, EIETD, and RPD were obtained from a random “donor” (another reported episode) using hot deck imputation.

There are several variables in the 2007 Injury/Poisoning Episode file that supply information about the imputed data and about the consistency of the episode date information provided by respondents. The variable IMPMETH indicates which episodes have a value for RPD that is based on a specific day, month, and year of the episode that was provided or was deduced from information provided by the respondent (i.e., no imputation was needed) and which episodes have a value for RPD that was imputed. Flag variables have been added to the file to indicate whether the elapsed time (RPD) or the elapsed time interval boundaries (BIETD and EIETD) fall within the 91-day reference period mentioned in family level questions FIJ.010\_01.000 and FIJ.020\_00.000. This was done because it is possible that the respondent provided inconsistent information (i.e., reported that the injury or poisoning occurred during the 91-day reference period mentioned in the family level questions, and then, in follow-up questions about the episode date, reported that the injury or poisoning occurred beyond the 91-day reference period mentioned in the family level questions). Also, the elapsed time interval boundaries and imputed values of the elapsed time were not constrained to be  $\leq 91$ ; they were only constrained to be consistent with the date information reported by the respondent. Variable ETFLG indicates whether the elapsed time (RPD) is  $\leq 91$  days. Variable BEIFLG indicates whether the boundaries (BIETD and EIETD) of the elapsed time interval are  $\leq 91$  days. These flags were created for convenience so that analysts can decide which version of inconsistently-reported date information to use. Analysts may also choose to re-impute values of RPD that are greater than 91, constraining them to be within the 91-day limit as well as within the elapsed time interval.

## **II. Verbatim Injury/Poisoning Episode File**

The Verbatim Injury/Poisoning Episode file contains edited narrative text descriptions of the injury or poisoning provided by the respondent and includes a description of how the injury or poisoning happened and “other specified” responses for the body part injured, the kind of injury, the place the person received medical care, the cause of the poisoning, and the activity at the time of the injury/poisoning. (The pre-edited responses are “verbatim” only insofar as the interviewer could type the information and condense it to fit the 300 character field.) Editing was done only to protect the injured or poisoned person’s confidentiality. Text descriptions used to replace original text that could have resulted in a breach of confidentiality are surrounded by arrows (<>). Grammatical and/or spelling errors were not corrected. The codes of “R,” which represents “Refused;” “D” or “DK,” which represent “Don’t know;” and “N,” which represents “No more information,” have also been left in the file. The following types of changes were made to the file in order to protect the injured or poisoned person’s confidentiality:

- Person names (first, middle, and/or surnames or initials) were replaced with <He> or <She>;

- Names of commercial operations were replaced with a general category (e.g., the name of a restaurant that serves fast food would be replaced with <fast food restaurant>);
- All place names including cities, counties, states, and street addresses were removed;
- The detailed description of an occupation was replaced with a more general category using the North American Industrial Classification System (NAICS) as a guide;
- Brand names were replaced with a generic term for the product (e.g., the brand name of a car would be replaced with <motor vehicle>);
- Text that indicated unusual personal behavior or events was modified to make it less remarkable;
- Any group or organization that was known to have a register of its members was replaced with a generic term.

## **2007 National Health Interview Survey Sample Child File**

The Sample Child section of the 2007 NHIS covers additional subject areas not included in the Family Core. Moreover, the questions in the Sample Child section are more specific and are intended to gather more detailed information than those in the Family Core. Sample children do not self-report; instead a knowledgeable adult (typically a parent or guardian) answers questions on the sample child's behalf. More detail about the sections comprising the Sample Child File is discussed below.

### **I. Child Conditions, Limitation of Activity and Health Status Section (CHS)**

The Child Conditions, Limitation of Activity, and Health Status Section (CHS) of the 2007 NHIS contains information on conditions, limitations of activity, health status, and mental health. The CHS includes questions on the following health conditions: mental retardation, developmental delays, Attention Deficit Hyperactivity Disorder (ADHD) or Attention Deficit Disorder (ADD), Down's syndrome, cerebral palsy, muscular dystrophy, cystic fibrosis, sickle cell anemia, autism, diabetes, arthritis, congenital and other heart disease, asthma, allergies, colitis, anemia, ear infections, seizures, headaches, stuttering, and stammering. A question about whether the sample child still has asthma is included. This section also contains a question used to determine the number of school-loss days reported during the 12 months prior to the interview. In addition, respondents were asked about hearing and vision loss; if a health problem requires the sample child to use special equipment such as a brace, wheelchair, or hearing aid; whether the sample child's health is better, worse, or the same compared with 12 months ago; and whether the sample child currently has a problem that has required prescription medication for at least three months. Lastly, there are questions about the sample child's height and weight.

In 2007, child mental health questions derived from the Child Behavior Checklist for children ages 2-3 years remain in the CHS section. The items in the checklist were chosen for their ability to discriminate between children who have not received mental health services in the preceding 12 months and those who have, by using demographically-matched normative and clinical samples for boys and girls. Each set of items can be viewed as comprising a scale with each item scored as either "0," "1," or "2." More information on the scale derived from the Child Behavior Checklist is included in Appendix IV of this document.

In 2007, in conjunction with the Child Complementary and Alternative Health supplement (CAL), several supplemental condition questions were embedded in the CHS section. These questions covered a variety of conditions and were asked for all sample children: LUNGYR (lung problem), CANCERYR (cancer), NEUROYR (neurological problem), URINYR (urinary problem), GUMYR (gum disease), FLUYR (flu/pneumonia), SINUSYR (sinusitis), and STREPYR (strep throat/tonsillitis).

Additional questions were asked for sample children two years of age and under: ALOTHYR1 (allergies), ABDOMYR1 (abdominal pain), BACKYR1 (back/neck pain), PNOTHYR1 (chronic pain), FATIGYR1 (fatigue), FEVRYR1 (fever), COLDYR1 (head/chest

cold), THOTHYR1 (sore throat, other than strep/tonsillitis), ACIDYR1 (acid reflux), NAUSYR1 (nausea/vomiting), CONSPYR1 (recurring constipation), OVRWTYR1 (overweight), ACNEYR1 (acne), WARTSYR1 (warts), and SKOTHYR1 (skin problem other than eczema/acne/warts).

A parallel set of questions was asked for sample children 3-17 years of age: ALOTHYR2 (allergies), ABDOMYR2 (abdominal pain), BACKYR2 (back/neck pain), PNOTHYR2 (chronic pain), FATIGYR2 (fatigue), FEVRYR2 (fever), COLDYR2 (head/chest cold), THOTHYR2 (sore throat, other than strep/tonsillitis), ACIDYR2 (acid reflux), NAUSYR2 (nausea/vomiting), CONSPYR2 (recurring constipation), OVRWTYR2 (overweight), ACNEYR2 (acne), WARTSYR2 (warts), and SKOTHYR2 (skin problem other than eczema/acne/warts). Two additional questions not asked for younger children were asked for sample children aged 3-17 years: HEADYR2 (headache) and INSOMYR2 (insomnia).

Four additional mental health questions were asked for sample children 4-17 years of age: DEPRSYR (depression), PHOBYR (phobia/fears), C\_ANXYR (anxiety/stress), and INCONTYR (incontinence/bed wetting). A question about menstrual problems was asked for all female sample children aged 10-17 years: MENSTYR (menstrual problems/PMS).

## Technical Notes

Beginning in 2004 there has been a 10-fold increase in the number of sample children who were reported to have cerebral palsy (34 cases reported in 2003; 311-353 cases reported each year from 2004-2006). This increase is believed to be due to interviewer error arising from operational differences in the CASES instrument used prior to 2004 and the Blaise instrument used beginning in 2004. Beginning in Quarter 3 of 2007 an interviewer note was added to the Blaise instrument to address this difference, and the number of cerebral palsy cases reported in 2007 fell to 159. Although this variable is included in the file, we suggest that it **not** be used for analysis.

Several questions pertaining to child behavior are used to create recodes; only the recodes are included in the Public Use file. The background and usage of the mental health indicators can be found in Appendix IV.

Regarding the CHS data on colds and intestinal illnesses, analysts should keep in mind that the questions are measuring fairly broad symptoms and illnesses. Furthermore, these may be a result of either acute or chronic conditions (e.g., irritable bowel syndrome or respiratory allergies). These data are best used to measure trends over time.

## II. Child Health Care Access and Utilization Section (CAU)

The Child Health Care Access and Utilization Section (CAU) of the 2007 NHIS contains information on access to health care, dental care, and health care provider contacts. The questions pertaining to access to health care include: having a usual place for sick care; having a usual place for routine/preventive care; change in place of care; reasons for a delay in getting medical care; and the inability to afford medical care. A question on dental care asked about the length of time since last dental visit.

Questions regarding health care provider contacts include visits and telephone contacts to or from medical doctors and other health care professionals (such as chiropractors) in the past 12 months. As with the FAU section discussed previously, the category of “health care professional” has been expanded to include chiropractors, various types of therapists, psychiatrists, psychologists, and social workers; moreover, contacts or visits are not restricted to medical doctors or professionals working with/for a medical doctor. Note that questions about home care are asked independently of other types of health care visits. In addition, the reference period for all health care contacts is the past 12 months. Lastly, a separate question is asked about the number of visits to a hospital emergency room in the past 12 months.

### **III. Child Mental Health Brief Section (CMB)**

As part of a collaborative agreement with the National Institute of Mental Health (NIMH), the Strengths and Difficulties Questionnaire (SDQ) was first used in 2001 in a Child Mental Health Supplement in the CAU section. The SDQ is a behavioral screening questionnaire for children ages 4 to 17 years with extended questions that provide information on the duration of a child’s problem and the impact that the problem has on the child and his/her family. It is copyrighted by Dr. Robert Goodman, London, England and is used with his permission. In 2002 the long version of the SDQ was deleted from the CAU section, and a short version of the SDQ was added to the CHS core.

In 2003 the short version of the SDQ was dropped from the CHS section, and the long SDQ was reinserted into the CAU section. The six items from the short SDQ in 2002 reverted to their original names and question numbers in the long SDQ in 2003 as follows: CSCL2\_C2 in 2003 (CMHMF12 in 2002), CSCL2\_E2 in 2003 (CMHMF13 in 2002), CSCL3\_E3 in 2003 (CMHMF14 in 2002), CSCL5\_P5 in 2003 (CMHMF15 in 2002), CSCL5\_H5 in 2003 (CMHMF16 in 2002), and CSCL6 in 2003 (CMHDIFF in 2002).

In 2004, the long SDQ was transferred from the CAU section to a newly created section, Child Mental Health (CMH). The question/answer wording and the question order remained the same as in 2003. Variable names (except CSCL7) have been changed to accommodate the new editing system. The question numbers have also been changed to reflect the new question numbering system and the new section name.

In 2005, the long SDQ (CMH section) was dropped. The short SDQ, a subset of the long SDQ which was originally fielded in 2002, was reinserted in the NHIS, the Child Mental Health Brief Section or CMB. For the short SDQ items, the question/answer wording did not change and variable names remained the same as in 2004. However, question numbers were changed to reflect placement in a new section.

In 2007, the CMB section remained the same as in 2006. More detailed information on the SDQ is provided in Appendix V.

Note that other child mental health questions derived from the Child Behavior Checklist for children ages 2-3 years remain in the CHS section.

### **Important Note**

**The original numbering system of the response categories in the instrument has been modified in the Variable Layout Report for all variables in the CMB section. In order to correspond with the SDQ scoring system detailed in Appendix V, all variables with original answer codes of 1, 2, 3 in the instrument were changed to 0, 1, 2 in the data file, Variable Layout Report, and Variable Frequency Report; all variables with original answer codes of 1, 2, 3, 4 in the instrument were changed to 0, 1, 2, 3 in the data file, Variable Layout Report, and Variable Frequency Report.**

#### **IV. Child Mental Health Services (CMS)**

In 2007 NHIS information was collected about mental health services for sample children 4-17 years of age who have difficulties with emotion, concentration, behavior, or being able to get along with others. The CMS includes information about a health care provider visit for mental health services, medication prescribed for difficulties with concentration, hyperactivity or impulsivity, receipt of any non-pharmacologic treatment or help for difficulties with concentration, hyperactivity or impulsivity, and the type of provider for this treatment—pediatric or general medical care practice, mental health private practice, mental health clinic or center, the child’s school or some other place.

#### **V. Child Influenza Immunization Section (CFI)**

The Child Influenza Immunization Section (CFI) was included for the first time in the 2005 NHIS. This section contains information on receipt of a flu vaccination in the past 12 months; month and year of the most recent flu vaccination; receipt of nasal flu spray vaccination in the past 12 months; and month and year of most recent nasal flu spray vaccination. These questions were also administered to all sample adults (see the AAU section). No changes were made in 2007.

**CAUTION.** The 2007 Sample Child public use files have two records from a singleton PSU, i.e., only one PSU was drawn---see Appendix III Conceptual NHIS design for 2007 for discussion of primary sampling units. These same two respondents are also in the Child Complementary and Alternative Medicine supplement public use file. Contemporary software for complex survey design variance estimation such as SAS, SPSS and SUDAAN may estimate variances that are marginally different from estimates obtained from including dummy records for the two singleton PSU respondents. Users of some software (e.g., Stata 9) may get missing values for variance estimates. Users of SUDAAN should specify the MISSUNIT option in the NEST statement to prevent errors that stop processing of data. A mini-file with two dummy records for the singleton PSUs, assigning a

record weight equal to 1 to each, will be released in August 2008 along with instructions and documentation describing the potential impact of singleton PSUs on variance estimation. Users encountering software problems arising from singleton PSUs can append the dummy record file to the public use files.

## **2007 National Health Interview Survey Child Complementary and Alternative Medicine File (CAL)**

The Child Complementary and Alternative Medicine Supplement (CAL) collected information about sample children on their use of 17 non-conventional health care practices. The modalities covered in the supplement include acupuncture, ayurveda, biofeedback, chelation therapy, chiropractic or osteopathic manipulation, energy healing therapy, hypnosis, massage, naturopathy, traditional healers, movement therapies, herbal and non-vitamin supplements, vitamins and minerals, homeopathy, special diets, yoga/tai chi/qi gong, and relaxation techniques.

While some questions were tailored to the specific health care practice, others were asked for most of the 17 modalities. The topics covered include: whether the sample child used the modality in the past 12 months; whether the sample child had seen a provider or practitioner for the modality; whether the modality was used to treat a specific health problem or condition, and, if so, what health problems or conditions were treated; and whether the modality was used to improve athletic/sports performance. Given the number of modalities and questions, a separate stand-alone data file was created for the supplement rather than append it to the 2007 Sample Child file. The Sample Child weight is included on the CAL file and should be used for all the variables on this file.

Although the questions in the CAL Supplement were asked about all sample children, some respondents did not complete the supplement. Respondents who did not give responses to any of the questions in the supplement had a value of “not ascertained” inserted in the appropriate fields in this file. These children were given a coded value of “1” for the record completion status variable (COMPSTAT), which indicates that their record only contains responses of “not ascertained.” In other words, these children (n=98, or 1.0% of sample children) are retained in the file, but they are coded as “8” in all the relevant fields of the file. All children who met the criteria for completing the supplement were given a coded value of “2” (all answers refused or don’t know), “3” (all answers refused, DK, or not ascertained), or “4” (at least one valid answer) for the record completion status variable.

### **Back Coding**

In the CAL Supplement, respondents were given the opportunity to choose “Other specify” as a response to the questions about what health problems or conditions were treated with specific types of complementary and alternative medicine (CAM). If “Other specify” was selected, the respondent was then asked to give a verbatim response indicating the health problem or condition treated with the specified type of CAM. After reviewing the verbatim responses, it was found that some of these responses could be back coded (added) into existing categories, while other responses occurred so frequently that an additional category needed to be created.

In the CAL file, one additional health problem or condition response category, other musculoskeletal, was added to the health problem or condition list for each therapy, as part of the

editing process. When possible, verbatim responses were back coded into existing categories or the new other musculoskeletal category. In all cases, the original verbatim text was retained in the original variables. As a result, the frequency count for the “other specify” variable for each CAM therapy may not match the frequency for its corresponding verbatim variable.

**CAUTION.** The 2007 Sample Child public use files have two records from a singleton PSU, i.e., only one PSU was drawn---see Appendix III, Conceptual NHIS design for 2007, for discussion of primary sampling units. These same two respondents are also in the CAL supplement public use file. Contemporary software for complex survey design variance estimation such as SAS, SPSS and SUDAAN may estimate variances that are marginally different from estimates obtained from including dummy records for the two singleton PSU respondents. Users of some software (e.g., Stata 9) may get missing values for variance estimates. Users of SUDAAN should specify the MISSUNIT option in the NEST statement to prevent errors that stop processing of data. A mini-file with two dummy records for the singleton PSUs, assigning a record weight equal to 1 to each, will be released in August 2008 along with instructions and documentation describing the potential impact of singleton PSUs on variance estimation. Users encountering software problems arising from singleton PSUs can append the dummy record file to the public use files.

## 2007 National Health Interview Survey Sample Adult File

The Sample Adult section of the 2007 NHIS covers many of the subject areas included in the Family Core. However, the questions in the Sample Adult section are more specific and are intended to gather more detailed information. In addition, sample adults generally respond for themselves, although in a small number of cases, proxy responses are allowed if the selected adult had a physical or mental condition prohibiting him/her from responding. The variable PROX1 indicates those cases where information was obtained from a proxy respondent. The eight sections comprising the Sample Adult section are discussed below.

### I. Adult Demographics Section (ASD)

**Users are advised that the two-digit recodes based on Census codes derived from the 1987 SIC/SOC have been discontinued in the 2005, 2006, and 2007 NHIS.** The 2007 NHIS file contains two-digit recodes based on Census codes derived from the 2002 NAICS and the 2000 SOC. Please see the section below entitled “Industry and Occupation Coding” for additional information.

The Adult Socio-Demographics (ASD) section contains information regarding the occupation, industry, workplace, and employment conditions of currently employed sample adults as well as those who have ever worked (e.g., retired persons).

Sample adults aged 18 years and older who were “working at a job or business,” “with a job or business but not at work,” or “working, but not for pay, at a job or business” during the week prior to their interview were asked a series of questions about their job and work status during the week prior to the interview. In addition, those sample adults who said that they were “looking for work” or “not working and not looking for work” during the week prior to the interview were asked if they had “ever held a job or worked at a business.” Sample adults who responded affirmatively were then asked the occupation, industry and work status questions in the ASD section. Note that sample adults who had ever worked and were either retired or 65 years of age or older were asked about the job they had held the longest, whereas sample adults who had ever worked, were younger than 65 years of age, and were not retired were asked about their most recently held job. In a subsequent question, currently employed sample adults were asked if their current job was also the job they had held for the longest time. Likewise, sample adults who had ever worked and were not retired were asked if their most recently held job was also the job they had held for the longest.

Additional questions in the ASD section ask sample adults to describe their current/most recent/longest-held employment situation (whether they were employed by a private company or business, the federal government, a state or local government, self-employed in their own business or professional practice, or working without pay in a family business or farm), the number of full and part time employees at their workplace, how long they had worked at their current/most recent/longest-held job, whether they were paid by the hour, and whether they received paid sick leave. Respondents who indicated that they were self-employed at their current/most

recent/longest-held job were asked whether they had an incorporated business. Currently employed sample adults were asked whether they were working at more than one job.

Users should be aware that DOINGLWA and WHYNOWKA are the ASD equivalents of DOINGLWP and WHYNOWKP in the FSD section. For the majority of respondents, DOINGLWA and DOINGLWP will have identical values (and, likewise, WHYNOWKA and WHYNOWKP). However, it is nevertheless possible that DOINGLWA and DOINGLWP (and WHYNOWKA and WHYNOWKP) may have inconsistent values across the Sample Adult and Person data files. Users wishing to reconcile any discrepant values are advised to use the values of DOINGLWA and WHYNOWKA (rather than DOINGLWP and WHYNOWKP, respectively), since the information obtained from the family respondent during the FSD portion of the interview (and reflected in DOINGLWP and WHYNOWKP) was subsequently confirmed and corrected by the sample adult during his or her interview (as reflected in DOINGLWA and WHYNOWKA).

With the exception of BUSINC1A, WRKLONGH, and ONEJOB, the universe for all variables in the 2007 ASD section includes currently employed and ever employed sample adults. Variables with smaller universes (e.g., currently employed sample adults only) that were included in previous years' data files have been eliminated. Users wishing to replicate those variables are advised to use DOINGLWA to identify the subset of currently employed sample adults (i.e., DOINGLWA = 1, 2, or 4).

### Industry and Occupation Coding

During the course of the interview, verbatim responses were obtained from each eligible respondent regarding his/her industry and occupation. This information was subsequently reviewed by U.S. Census Bureau coding specialists, who assigned appropriate industry and occupation codes. These 3-digit codes, developed by U.S. Census Bureau staff for use in Federal surveys, were consistent with the structures of the Standard Industrial Classification (SIC) and Standard Occupation Classification (SOC) (but were not actual SIC and SOC codes). Prior to the 1997 NHIS, the codes were included on all NHIS public use data files. However, a review of NHIS data suggested that the level of detail contained in the codes could compromise respondent confidentiality. Consequently, beginning in 1997, the 3-digit codes were restricted to in-house NHIS data files, and DHIS staff created several 2-digit industry and occupation recodes that could be included on the public use data files. The latter recodes were based on occupation and industry groups and subgroups consistent with the existent SIC and SOC structures.

Changes in the U.S. economy led to changes in the SIC and SOC classifications. After an extensive period of review, the standard industry and occupation classifications – and the corresponding 3-digit Census codes used by the NHIS and other Federal surveys – were replaced by the North American Industrial Classification System (NAICS) and a revamped SOC (referred to subsequently as “New SOC”). Accordingly, the Census Bureau has developed new 4-digit industry and occupation codes to replace the obsolete 3-digit codes.

Most Federal surveys have switched to the new classification systems based on NAICS/New SOC. As a result, the 2004-2007 NHIS in-house data files contain 4-digit 2002 Census codes for industry and occupation consistent with the 2002 NAICS and 2000 New SOC. As with the 2004-2006 NHIS public use data files, the 2007 NHIS public use data files contain 2-

digit industry and occupation recodes based on these 4-digit Census codes. The 2004 NHIS public use data contained a second set of 2-digit industry and occupation recodes (OCCUP1A, OCCUP2A, INDSTR1A, and INDSTR2A) based on the 3-digit 1990 Census codes (and, in turn, the 1987 SIC and 1980 SOC); these were dropped in 2005 (and after).

Users are advised that the previous coding scheme based on the 3-digit Census codes and the new coding scheme based on the 4-digit Census codes are entirely different classification systems that are not compatible with one another. Moreover, crosswalks showing how these systems compare to one another are not available at this time. However, the coding categories for these recodes are provided in the Industry and Occupation Appendices (following the Variable Layout Report for the Sample Adult file), and additional information is available on-line (see the final paragraph in this section).

While the 2007 NHIS Sample Adult public use file does not include the 4-digit Census codes, it does include a detailed occupation recode (OCCUPN1) with 94 distinct categories, while the associated simple recode (OCCUPN2) has 23 categories. These categories are derived from the 2000 New SOC Occupation Subgroups and Major Occupation Groups, respectively, as determined by the U.S. Census Bureau and the Bureau of Labor Statistics. The detailed industry recode (INDSTRN1) informed by the 2002 NAICS has 79 distinct categories, while the associated simple recode (INDSTRN2) has 21 categories. These categories are derived from the 2002 NAICS Industry Subsectors and Sectors, respectively, as identified by Census.

For more information about the 2002 NAICS, please refer to <http://www.census.gov/epcd/www/naics.html>. For more information about the 2000 SOC, please refer to <http://www.bls.gov/soc/home.htm>. To obtain a complete list of the NAICS Sectors and Subsectors and the SOC Major Groups and Subgroups, please refer to the Industry/Occupation Documentation for the Sample Adult data file at: [http://www.cdc.gov/nchs/about/major/nhis/nhis\\_2007\\_data\\_release.htm](http://www.cdc.gov/nchs/about/major/nhis/nhis_2007_data_release.htm)

## **II. Adult Conditions Section (ACN)**

The ACN section of the 2007 NHIS obtains information from the sample adult as to whether he or she has, or has had, a selected number of medical conditions. In most instances, sample adults were asked whether a doctor or other health professional had told them that they had the condition in question (joint symptoms, pain, hearing, vision impairment, and tooth loss are the exceptions). Respondents are also asked about head colds and intestinal illness which began in the 2 weeks prior to the interview, and women age 18-49 are asked about current pregnancy status. In addition, the section contains information about the sample adult's current mental or emotional health (whether he or she experienced feelings of sadness, nervousness, restlessness, hopelessness, worthlessness, or that everything was an effort in the past 30 days), and the extent to which these feelings interfered with his or her life or daily activities (Kessler's "K6" screen for nonspecific psychological distress). For more information about Kessler's K6 please refer to [http://www.hcp.med.harvard.edu/ncs/k6\\_scales.php](http://www.hcp.med.harvard.edu/ncs/k6_scales.php). Table 6 shows the specific health-related conditions covered in this section, as well as the various reference periods covered by the questions.

**Table 6. Sample Adult File: Conditions and Reference Periods**

	Condition	Ever	12 mos.	3 mos.	30 days	2 weeks	Now	Other
ACN.010	Hypertension	X						
ACN.020	Hypertension 2+ visits							Twice
ACN.031	Coronary heart disease	X						
ACN.031	Angina pectoris	X						
ACN.031	Heart attack (MI)	X						
ACN.031	Other heart condition or heart disease	X						
ACN.031	Stroke	X						
ACN.031	Emphysema	X						
ACN.080	Asthma	X						
ACN.085	Asthma still have						X	
ACN.090	Asthma episode / attack		X					
ACN.100	Asthma ER visit		X					
ACN.110	Ulcer ever told	X						
ACN.120	Ulcer recent		X					
ACN.130	Cancer any	X						
ACN.140	Cancer kind	X						
ACN.150	Cancer when							Age
ACN.160	Diabetes	X						
ACN.170	Diabetes when							Age
ACN.180	Insulin						X	
ACN.190	Oral agents/pills						X	
ACN.201	Hay fever		X					
ACN.201	Sinusitis		X					
ACN.201	Chronic bronchitis		X					
ACN.201	Weak kidneys		X					
ACN.201	Liver condition		X					
ACN.250	Joint symptoms				X			
ACN.260	Joints affected				X			
ACN.270	Joint symptoms chronic			X				
ACN.280	Joints doctor consult	X						
ACN.290	Arthritis (arthritis, gout, fibromyalgia, rheumatoid arthritis, lupus) diagnosis	X						
ACN.295	Limited in activities due to arthritis/joint symptoms						X	
ACN.300	Neck pain			X				

	Condition	Ever	12 mos.	3 mos.	30 days	2 weeks	Now	Other
ACN.310	Back pain			X				
ACN.320	Leg pain			X				
ACN.331	Jaw, face pain			X				
ACN.331	Migraine			X				
ACN.350	Head/chest cold					X		
ACN.360	Intestinal illness					X		
ACN.370	Pregnant						X	
ACN.430	Vision impairment						X	
ACN.440	Blind						X	
ACN.451	Lost all teeth						X	
ACN.471	Sad				X			
ACN.471	Nervous				X			
ACN.471	Restless				X			
ACN.471	Hopeless				X			
ACN.471	Everything an effort				X			
ACN.471	Worthless				X			

The cancer questions were asked in a format that allowed a respondent who reported having had cancer to specify up to three types of cancer as well as to indicate that he/she had had more than three different cancers. The responses were recorded with the codes indicated in the questionnaire and were then transformed into “mentioned”/ “not-mentioned” variables during editing. These variables (CNKIND1-31) assign to every sample adult who reported having ever had cancer either a “mentioned,” if he/she specified that particular cancer, a “not mentioned,” if he/she did not specify that cancer, or a “refused,” “don’t know,” or “not ascertained,” if there was no information for any of the cancers. Thus, a sample adult may have a code in each of the cancer variables, but can have only up to three “mentions,” with a fourth mention possible for the variable CNKIND31 (“More than 3 kinds”).

Age questions CANAGE1-30 and DIBAGE (“How old were you when you were diagnosed [with this condition]?”) are “top coded” to 85+ years to insure confidentiality among the oldest respondents. The recode DIFAGE2 (“How long have you had diabetes” [AGE minus DIBAGE]) is calculated prior to top coding AGE and DIBAGE, but is itself top coded to 83+ years to insure confidentiality. The answers to the age questions were not edited for reasonableness, and some respondents appear to have given the length of time since they were diagnosed rather than their age at diagnosis.

In 2007, a new core diabetes question was added: DIBPRE1 (“Have you EVER been told by a doctor or other health professional that you have any of the following: prediabetes, impaired fasting glucose, impaired glucose tolerance, borderline diabetes, or high blood sugar?”).

A new supplemental hearing question AHEARST1 (“Is your hearing excellent, good, a little trouble hearing, moderate trouble, a lot of trouble, or are you deaf?”) replaces the core question AHEARST. AHEARST1 has more response categories than AHEARST. The core question HEARID was also dropped in 2007, and a number of supplemental questions about hearing, hearing loss, and hearing aid use were included: HRWORS, HRWHICH, HRRIGHT, HRLEFT, HRWHISP, HRTALK, HRSHOUT, HRSPEAK, HRBACK, HRFRUST, HRSAFETY, HEARAGE1, HRSUDDEN, HRCAUS1, HRFAM, HRPROBHP, HRENT, HRAUD, HRTEST, HRCOCREC, HRCOCNOW, HRAIDNOW, HRAIDLNG, HRAIDYR, HRAIDEV, HRAIDREC, HRAIDLGP, HRAIDOFT, HRAIDNT1-9, HREALDS, HREALDTP1-9, HRTIN, HRTINLNG, HRTINOFT, HRTINMUS, HRTINSLP, HRTNPROB, HRTINDIS, HRTINRM, HRREMT01-15P, HRFIRE, HRFIRTYR, HRTOTR, HR12MR, HRFRPROT, HRWRKNOS, HRWRKTOT, HRWRKYR, HRWKPROT, HRLESNOS, HRLESTP01-10, HRLESYR, and HRLSPROT.

Major changes were made in 2002 to core questions about arthritis and joint symptoms, and those questions remained unchanged through 2007. Users are advised to read the 2002 Survey Description document to learn about those changes. Because of those changes, any comparisons of 2002-2007 arthritis and joint symptom data with data prior to 2002 should be undertaken with caution.

In conjunction with the Adult Complementary and Alternative Health supplement (ALT), several supplemental condition questions were embedded in the ACN section in 2007. Some of these questions elaborate on existing core condition questions and place the condition in a 12 month timeframe: HYPYR, CHDYR, ANGYR, MIEVYR, HRTYR, STRYR, and EPHYR. ARHTYR1-6 (type of arthritis) expands on the existing core arthritis questions. Several new conditions are also ascertained, some of which are also placed in a 12 month timeframe: CHLEV, CHLYR (high cholesterol), GUMDISEV, GUMDISYR (gum disease), PHOBIAEV, PHOBIAYR (phobias or fears), AFLUPNEV, AFLUPNYR (influenza or pneumonia), PRCIREV, PRCIRYR (poor circulation in legs), UREV, URYR (urinary problems), ADDHYP (Attention Deficit Disorder or hyperactivity), AUTISM, BIPDIS (bipolar disorder), DEMENTIA, MANIAPSY (mania or psychosis), SCHIZPHN (schizophrenia), SEIZURES, BOWLEV, BOWLYR, BOWLTYR1-3 (inflammatory bowel disease, irritable bowel, or severe constipation), ACIDRYR (acid reflux or heartburn), HACHEYR (headaches), MEMLOSyr (memory loss), SPNYR (sprains), DENYR (dental pain), ALCTOBYR (excessive use of alcohol or tobacco), SUBABYR (substance abuse), SKNYR (skin problems), INSYR (insomnia), FATYR (daytime sleepiness), DEPYR (depression), ANXYR (anxiety), MENSYR (menstrual problems), MENOYR (menopausal problems), GYNYR (gynecologic problems), and PROSYR (prostate problems or impotence).

### **III. Adult Health Status and Limitation of Activity Section (AHS)**

The Adult Health Status and Limitation of Activity component of the Sample Adult file contains information from respondents on illness behavior, health status, special equipment, limitations in functional activities, and the conditions underlying such limitations. While the AHS section may seem similar to the FHS section in the Person file, the questions in these sections have a somewhat different focus. For example, both sections asked about the ability to walk without

special equipment. However, the walking limitation question in the FHS section (FHS.210) only captured whether a person has difficulty walking without using special equipment. In contrast, the Sample Adult question on walking (AHS.091) asked about the degree of difficulty the respondent has walking a specified distance (a quarter mile, or about three city blocks) by him/herself and without using any special equipment.

The 2007 AHS time variables and recodes, which indicate how long respondents have had the condition(s) causing their limitation(s), were processed using procedures similar to those used in 2002-2006. Substantively, the 2002-2007 variables and recodes are similar to those from previous years (1997-2001), but the detailed unknown categories that were included in the earlier data were collapsed into broader categories starting in 2002.

### Health Indicators: Illness Behavior and Health Status

The first questions in this section determined the number of days the respondent took off from work or spent in bed due to illness or injury during the 12 months prior to the interview. In addition, respondents were asked to compare their health now (whether it is better, worse, or the same) to their health 12 months ago.

### Limitation of Functional Activities

The functional limitation questions in the AHS section asked the respondent to indicate the *degree of difficulty* he/she would have in performing specific physical tasks (e.g., walking a quarter of a mile, walking up ten steps, standing for two hours, carrying a ten pound object, etc.), and engaging in social activities and recreation (e.g., going shopping, attending club meetings, visiting friends, sewing, reading, etc.) without the assistance of another person or using special equipment. This is in sharp contrast to the questions in the FHS section, which allow only “yes” or “no” responses to questions inquiring whether household members needed help from another person with personal care needs (e.g., bathing, dressing, eating, etc.) or in handling routine tasks (doing everyday chores or shopping).

As in FHS, if the sample adult reported difficulty with any of these 12 activities, he/she was then asked what condition(s) cause the difficulty, as well as how long he/she has had the condition. The format of these condition data is similar to that found in the FHS section.

### Conditions

Each sample adult indicating any functional limitation (regardless of the degree of the limitation) is asked about the condition(s) or health problem(s) associated with that limitation, as well as the amount of time he/she has had the condition. Sample adults were given the following fixed response categories: “vision/problem seeing,” “hearing problem,” “arthritis/rheumatism,” “back or neck problem,” “fractures, bone/joint injury,” “other injury,” “heart problem,” “stroke problem,” “hypertension/high blood pressure,” “diabetes,” “lung/breathing problem,” “cancer,” “birth defect,” “mental retardation,” “other developmental problem (e.g., cerebral palsy),” “senility,” “depression/anxiety/emotional problem,” and “weight problem. Starting in 2001 and continuing in 2007, if the sample adult was limited by a condition not listed in one of these 18

fixed categories, the interviewer accessed a second screen containing 17 additional condition categories and two “other impairment problem” categories. These conditions were not read aloud to respondents, but if the sample adult’s condition was limited by one of these 17 conditions, the interviewer recorded this information. If the sample adult was limited by a condition not included in one of the 18 fixed categories or on the interviewer’s computer screen, then the interviewer entered a 50-character verbatim response for one or both of the “other impairment problem” categories.

The AHS condition data were edited very much like the condition data in FHS. The verbatim responses recorded by interviewers in one or both of the 50-character fields indicating “other impairment problem,” as well as those in the 17 additional general categories seen by the interviewers, were subsequently analyzed during data processing. While most respondents named “other” conditions that did not fall into the 18 fixed response categories as originally specified in the instrument, some respondents named conditions that should have been included in one of the fixed categories. In the latter case, these “other” responses were assigned codes during data processing corresponding to the appropriate category. An additional 16 *ad hoc* categories were created, and were assigned numbers 19\_ thru 34\_. (Note: Due to a naming convention error in 2002 and 2003 these same *ad hoc* categories were assigned numbers 19 thru 34 without an underscore.) Any verbatim conditions that could not be back-coded to one of the 18 fixed categories or recoded to one of the *ad hoc* categories remained in the “other impairment” categories, and were renumbered “90” and, if necessary, “91.” In addition, responses in the 17 general categories seen only by the interviewer were also back-coded and categorized into 8 of the *ad hoc* categories. The resulting 34 output categories were generally based on the International Classification of Diseases, Ninth Revision, Clinical Modification (see the FHS section).

These specific condition categories were subsequently transformed into variables indicating whether or not the condition was responsible for the respondent’s difficulty with any functional activity (a mention/not-mention format). Because the 16 *ad hoc* categories were not included on the flash cards given to respondents during the course of the interview, it is possible that frequencies obtained for these conditions may be underestimates. Therefore, these variables should be analyzed with care. Moreover, none of the AHS condition variables (AFLHCA1 through AFLHCA34\_) should be used to estimate prevalence rates for the conditions they represent, because only those sample adults with a previously reported functional limitation were eligible for the condition questions that followed. Analysts who are interested in estimating the prevalence of particular conditions are referred to the Sample Adult Conditions (ACN) section (above).

## Recodes

The recode FLA1AR is a summary measure that indicates sample adults who reported *any* difficulty with *any* one or more of the functional activities discussed during the course of the AHS section of the interview. In other words, individuals who indicated *any* degree of difficulty in FLWALK, FLCLIMB, FLSTAND, FLSIT, FLSTOOP, FLREACH, FLGRASP, FLCARRY, FLPUSH, FLSHOP, FLSOCL, *or* FLRELAX are coded “1” for FLA1AR. This variable includes three response levels: “1” for limited, “2” for not limited, and “3” for unknown if limited. ALCHRONR is based on FLA1AR but adds the additional criterion of whether at least one of the reported causal conditions is a chronic condition. The AHS section also includes time recodes and

chronic recodes for each of the 36 categories, which are very similar to those used in the FHS section described above.

#### Technical Notes

The condition variable AFLHCA31\_ includes any causal condition that specifically mentioned “surgery” or “operation,” or otherwise indicates a medical treatment as the causal condition (either ongoing or occurring within the last year). The condition variable AFLHCA33\_ includes any causal condition that specifically and solely mentioned “fatigue,” “weakness,” “lack of strength,” “tiredness,” “exhaustion,” etc. without reference to any particular part of the body. Lastly, the condition variable AFLHCA34\_ includes any causal condition that specifically and solely mentioned “pregnancy,” “pregnant,” or “childbirth.”

#### **IV. Adult Health Behaviors Section (AHB)**

The AHB section of the NHIS Sample Adult questionnaire contains questions related to cigarette smoking, leisure-time physical activity, alcohol use, height, weight, and sleep. With the exception of a question added in 2004 on sleep, all health behavior questions have been in the NHIS Sample Adult core questionnaire since 1997.

##### Smoking

Current smokers are defined as persons who have ever smoked 100 cigarettes and who currently smoke every day or some days. Since 2004, there is only one smoking status recode on the data file (SMKSTAT2), rather than three recodes during data years 1997-2003.

##### Leisure-time Physical Activity

The section on leisure-time physical activity is introduced with the following statement: “The next questions are about physical activities (exercise, sports, physically active hobbies...) that you may do in your LEISURE time.” In this section, respondents are asked to summarize their *usual* leisure-time physical activity – both in terms of frequency and duration. This requires some mental calculations by the respondent. Responses can be offered in terms of any time unit the respondent volunteers (times per day, per week, per month, or per year). A recode converting all responses into frequency in times per week is provided for each type of activity. The set of leisure-time physical activity questions included every year in the sample adult core module is: frequency and duration of vigorous activities, frequency and duration of light or moderate activities, and frequency of strengthening activities. The questions on leisure-time physical activity are used for tracking Healthy People 2010 Objectives 22.1-22.4 and in NHIS Early Release and in *Health, United States* (beginning in 2005).

##### Alcohol Use

Lifetime drinking status was assessed for all sample adults. Questions related to current drinking behavior were asked of all respondents who had had at least 12 drinks in their lifetime. Respondents were permitted to answer in terms of the number of days they drank per week, per

month, or per year. Standardized variables that convert the various time unit responses to standardized units-- that is, days per week (ALC12MWK) and days per year (ALC12MYR)- are provided.

A question asking how often the respondent had five or more drinks in one day during the past year was asked of all adults who drank at least once in the past year. The responses were not edited for consistency with the respondent's *usual* quantity or frequency of alcohol consumption because there was no basis for evaluating which one might be the more accurate. Note that the questions related to quantity of alcohol consumption are phrased in terms of the number of drinks consumed in a *day* and not the number of drinks consumed *at a sitting*.

ALCSTAT, a new recode begun in 2004, classifies lifetime and current drinking status for all sample adults. It replaced ALCSTAT1 (1997-2003) and ALC7STAT (2001-2003) and captured, in a single variable, all of the information contained in these two earlier recodes.

ALCSTAT is consistent with the classification of lifetime and current drinking status shown annually in *Health, United States, 2007*. The category "current drinker, level unknown" is slightly different from the category of the same name in the earlier variable, ALC7STAT. Since 2004, adults who said they did not know how often they drank were not asked the question about usual number of drinks (ALCAMT) and are classified as "drinking status unknown" in ALCSTAT. In contrast, in the earlier variable (ALC7STAT), adults who said they did not know the frequency of their alcohol consumption were asked the question about number of drinks (ALCAMT); those few (less than 0.5% of sample adults) who answered the second question without having answered the first were classified as "current drinkers, level unknown" in ALC7STAT.

Since 2004, the category "former drinker, frequency unknown" (ALCSTAT=4) includes former drinkers for whom information is not available on whether or not they had 12 or more drinks in any one year. Previously, in ALC7STAT, this category of former drinker was combined with "unknown drinking status" (ALC7STAT=9). ALCSTAT can be created by the data user relatively easily for data years in which both ALCSTAT1 and ALC7STAT appear (2001-2003). Creating ALCSTAT for data years prior to 2001 can be done, but the coding is quite complex.

## Body Weight and Height

Sample adults were asked their current height and weight. **No physical measurements were taken.** In cases where very large or very small values were reported for either height or weight, the data for both variables were changed to "96" or "996" ("Not available") on the public use data file. This was done in order to protect the confidentiality of NHIS respondents who might be identifiable by their unusual physical characteristics. National estimates based on physical measurements, such as those available from NCHS' National Health and Nutrition Examination Survey (NHANES), may differ from those available from the NHIS, which are self-reported.

Respondents have the option of reporting their height and weight in either U.S. Customary (lbs/oz; ft/in) or metric (kg; m/cm) format. Less than 1% of respondents report in metric format. Metric responses on height and weight are converted into U.S. Customary format for inclusion on the micro data file. Since 2006, the factor used to convert metric values from

centimeters to inches has been expanded to 2.54 for greater precision. The conversion factor was rounded to 2.5 during 1997-2005. For the earlier data years, estimates of height in feet and inches will be slight overestimates for respondents who initially reported their height in meters and/or centimeters (e.g., n=216, or less than 1% of adults in 2005).

Body Mass Index (BMI), a measure of body weight relative to height, was calculated using the formula:  $BMI = \text{kilograms} / \text{meters}^2$ . Kilograms and meters were derived from (U.S. Customary) pounds and inches using the following factors: 1 kilogram=2.20462 pounds; 1 meter=39.37008 inches. BMI was calculated for all persons who provided height and weight, including those for whom specific height and weight values were changed to “96” and “996” (not available) on the public use file for reasons of confidentiality. The values for the BMI include two implied decimals.

Individual values for height and weight have been included in the public use data files as long as the values did not fall at the extremes— the lowest 1½ percent or highest 1½ percent of records— in order to protect respondent confidentiality. Publicly releasable ranges for weight changed beginning in 2006. For details concerning these and other changes, especially regarding body weight, please review the section entitled “Body Weight and Height” within the AHB section of the 2006 NHIS Survey Description Document and Appendix VIII.

The following classification of body weight status for both men and women, established by the World Health Organization, is used in the NHIS data files: underweight (BMI < 18.5); healthy weight (18.5 <= BMI < 25); overweight, but not obese (25 <= BMI < 30); overweight, including obese (BMI >= 25); and obese (BMI >= 30).

## Sleep

A question asking about usual number of hours of sleep, first introduced in the Sample Adult Core in 2004, continues in 2007. Prior to 2004, a question about sleep was last asked in the NHIS in 1990 as part of the Health Promotion and Disease Prevention Supplement.

## **V. Adult Health Care Access and Utilization Section (AAU)**

The core Adult Health Care Access and Utilization (AAU) section of the 2007 NHIS has remained largely unchanged since 1997 and contains information on access to health care, dental care, health care provider contacts, and immunizations.

Questions regarding access to health care include having a usual place for sick care, having a usual place for routine/preventive care, change in the place of care, any delays in getting medical care, and instances of being unable to afford medical care. The question about the reason for delaying care focused on such access issues as transportation, getting an appointment, and waiting time prior to actually seeing the doctor. A question on dental care asked about the length of time since last dental visit.

Respondents were asked about health care provider contacts, including questions about doctor contacts during the past 12 months. Doctor visit probe questions allow for visits not only from medical doctors but from a variety of other health care professionals, including chiropractors. Questions about home care are included as well as a question asking about the number of visits to a hospital emergency room in the past 12 months. There is also a question that asks how long it has been since the respondent has seen or talked to a doctor.

There are several questions related to adult immunizations: flu shot including the month and year received, nasal spray flu vaccine including the month and year received, pneumonia vaccination, and the hepatitis B vaccine. Additional questions inquire whether the respondent had ever had chickenpox or had chickenpox in the past 12 months, had ever had hepatitis, or had ever lived with someone with hepatitis.

### Technical Notes

Analysts are advised to read the notes in the Dataset Documentation for further information pertaining to any changes that may have occurred and to compare the 2007 Dataset Documentation to documentation from the 2006 (and earlier) NHIS for any other changes that may have occurred over time to the variables in this section.

## **VI. Adult AIDS Section (ADS)**

This section contains a series of questions related to testing for HIV, the virus that causes AIDS. Respondents were asked whether they had ever donated blood and whether they had a blood test for HIV, their main reasons for getting or not getting tested, when they had their last test, the number of times they had been tested, and where (the location/facility) the testing was done. The section also contains questions on respondents' plans for being tested in the future and their reasons for those plans, as well as their perceived personal risk for getting AIDS. With some modifications and additions, these questions are similar to those asked in the AIDS Knowledge and Attitudes Supplements that were included in the NHIS from 1987 to 1995.

Beginning in 2000, questions on sexually transmitted diseases (STDs) and tuberculosis (TB) are included in this section. These questions asked respondents whether they had an STD other than HIV or AIDS, whether they saw a doctor or health professional, and the location/facility to which they went to be checked. In addition to STD questions, respondents were also asked about TB, whether they had heard of it, how much they knew about it, and if they knew anyone personally with the disease. In addition, respondents were asked about their perceived personal risk of getting TB, and if, in their opinion, TB could be cured.

## **2007 National Health Interview Survey Adult Complementary and Alternative Medicine File (ALT)**

The Adult Complementary and Alternative Medicine Supplement (ALT) collected information from sample adults on their use of 18 non-conventional health care practices. The modalities covered in the supplement include acupuncture, ayurveda, biofeedback, chelation therapy, chiropractic or osteopathic manipulation, energy healing therapy, hypnosis, massage, naturopathy, traditional healers, movement therapies, herbal and non-vitamin supplements, vitamins and minerals, homeopathy, special diets, yoga/tai chi/qi gong, relaxation techniques, and prayer for health reasons.

While some questions were tailored to the specific health care practice, others were asked for most of the 18 modalities. The topics covered include: whether the respondent ever used the modality, and, if so, whether the modality was used in the past 12 months; number of times the respondent had seen a practitioner for the modality; amount paid out-of-pocket for the modality; whether the modality was used to treat a specific health problem or condition, and, if so, what health problems or conditions were treated and for which one of the health problems or conditions the modality was used the most; whether the respondent received any of a specified set of traditional medical treatments for the condition treated most by the modality, and, if so, when the conventional medical treatments were received in relation to when the modality was used; reasons for using the modality; and disclosure of use to a specified set of conventional medical professionals. Given the number of modalities and questions, a separate stand-alone data file was created for the supplement rather than append it to the 2007 Sample Adult file. The Sample Adult weight is included on the ALT file and should be used for all the variables on this file.

Although all sample adults were to be asked the questions in the ALT Supplement, some persons did not complete the supplement. Persons who did not give responses to any of the questions in the supplement had a value of “not ascertained” inserted in the appropriate fields in this file. These persons were given a coded value of “1” for the record completion status variable (COMPSTAT), which indicates that their record only contains responses of “not ascertained.” In other words, these respondents (n=556, or 2.4% of sample adults) are retained in the file, but they are coded as “8” in all the relevant fields of the file. All persons who met the criteria for completing the supplement were given a coded value of “2” (all answers refused or don’t know), “3” (all answers refused, DK, or not ascertained), or “4” (at least one valid answer) for the record completion status variable.

### **Back Coding**

In the ALT Supplement, respondents were given the opportunity to choose “Other specify” as a response to the questions about what health problems or conditions were treated with specific types of complementary and alternative medicine (CAM). If “Other specify” was selected, the respondent was then asked to give a verbatim response indicating the health problem or condition treated with the specified type of CAM. After reviewing the verbatim

responses, it was found that some of these responses could be back coded (added) into existing categories, while other responses occurred so frequently that additional categories needed to be created.

In the ALT file, five additional health problem or condition response categories (anemia and other blood disorders, other allergies, other infectious diseases and problems of the immune system, other musculoskeletal, and stress) were added to the existing health problem or condition list for each therapy, as part of the editing process. When possible, verbatim responses were back coded into existing categories or one of the new categories. In all cases, the original verbatim text was retained in the original variables. As a result, the frequency count for the "other specify" variable for each CAM therapy may not match the frequency for its corresponding verbatim variable. Also, the frequency counts for the questions about the condition treated most often with each CAM therapy have not been changed to reflect back coding the "other specify" responses.

## **2007 National Health Interview Survey Coverage Section: Telephone Questions**

The purpose of the cellular telephone questions is to track the prevalence of wireless telephones in American families over time, allowing researchers to analyze the demographic characteristics of families who have substituted wireless service for landline home telephones. Having these data from a large population-based survey such as the NHIS provides useful information about potential bias from undercoverage in random digit dial telephone surveys, using only land-line telephone numbers in their sampling frames.

In 2007 the cellular telephone questions were modified. In 2003-2006 all cellular phone questions resided in the Recontact Section (REC) at the end of the survey; in 2007 all cellular phone questions were moved to the Coverage Section (COV.330 – COV.337) at the beginning of the survey. Appendix VIII lists the cellular telephone variable changes. The 2007 survey instrument can be found on the DHIS Web site at: <http://www.cdc.gov/nchs/nhis.htm>.

## Guidelines for Citation of Data Source

With the goal of mutual benefit, the National Center for Health Statistics (NCHS) requests that recipients of NHIS data files cooperate in certain actions related to their use. Any published material derived from the 2007 data should acknowledge CDC/NCHS, National Health Interview Survey as the original source. The suggested citation to appear at the bottom of all tables and graphs is as follows:

Data Source: CDC/NCHS, National Health Interview Survey, 2007

In a bibliography, the suggested citation should read:

National Center for Health Statistics. *Data File Documentation, National Health Interview Survey, 2007 (machine readable data file and documentation)*. National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, Maryland. 2008.

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

Achenbach and Edelbrock. *Manual for the Child Behavior Checklist and Revised Child Behavior Profile*. University of Vermont, Department of Psychiatry. 1983.

DeNavas-Walt C, Proctor BD, Smith J. U.S. Census Bureau. Current Population Reports, P60-233. *Income, Poverty, and Health Insurance Coverage in the United States: 2006*. U.S. Government Printing Office. Washington, DC. 2007.

Metropolitan Life Insurance Company. 1983 Metropolitan height and weight tables. *Statistical Bulletin Metropolitan Insurance Company* 64:1. 1983.

National Center for Health Statistics. *Health, United States, 2004*. Hyattsville, Maryland. 2004.

National Center for Health Statistics. *National Health Interview Survey: Research for the 1995-2004 Redesign*. Vital Health Stat 2(126). 1999.

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

U.S. Department of Health and Human Services. *Healthy People 2010. 2nd Ed. With Understanding and Improving Health and Objectives for Improving Health*. 2 vols. Washington: U.S. Government Printing Office, November, 2000.

Warner M, Schenker N, Heinen MA, Fingerhut LA. The effects of recall on reporting injury and poisoning episodes in the National Health Interview Survey. *Injury Prevention*, 11(5):282-7. 2005.

## Appendix I

### Calculation of Response Rates for the 2007 NHIS

The 1997 redesigned NHIS incorporated a change from the previous paper and pencil questionnaire to a new computer assisted personal interviewing (CAPI) system. The response rates calculated here pertain to the Basic Module questions in the 2007 NHIS.

The Basic Module collects basic information on the household and all family members. In addition, for each family, more detailed information is collected on one sample adult, and one sample child, if any.

#### Household Response Rate

$$\frac{\text{(Interviewed Households)}}{\text{(Interviewed Households + Type A Non-Response Households)}}$$

The Household (HH) response rate is calculated by dividing the number of responding households by the number of households that are in-scope or eligible for the survey. Note that Type A non-response households are eligible households that were not interviewed for a variety of reasons: language problems; no one was at home after repeated contact attempts; family temporarily absent; refusal; household records rejected for insufficient data; household records rejected for other CAPI related problems; or other reasons for no interview.

#### Conditional Family Response Rate

$$\frac{\text{(Interviewed Families)}}{\text{(Interviewed Families + Rejected Families(from Interviewed HHs))}}$$

Family Core data were collected from the respondent about all persons in the family. The response rates for the Family Core can be calculated in two ways: conditionally and finally. The conditional Family response rate is the rate only for those families identified as eligible and does not take into account household non-response. The conditional Family response rate is calculated by dividing the number of responding families by the number of families that are eligible for the survey, that is, from interviewed households. Note that a household can have multiple families, and rejected families are families that were deleted from interviewed households because of insufficient data.

### **Final Family Core Response Rate**

$$\frac{\text{(Interviewed Families)}}{\text{(Interviewed + Rejected Families (from Interviewed HHs))}} \text{(Household Response Rate)}$$

The final Family response rate is the rate for those families identified as eligible that takes into account household non-response. The final Family response rate is calculated by dividing the number of responding families by the number of families that are eligible for the survey, that is, from interviewed households, and then multiplying this quotient by the Household response rate.

### **Conditional Sample Child Response Rate**

$$\frac{\text{(Interviewed Sample Children)}}{\text{(Eligible Sample Children)}}$$

The response rates for the Sample Child section can be calculated in two ways: conditionally and finally. The conditional Sample Child response rate is the rate only for sample children and does not take into account household or family non-response. The conditional Sample Child response rate is calculated by dividing the number of responding sample children by the number of eligible sample children from interviewed families

### **Final Sample Child Response Rate**

$$\frac{\text{(Interviewed Sample Children)}}{\text{(Eligible Sample Children from Interviewed Families)}} \text{(Final Family Response Rate)}$$

The final Sample Child response rate is the rate for sample children that takes into account household and family non-response. The final Sample Child response rate is calculated by dividing the number of responding sample children by the number of sample children who are eligible for the survey, that is, from interviewed families, and then multiplying this quotient by the final Family response rate.

### Conditional Sample Adult Response Rate

$$\frac{\text{(Interviewed Sample Adults)}}{\text{(Eligible Sample Adults)}}$$

The response rates for the Sample Adult section can be calculated in two ways: conditionally and finally. The conditional Sample Adult response rate is the rate only for those sample adults identified as eligible and does not take into account household or family non-response. The conditional Sample Adult response rate is calculated by dividing the number of responding sample adults by the number of eligible sample adults from interviewed families.

### Final Sample Adult Response Rate

$$\frac{\text{(Interviewed Sample Adults)}}{\text{(Eligible Sample Adults from Interviewed Families)}} \text{(Final Family Response Rate)}$$

The final Sample Adult response rate is the rate for those sample adults identified as eligible that takes into account household and family non-response. The final Sample Adult response rate is calculated by dividing the number of responding sample adults by the number of sample adults who are eligible for the survey, that is, from interviewed families, and then multiplying this quotient by the final Family response rate.

**Appendix I, Table 1. Response Rates for the 2007 NHIS**

Household	87.1%
Family - Conditional	99.4%
Family - Final	86.6%
Sample Child - Conditional	88.4%
Sample Child – Final	76.5%
Sample Adult - Conditional	78.3%
Sample Adult – Final	67.8%

## Calculation of Response Rates for Combined NHIS Data Years

The response rates for combined NHIS data years are calculated in the same basic way as for a single year. The following examples are shown for two years of data. Similar methods apply for multiple years of data in 1997 and beyond.

### Household Response Rate for Combined Data Years

$$\frac{\text{(Interviewed Households for Years 1 and 2)}}{\text{(Interviewed Households for Years 1 and 2 + Type A Non-Response Households for Years 1 and 2)}}$$

The Household response rate for combined data years is calculated by dividing the number of responding households for Years 1 and 2 by the number of households that are in-scope or eligible for the survey for Years 1 and 2. Note that Type A non-response households are eligible households that were not interviewed for a variety of reasons: language problems; no one was at home after repeated contact attempts; family temporarily absent; refusal; household records rejected for insufficient data; household records rejected for other CAPI related problems; or other reasons for no interview.

### Conditional Family Response Rate for Combined Data Years

$$\frac{\text{(Interviewed Families for Years 1 and 2)}}{\text{(Interviewed Families for Years 1 and 2 + Rejected Families for Years 1 and 2)}}$$

Family Core data were collected from the respondent about all persons in the family. The response rates for the Family Core can be calculated in two ways: conditionally and finally. The conditional Family response rate is the rate only for those families identified as eligible and does not take into account household non-response. The conditional Family response rate for combined data years is calculated by dividing the number of responding families in Years 1 and 2 by the number of families that are eligible for the survey in Years 1 and 2, that is, from interviewed households in Years 1 and 2. Note that a household can have multiple families, and rejected families are families that were deleted from interviewed households because of insufficient data.

### Final Family Response Rate for Combined Data Years

$$\frac{(\text{Interviewed Families for Years 1 and 2})}{\left( \begin{array}{l} \text{Interviewed Families for Years 1 and 2+} \\ \text{Rejected Families from Interviewed HHs for Years 1 and 2} \end{array} \right)} \left( \begin{array}{l} \text{Household Response} \\ \text{Rate for Years 1 and 2} \end{array} \right)$$

The final Family response rate is the rate for those families identified as eligible that takes into account household non-response. The final Family response rate for combined data years is calculated by dividing the number of responding families for Years 1 and 2 by the number of families that are eligible for the survey for Years 1 and 2, that is, from interviewed households for Years 1 and 2, and then multiplying this quotient by the Household response rate for the combined data years.

### Conditional Sample Child Response Rate for Combined Data Years

$$\frac{(\text{Interviewed Sample Children for Years 1 and 2})}{(\text{Eligible Sample Children for Years 1 and 2})}$$

The response rates for the Sample Child section can be calculated in two ways: conditionally and finally. The conditional Sample Child response rate is the rate only for sample children and does not take into account household or family non-response. The conditional Sample Child response rate for combined data years is calculated by dividing the number of responding sample children for Years 1 and 2 by the number of eligible sample children from interviewed families for Years 1 and 2.

### Final Sample Child Response Rate for Combined Data Years

$$\frac{(\text{Interviewed Sample Children for Years 1 and 2})}{(\text{Eligible Sample Children from Interviewed Families for Years 1 and 2})} \left( \begin{array}{l} \text{Final Family Response} \\ \text{Rate for Years 1 and 2} \end{array} \right)$$

The final Sample Child response rate is the rate for sample children that takes into account household and family non-response. The final Sample Child response rate for combined data years is calculated by dividing the number of responding sample children for Year 1 and Year 2 by the number of sample children who are eligible for the survey, that is, from interviewed families for Year 1 and Year 2, and then multiplying this quotient by the final Family response rate for the combined data years.

**Conditional Sample Adult Response Rate for Combined Data Years**

$$\frac{\text{(Interviewed Sample Adults for Years 1 and 2)}}{\text{(Eligible Sample Adults for Years 1 and 2)}}$$

The response rates for the Sample Adult section can be calculated in two ways: conditionally and finally. The conditional Sample Adult response rate is the rate only for those sample adults identified as eligible and does not take into account household or family non-response. The conditional Sample Adult response rate for combined data years is calculated by dividing the number of responding sample adults for Years 1 and 2 by the number of eligible sample adults from interviewed families for Years 1 and 2.

**Final Sample Adult Response Rate for Combined Data Years**

$$\frac{\text{(Interviewed Sample Adults for Years 1 and 2)}}{\text{(Eligible Sample Adults from Interviewed Families for Years 1 and 2)}} \left( \begin{array}{l} \text{Final Family Response} \\ \text{Rate for Years 1 and 2} \end{array} \right)$$

The final Sample Adult response rate is the rate for those sample adults identified as eligible that takes into account household and family non-response. The final Sample Adult response rate for combined data years is calculated by dividing the number of responding sample adults for Years 1 and 2 by the number of sample adults who are eligible for the survey, that is, from interviewed families for Years 1 and 2, and then multiplying this quotient by the final Family response rate for the combined data years.

**Appendix I, Table 2. Number Eligible/Interviewed, 2007 NHIS**

<b>File</b>	<b>Eligible</b>	<b>Interviewed</b>
<b>Household</b>	33,615	29,266
<b>Family/Person</b>	30,081	29,915
<b>Sample Child</b>	10,658	9,417
<b>Sample Adult</b>	29,875	23,393

**Appendix I, Table 3. Number Eligible/Interviewed, 2006 NHIS**

<b>File</b>	<b>Eligible</b>	<b>Interviewed</b>
<b>Household</b>	33,468	29,204
<b>Family/Person</b>	29,974	29,868
<b>Sample Child</b>	10,853	9,837
<b>Sample Adult</b>	29,825	24,275

**Appendix I, Table 4. Number Eligible/Interviewed, 2005 NHIS**

<b>File</b>	<b>Eligible</b>	<b>Interviewed</b>
<b>Household</b>	44,540	38,509
<b>Family/Person</b>	39,463	39,284
<b>Sample Child</b>	13,906	12,523
<b>Sample Adult</b>	39,227	31,428

**Appendix I, Table 5. Number Eligible/Interviewed, 2004 NHIS**

<b>File</b>	<b>Eligible</b>	<b>Interviewed</b>
<b>Household</b>	42,089	36,579
<b>Family/Person</b>	37,653	37,466
<b>Sample Child</b>	13,538	12,424
<b>Sample Adult</b>	37,388	31,326

**Appendix I, Table 6. Number Eligible/Interviewed, 2003 NHIS**

<b>File</b>	<b>Eligible</b>	<b>Interviewed</b>
<b>Household</b>	40,266	35,921
<b>Family/Person</b>	37,126	36,573
<b>Sample Child</b>	13,275	12,249
<b>Immunization</b>	13,275	11,665
<b>Sample Adult</b>	36,524	30,852

**Appendix I, Table 7. Number Eligible/Interviewed, 2002 NHIS**

<b>File</b>	<b>Eligible</b>	<b>Interviewed</b>
<b>Household</b>	40,377	36,161
<b>Family/Person</b>	37,458	36,831
<b>Sample Child</b>	13,570	12,524
<b>Immunization</b>	13,865	13,611
<b>Sample Adult</b>	36,787	31,044

**Appendix I, Table 8. Number Eligible/Interviewed, 2001 NHIS**

<b>File</b>	<b>Eligible</b>	<b>Interviewed</b>
<b>Household</b>	43,797	38,932
<b>Family/Person</b>	40,227	39,633
<b>Sample Child</b>	14,766	13,579
<b>Immunization</b>	15,000	14,709
<b>Sample Adult</b>	39,564	33,326

**Appendix I, Table 9. Number Eligible/Interviewed, 2000 NHIS**

<b>File</b>	<b>Eligible</b>	<b>Interviewed</b>
<b>Household</b>	43,437	38,633
<b>Family/Person</b>	39,998	39,264
<b>Sample Child</b>	14,711	13,376
<b>Immunization</b>	14,890	14,618
<b>Sample Adult</b>	39,201	32,374

**Appendix I, Table 10. Number Eligible/Interviewed, 1999 NHIS**

<b>File</b>	<b>Eligible</b>	<b>Interviewed</b>
<b>Household</b>	42,882	37,573
<b>Family/Person</b>	38,845	38,171
<b>Sample Child</b>	14,217	12,910
<b>Immunization</b>	14,178	13,881
<b>Sample Adult</b>	38,117	30,801

**Appendix I, Table 11. Number Eligible/Interviewed, 1998 NHIS**

<b>File</b>	<b>Eligible</b>	<b>Interviewed</b>
<b>Household</b>	42,440	38,209
<b>Family/Person</b>	39,559	38,773
<b>Sample Child</b>	14,619	13,645
<b>Prevention Sample Child</b>	13,645	13,610
<b>Immunization</b>	15,041	14,775
<b>Sample Adult</b>	38,729	32,440
<b>Prevention Sample Adult</b>	32,440	31,882

**Appendix I, Table 12. Number Eligible/Interviewed, 1997 NHIS**

<b>File</b>	<b>Eligible</b>	<b>Interviewed</b>
<b>Household</b>	43,370	39,832
<b>Family/Person</b>	41,291	40,623
<b>Sample Child</b>	15,244	14,290
<b>Immunization</b>	15,558	15,402
<b>Sample Adult</b>	40,552	36,116

## Appendix II

### Race and Hispanic Origin in the 2007 NHIS

#### Background

For over 20 years, the National Health Interview Survey (NHIS) had collected information on the race and Hispanic origin or ethnicity of its respondents, following guidelines set forth by the Office of Management and Budget in a policy known as OMB Directive 15 (Office of Management and Budget, 1977). The NHIS relied on respondents to provide self-identified race and ethnicity information (proxy information is reported for children and non-present household members), although interviewer-observed race was also recorded through 1996, the last year of the paper questionnaire. NHIS data are routinely tabulated by race and ethnicity in NCHS publications such as Summary Health Statistics, Health U.S., and Advance Data reports.

In response to the changing demographics of the U.S. population, the OMB revised Directive 15 in 1997 after an extensive period of research and public commentary. The new race and ethnicity standards allow respondents to the Census and federal surveys to indicate more than one race group in answering questions on race. A complete description of the new OMB guidelines on the collection of racial and ethnic data, including descriptions of the new race categories, the ordering of race and ethnicity questions, and guidelines for the tabulation and publication of data under the new standards, can be found on the OMB Web site: <http://www.whitehouse.gov/omb/inforeg/statpolicy.html>. In accordance with this requirement, the NHIS became fully compliant with the new race and ethnicity standards with the fielding of the 1999 questionnaire, although the NHIS had been following some aspects of the new guidelines for many years. This policy was expected to be fully implemented across the federal statistical system beginning with the 2003 calendar year.

As noted previously, the U.S. Census Bureau is the data collection agent for the NHIS, as it is for a number of other federal surveys. The Census Bureau also provides the control totals for race/ethnicity (along with sex and age) that are used in the post-stratification adjustment of the person weights in the NHIS data file. In order to maintain consistency with the Census Bureau procedures for collecting and editing data on race and ethnicity, the NHIS made major changes to its editing procedures in the 2003 data year. Beginning in the 2003 NHIS, “Other race” was no longer available as a separate race response. This response category was treated as missing, and the race was imputed if this was the only race response. In cases where “Other race” was mentioned along with one or more OMB race groups, the “Other race” response was dropped, and the OMB race group information was retained. These editing changes are consistent with the procedures that the Census Bureau uses to create the Modified Race Data Summary File, which is the data file that provides the population control totals used in weighting the NHIS data. More information about the Modified Race Data Summary File and the editing procedures used to create it, can be found at the following Web site: <http://www.census.gov/popest/archives/files/MRSF-01-US1.html>. These editing procedures remain in effect for the 2007 data file. Please refer to the 2007 Variable Layout Report for more information.

## **Race and Hispanic Origin Questions in the National Health Interview Survey**

The 2007 NHIS included two questions about Hispanic Origin:

“Do/Does {you/name} consider {yourself / himself / herself} Hispanic / Latino?” [HHC.170], and

“Please give me the number of the group that represents your Hispanic Origin or ancestry” [HHC.180; response categories shown to the respondent on a flashcard].

There were no changes in the wording of the 2007 Hispanic origin question, but some responses were imputed, and the variable name is labeled as HISPAN\_I to indicate this fact (see section on the imputation of race and ethnicity later in this document).

The 2007 NHIS also included two questions to obtain information on a respondent’s race:

“What race {does/do} {name/you} consider {himself/herself/yourself} to be? Please select one or more of the following groups.” [HHC.200; response categories shown to the respondent on a flashcard], and

“Which one of these groups, that is (FR: READ GROUPS) would you say BEST represents {your/name’s} race?” [HHC.220; response categories given are read back to the respondent by the interviewer].

The first question is asked of all respondents, while the second question is asked only of those respondents who give more than one response to the first question. Although the wording and placement of these two questions are essentially the same as they had been in the NHIS for many years, there were changes made to the response categories beginning in 1999. In compliance with the new race and ethnicity data collection standards, the category “Asian and Pacific Islander” is now split into two categories, “Asian” and “Native Hawaiian and Other Pacific Islander.” Because confidentiality regulations on minimum sample size do not permit the NHIS to release data for Native Hawaiians and Other Pacific Islanders or some Asian subgroups separately, public use data are provided for the three largest Asian subpopulation groups, while the “Other Pacific Islander” and “Other Asian” categories combine the remaining groups that cannot be shown separately.

### **2007 Race and Hispanic Origin Variables**

The following table (Appendix II, Table I) summarizes the Hispanic origin and race variables in the 2007 data file. Details on the specific response categories for the race questions and additional details on these variables can be found in the 2007 public use Variable Layout Report, and users are strongly urged to read these descriptions carefully to determine how and when the variables should be used in analysis. Data users are also encouraged to check the Variable Frequency Report to examine the unweighted data for these variables before computing weighted estimates.

**Appendix II, Table 1. 2007 NHIS Race/Ethnicity Variable Names and Description**

<b>2007 Variable Name</b>	<b>Description</b>
<b>ORIGIN_I</b>	Hispanic origin/ancestry with imputed values for some records
<b>ORIGIMPT</b>	Hispanic origin imputation flag
<b>HISPAN_I</b>	Type of Hispanic origin/ancestry with imputed values for some records
<b>HISPIMPT</b>	Type of Hispanic origin imputation flag
<b>RACERPI2</b>	Contains 4 of 5 OMB race groups; values imputed for some records. Does not include “Other race” category.
<b>MRACRPI2</b>	Detailed race variable; multiple race persons not selecting a primary race group in separate category. Values were imputed for some records. Does not include “Other race” category.
<b>MRACBPI2</b>	See section below on bridging; values were imputed for some records. “Other race” category included for bridging purposes.
<b>RACRECI3</b>	Variable that contains 4 race categories used in post-stratification and weighting. New category added to reflect changes in sample design. Values imputed for some records.
<b>RACEIMP2</b>	Imputation flag for use in determining which cases were imputed for the race variables. New categories added to account for new editing procedures.
<b>HISCODI3</b>	Same categories as RACRECI3, crossed with ORIGIN_I (Hispanic/non-Hispanic); values were imputed for some records.
<b>ERIMPFLG</b>	Summary race/ethnicity imputation flag – indicates that either race or ethnicity or both race and ethnicity were imputed.

**Procedures for Imputation of Ethnicity and Race in the NHIS**

Prior to the 2000 NHIS, race recodes #1 and #2 were created using a crude imputation method that assigned a race to persons with missing values for the variable MAINRACE. Under these procedures, in the 1996 and earlier NHIS, if an observed race were recorded by the interviewer, it was used to code a race value. If there were no observed race value, all persons with a missing value for MAINRACE who were identified as Hispanic (on the Hispanic origin question) were coded as “White,” and those who were identified as non-Hispanic were coded as “Other race.” Beginning with the 1997 NHIS, observed race was no longer collected. Therefore, the race imputation procedures for all persons with missing values for the variable MAINRACE in 1997-1999 matched the imputation procedures for earlier years when no observed race values were recorded.

In an effort to improve the quality of data on ethnicity and race in the NHIS, hot-deck imputation of selected race and ethnicity variables was done for the first time in the 2000 NHIS and continued to be used for the 2007 NHIS data. Changes implemented in the 2003 imputation procedures remain in effect for 2007. Records for persons for whom “Other race” was the only race mentioned were treated as having missing data on race, and were added to the pool of records for which selected race and ethnicity variables are imputed.

The variables ORIGIN (whether or not the respondent is of Hispanic origin), HISPTY01-HISPTY10 (type of Hispanic origin), RACE1-RACE5 (each of 5 possible race mentions), and MAINRACE (primary race selection for persons reporting more than one race) with missing values were imputed (note that the pre-imputation variable names are used in this description because the names were not changed until the imputation was completed). The imputation was carried out in two stages.

Stage 1 imputation was used for households in which some persons had missing values and some persons had valid entries for ethnicity and race variables (imputation within households). Stage 1 imputation was based on the hot-deck imputation procedures developed for the Decennial Census Dress Rehearsal (conducted in 1998), which were adapted to utilize NHIS family relationship variables for imputation of the missing ethnicity and race data. Additional imputation procedures for “Other race” responses were adapted from the Census Bureau’s Modified Race Data Summary File editing specifications for use with the NHIS race data.

Stage 2 imputation was used for households in which all persons had missing values for ethnicity and race variables (imputation between households). The specifications obtained from Census which were the basis of Stage 1 imputation did not contain information on the imputation of race and ethnicity between households. Therefore, staff in DHIS and NCHS’ Office of Research and Methodology developed the specifications for the between-household imputation, using the secondary sampling unit (SSU) as the geographic unit for selecting donors.

**1. Stage 1 Imputation** - for households in which some persons had missing values, and some persons had valid entries for ethnicity and race variables.

- Step 1. Generate datasets based on NHIS household files for within-household imputation.
- Step 2. Preview the frequency distributions of the variables to be imputed.
- Step 3. Re-classify donors based on variables RRP (relationship of person to household reference persons) and DEGREE1-DEGREE7 (relationship variables - e.g., whether person is biological, step, foster, or in-law child of reference person).
- Step 4. Load donors' data to hot decks within each household, and conduct imputation for each donee in the same household. Donees are classified in twenty-six categories based on the relationship of the donees to the Reference Person in the household (see following section). The allocation sequence of donors for each type of donee is different, depending on the type of the donee, and the relationship between the donor and the donee.
- Step 5. Review the distributions of the imputed variables after imputation for comparison and analysis. Combine all records, and reclassify households for Stage 2 imputation.

**2. Stage 2 Imputation** - for households in which all persons had missing values.

- Step 1. The imputation was divided into three parts:
  - A) Imputation among Hispanic households (ORIGIN=1).
  - B) Imputation among Non-Hispanic households (ORIGIN=2).
  - C) Imputation for households with unknown Hispanic origin (ORIGIN=7, 8, 9).
- Step 2. Each part of the imputation complied with certain rules that are outlined in further detail in the Stage 2 imputation specification (not provided here). The combinations of imputed variables in each part are different.
- Step 3. After all imputations were completed, datasets from Stage 1 and Stage 2 were combined, records that were imputed were flagged for the in-house and public use data files, and comparisons of the distributions of the variables before and after imputation were examined.

### **Use of Imputation Flags**

Since hot-deck imputation procedures have been implemented on the NHIS race and ethnicity data, imputation flags have also been added to the data file. These flags allow data users to keep track of the number of cases for which race and/or ethnicity was imputed by the type of

original response. They also provide users with a means of accessing the data in their unimputed form. The flags also provide a mechanism for converting data back to the format in the data files prior to the implementation of imputation in 2000, which is critical for merging data files across survey years and maintaining trends in the data. There are four imputation flags on the 2007 public use data file: ORIGIMPT, HISPIMPT, RACEIMP2, and ERIMPFLG. These flags are described in Table 1 above.

**Users who wish to merge across data years or create trend data must recreate the race variables RACERPI2, MRACRPI2, and MRACBPI2 in the format they had in previous years by using the flag RACEIMP2.** Sample SAS code for using the imputation flags and merging across data years for the variable RACERPI2 (RACERP\_I in 2000-2002 and RACER\_P in 1999) is included below (the example uses 1999-2006 NHIS data, but other combinations of data years can be used with the appropriate adaptations to the code).

```
*****
*** merge 1999-2006 race variables using public use variables ****
*****;
*** recode 1999 data ***;

    if RACER_P in (97)          then RACEPU99=7;          /* refused */
else if RACER_P in (98)          then RACEPU99=8;          /* NA      */
else if RACER_P in (99)          then RACEPU99=9;          /* DK      */
else                                RACEPU99=RACER_P;

***** code to add imputed responses for 2000-2002 RACERP_I *****;

    if RACEIMPT in (1)          then RACP0002=7;          /* refused */
else if RACEIMPT in (2)          then RACP0002=8;          /* NA      */
else if RACEIMPT in (3)          then RACP0002=9;          /* DK      */
else                                RACP0002=RACERP_I;

***** code to add imputed responses for 2003-2006 RACERPI2 *****;

    if RACEIMP2 in (1)          then RACP0306=7;          /* refused */
else if RACEIMP2 in (2)          then RACP0306=8;          /* NA      */
else if RACEIMP2 in (3)          then RACP0306=9;          /* DK      */

else if RACEIMP2 in (4,5)        then RACP0306=5 ;          /* Other races1 */
else                                RACP0306=racerpi2;
```

1 Note that this category contains “Other race only,” “Unspecified Multiple race” and NHOPI persons.

```
*** combine 1999-2006 data into single variable *****;

    if RACEPU99 ne .            then RACE9906=RACEPU99;
else if RACP0002 ne .          then RACE9906=RACP0002;
else                            RACE9906=RACP0306;
```

## Bridging to the Old OMB Standards

The OMB tabulation guidelines for the new race and ethnicity standards recognize that the complete transition from the old standards to the new standards will take some time, and that many federal statistical systems have a primary mission to track data trends over time. During this transitional period, known as the “bridge,” it has been recommended that data systems tabulate data for publication under the new standards, while also providing a means for data users to bridge the new data back to the old standards. This will allow data users to examine differences, if any, in tabulating the data under the old and new standards, assist in the maintenance of data trends, and allow users to become accustomed to data tabulated under the new standard before the transition is complete. In the NHIS, the second race question (commonly known as the “follow-up question”) is used to create the bridge between data collected under the old standards and data collected under the new ones. The 2007 NHIS public use data file contains one bridge race variable to allow comparisons of 2007 data with data from previous years, and to enable merging the 2007 data with 1997-2006 data.

There was one major change to the race and ethnicity data in the 1999 NHIS (which is also true for 2000-2007) that occurred as a result of the creation of a bridge variable. NCHS confidentiality standards do not permit NCHS to release data that might lead to the inadvertent identification of individual respondents to the survey. Beginning with the 1999 survey (and continuing in 2007), data on “Asian” persons and “Native Hawaiian and Other Pacific Islander (NHOPI)” persons were collected separately according to the new OMB guidelines. Ideally, these two groups could be combined to recreate the old category “Asian and Pacific Islander (API)” as a bridge back to data collected under the old race standards. However, the NCHS Disclosure Review Board (DRB), consulting with DHIS analysts, determined that releasing data using an all-inclusive “Other Pacific Islander” category (which would include the Native Hawaiian, Samoan, Guamanian, and Other Pacific Islander groups) would pose a disclosure risk, especially when used in combination with other demographic and geographic information available on the file. For this reason, the decision was made to suppress the “Other Pacific Islander” category on all public use bridge variables. **This is important for data users to know because this change makes it impossible to bridge back to the old “Asian and Pacific Islander” category that existed in the 1998 and earlier NHIS surveys.** Data users who need this information for their analyses will have to contact the NCHS Research Data Center to obtain controlled access to non-released data.

### Creation and Editing of 2007 Race Variables

The variables RACRECI3 and MRACRPI2 correspond to the old OMB guidelines for collecting racial and ethnic data (see the Variable Layout Report for further descriptions of these variables). They were created in the same fashion as their previous NHIS counterparts (National Center for Health Statistics, 1996), with two exceptions. First, since observed race is no longer collected in the NHIS (beginning in 1997), it was not used to help classify persons with “Unknown” race on the RACRECI3 recode. Second, the recodes “White/Non-White” and “Black/Non-Black” were not created because they are no longer used in the weighting and tabulation of NHIS data. As in the past, smaller subgroups have been collapsed for confidentiality reasons.

Since the NHIS is now required to collect racial and ethnic data under the new OMB guidelines, new variables have been created to allow users to tabulate NHIS data by race variables that correspond to the new OMB guidelines. These variables conform to the new OMB race standards; therefore they are created independently of the follow-up race question (see the section of this appendix on Race and Hispanic Origin Questions in the National Health Interview Survey). The variable RACERPI2 was created using an algorithm that first coded the five race mentions from the survey into the single and multiple race group combinations (shown in bold/italicized and regular font, respectively) included in Table 2, below. All of the multiple race categories in the table were then collapsed into a single “Multiple race” category, and along with 4 of the 5 OMB single race categories, the variable RACERP\_I was created. The full algorithm is provided below so that our data users can better understand how this variable is derived.

### **Algorithm used to Create Single and Multiple Race Groups**

This SAS algorithm takes into account the new OMB categories: White, Black, American Indian/Alaskan Native (AIAN), Asian, and Native Hawaiian and Other Pacific Islander (NHOPI). In the NHIS, data are collected in 16 race categories: White, Black, Indian (American), Alaska Native, Native Hawaiian, Guamanian, Samoan, Other Pacific Islander (a verbatim mention that is back-coded to this category), Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Other Asian (a verbatim mention that is back-coded to this category). These can all be collapsed back to the OMB categories in the following fashion: *White*, *Black*, *AIAN* (includes Indian (American) and Alaska Native), *Asian* (includes Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese and Other Asian), and *NHOPI* (includes Native Hawaiian, Guamanian, Samoan and Other Pacific Islander).

Step 1: In the NHIS there are 5 possible mentions of race, which, when edited and cleaned, will become 5 race variables called RACE1, RACE2, RACE3, RACE4 and RACE5.

Step 2: Create and initialize the following variables to 0:

```
RACEW=0;  
RACEB=0;  
RACEAIAN=0;  
RACEASIA=0;  
RACENHPI=0;
```

Step 3: Set non-mutually exclusive conditions for recoding the 5 race variables, and set each of the above categories to the number designated:

```
IF ((RACE1=1) or (RACE2=1) or (RACE3=1) or (RACE4=1) or RACE5=1))  
then RACEW=1;  
* This sets RACEW to 1 if there is any mention of the race “White” in any of the  
5 race variables;
```

IF ((RACE1=2) or (RACE2=2) or (RACE3=2) or (RACE4=2) or RACE5=2))  
then RACEB=2;

*\*This sets RACEB to 2 if there is any mention of the race “Black” in any of the 5 race variables;*

IF ((RACE1=3) or (RACE2=3) or (RACE3=3) or (RACE4=3) or RACE5=3))  
then RACEAIAN=4;

*\*This sets RACEAIAN to 4 if there is any mention of the race “AIAN” in any of the 5 race variables;*

IF ((RACE1=4) or (RACE2=4) or (RACE3=4) or (RACE4=4) or RACE5=4))  
then RACEASIA=8;

*\*This sets RACEASIA to 8 if there is any mention of the race “Asian” in any of the 5 race variables;*

IF ((RACE1=5) or (RACE2=5) or (RACE3=5) or (RACE4=5) or RACE5=5))  
then RACENHPI=16;

*\*This sets RACENHPI to 16 if there is any mention of the race “NHOPI (Native Hawaiian and Other Pacific Islander)” in any of the 5 race variables;*

Step 4: RACEFULL=SUM(OF RACEW RACEB RACEAIAN RACEASIA  
RACENHPI);

The variables RACEW, RACEB, RACEAIAN, RACEASIA, and RACENHPI, are thus assigned the numbers 1, 2, 4, 8, and 16, which add up to a series of unique numbers corresponding to specific combinations of races. The value of RACEFULL tells which races (RACEW through RACENHPI) combined to give that number. For example, if RACEFULL=3, then only the sum of the values for RACEW=1 and RACEB=2 could have produced the number 3. Therefore, anyone with the value RACEFULL=3 falls into the “White/Black” race category. If RACEFULL=1, then those persons fall into the “White” category. This scheme accurately allocates persons with multiple Asian, AIAN, and NHOPI mentions. The full listing of categories and the numbers to which they correspond are included in the following table:

**Appendix II, Table 2. Coding Scheme for SAS Algorithm**

Coding Scheme for OMB Race Category Data (including single and multiple race mentions)		
<b># of Category (reported in SAS frequency distribution of RACEFULL)</b>	<b>Sum of Codes (breakdown of RACEFULL= SUM (OF RACEW+RACEB+ RACEAIAN+RACEASIA+ RACENHPI+RACEOTHR))</b>	<b>Resulting Category (used in the PROC FORMAT statement to label the categories in SAS)</b>
1	1+0+0+0+0+0	<i>White</i>
2	0+2+0+0+0+0	<i>Black</i>
3	1+2+0+0+0+0	White/Black
4	0+0+4+0+0+0	<i>AIAN</i>
5	1+0+4+0+0+0	White/AIAN
6	0+2+4+0+0+0	Black/AIAN
7	1+2+4+0+0+0	White/Black/AIAN
8	0+0+0+8+0+0	<i>Asian</i>
9	1+0+0+8+0+0	White/Asian
10	0+2+0+8+0+0	Black/Asian
11	1+2+0+8+0+0	White/Black/Asian
12	0+0+4+8+0+0	AIAN/Asian
13	1+0+4+8+0+0	White/AIAN/Asian
14	0+2+4+8+0+0	Black/AIAN/Asian
15	1+2+4+8+0+0	White/Black/AIAN/Asian
16	0+0+0+0+16+0	<i>NHOPI</i>
17	1+0+0+0+16+0	White/NHOPI
18	0+2+0+0+16+0	Black/NHOPI
19	1+2+0+0+16+0	White/Black/NHOPI
20	0+0+4+0+16+0	AIAN/NHOPI
21	1+0+4+0+16+0	White/AIAN/NHOPI
22	0+2+4+0+16+0	Black/AIAN/NHOPI

Coding Scheme for OMB Race Category Data (including single and multiple race mentions)		
# of Category (reported in SAS frequency distribution of RACEFULL)	Sum of Codes (breakdown of RACEFULL= SUM (OF RACEW+RACEB+ RACEAIAN+RACEASIA+ RACENHPI+RACEOTHR))	Resulting Category (used in the PROC FORMAT statement to label the categories in SAS)
23	1+2+4+0+16+0	White/Black/AIAN/NHOPI
24	0+0+0+8+16+0	Asian/NHOPI
25	1+0+0+8+16+0	White/Asian/NHOPI
26	0+2+0+8+16+0	Black/Asian/NHOPI
27	1+2+0+8+16+0	White/Black/Asian/NHOPI
28	0+0+4+8+16+0	AIAN/Asian/NHOPI
29	1+0+4+8+16+0	White/AIAN/Asian/NHOPI
30	0+2+4+8+16+0	Black/AIAN/Asian/NHOPI
31	1+2+4+8+16+0	White/Black/AIAN/Asian/NHOPI

*Data users should be aware that the variable RACEFULL and others derived from it are not available on public use data files for confidentiality reasons.* The recode RACERPI2 is a recode based on RACEFULL. Analysts who wish to use more detailed race data in their analyses should contact the NCHS Research Data Center or visit their web page:

<http://www.cdc.gov/nchs/r&d/rdc.htm> .

### Further Information

Although the race variables included in the 2007 file have been edited and tested, analytic and methodological work with these variables continues. NCHS is also evaluating other recodes for possible public release at a later date. If these analyses should result in changes to the 2007 NHIS race data, information about this will be placed on the NHIS website:

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

Additionally, the NHIS has a web site devoted exclusively to the race and Hispanic origin data from the survey. This site includes additional details on the NHIS race and Hispanic origin data, including more information on editing and imputation of the data, and links to documentation, questionnaires and other resources. We invite our users to visit this site:

<http://www.cdc.gov/nchs/about/major/nhis/rhoi/rhoi.htm>.

## References

Office of Management and Budget. *Provisional Guidance on the Implementation of the 1997 Standards for Federal Data on Race and Ethnicity, December 15, 2000 (Statistical Policy section)*. <http://www.whitehouse.gov/omb/inforeg/statpolicy.html>

Office of Management and Budget. Interagency Committee for the Review of Standards for Data on Race and Ethnicity. *Draft Provisional Guidance on the Implementation of the 1997 Standards for Federal Data on Race and Ethnicity, February 15, 1999*.

National Center for Health Statistics. *Computer Processing Procedures Calendar Year 1995, Volume 1 of 2, National Health Interview Survey, 1995*. National Center for Health Statistics, Hyattsville, MD (Producer). National Technical Information Service, U.S. Department of Commerce, Springfield, VA. 22161 (Distributor). 1996.

Office of Management and Budget. Office of Management and Budget Circular No. A48. *Standards and Guidelines for Federal Statistics*. Section 7H, Exhibit F, May 12, 1977.

Office of Management and Budget. *Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity*. Federal Register 62(210):58782-58790. 1997.

U.S Bureau of the Census. *Census 2000 Modified Race Data [MR (31)-CO.txt]*. 2002.

## Appendix III

### Variance Estimation and Other Analytic Issues, NHIS 2007

#### Introduction

The data collected in the NHIS are obtained through a complex, multistage sample design that involves stratification, clustering, and oversampling of specific population subgroups. The final weights provided for analytic purposes have been adjusted in several ways to permit calculation of valid estimates for the civilian, noninstitutionalized population of the United States. As with any variance estimation methodology, the techniques presented here involve several simplifying assumptions about the design and weighting scheme applied to the data. This appendix provides guidelines for data users based on simplified concepts of the NHIS sample design structure so that users may compute reasonably accurate standard error estimates.

Several software packages are available for analyzing complex samples. The Web site *Summary of Survey Analysis Software*, currently located at:

<http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html> ,

provides references for and a comparison of different software alternatives for the analysis of complex data. Analysts at NCHS generally use the software package SUDAAN<sup>®</sup> (Research Triangle Institute 2004) to produce accurate standard error estimates. In this appendix, examples of SUDAAN computer code are provided for illustrative purposes. Examples also are provided for the Stata, SPSS, SAS, R, and VPLX software packages. However, the appropriate application of these procedures is the ultimate responsibility of data users, and the example command code is *not* "guaranteed." Both the computer command code and methods are subject to change without notification to the user. NCHS recommends that NHIS data be analyzed under the direction of or in consultation with a statistician who is cognizant of sampling methodologies and techniques for the analysis of complex survey data.

**CAUTION.** Users are reminded that the use of standard statistical procedures that are based on the assumption that data are generated via simple random sampling (SRS) generally will produce incorrect estimates of variances and standard errors when used to analyze data from the NHIS. The clustering protocols that are used in the multistage selection of the NHIS sample require other analytic procedures, as described below. Analysts who apply SRS techniques to NHIS data generally will produce standard error estimates that are, on average, too small, and are likely to produce results that are subject to excessive Type I error.

#### Conceptual NHIS design for 2007

A new sample design was implemented in 2006. The 2007 NHIS sample came from the second year of the sample design. This appendix provides a brief outline of the new NHIS sample design. The new sample design is very similar to the previous sample design, which was in place

from 1995 to 2005. However, in order to accommodate the reduced NHIS funding level, the new sample design reduced the size of NHIS by about 13% relative to the previous sample design.

To achieve sampling efficiency and to keep survey operations manageable, cost-effective, and timely, the NHIS survey planners used multistage sampling techniques to select the sample of dwelling units for the NHIS. These multistage methods partition the target universe into several nested levels of strata and clusters. The NHIS target universe is defined as all dwelling units in the U.S. that contain members of the civilian noninstitutionalized population (households and noninstitutional group quarters such as college dormitories). As the NHIS is conducted in a face-to-face interview format, a simple random sample of dwelling units would be too dispersed throughout the nation; as a result, the costs of interviewing a simple random sample of 35,000 dwelling units would be prohibitive. Also, specific population subgroups, such as black, Hispanic, or Asian persons, would not be sampled sufficiently under a simple random sample design. To achieve survey objectives subject to resource constraints, the NHIS uses methods of clustering, stratification, and oversampling of specific population subgroups.

First, the target universe was partitioned into primary sampling units (PSUs), which are single counties, groups of adjacent counties (or equivalent jurisdictions), or metropolitan areas. These PSUs vary in population size and number of jurisdictions. Cost-effective field operations and efficient sampling result in those PSUs with the largest populations (e.g., the New York City metropolitan area) being sampled with certainty, and the smaller universe PSUs being represented by a sample. These smaller PSUs are called non-self-representing (NSR) or non-certainty PSUs. The universe of NSR PSUs is stratified geographically, for example by state, using multiple criteria consistent with NHIS objectives. Once these strata were defined, a sample of PSUs was selected; within most NSR strata, two PSUs were selected without replacement with probability proportional to population size, and the self-representing (SR) PSUs were selected with certainty. Within a few NSR strata with smaller population sizes, only one PSU was drawn.

The U.S. Bureau of the Census partitioned each selected NSR or SR PSU into substrata of Census blocks or combined blocks based on the concentrations of black, Asian, and Hispanic persons. These race and ethnicity density substrata were defined according to the population concentrations from the 2000 Decennial Census. New housing within a PSU was included as its own substratum in order to produce the most current sample of households. Each PSU could be partitioned into up to 21 substrata of dwelling units. Large metropolitan SR PSUs tend to have many substrata, while the NSR PSUs tend to have only a few.

Sampling within the PSU substrata is complex and involves clustering dwelling units within each substratum. These clusters form a universe of Secondary Sampling Units (SSUs). A systematic sample of SSUs is selected to represent each substratum.

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 proportion

of the NHIS sample that is assigned to be "screened" varies across the 21 substrata. For selected dwelling units, the NHIS collects some information about all persons living in the unit, and additional information is obtained for randomly selected persons living in the unit. For example, one adult per family is randomly selected for interview with the sample adult questionnaire.

In the previous NHIS sample design, all adults in a family had the same chance of being selected as the sample adult. In the new NHIS sample design, any black, Asian, or Hispanic adult aged 65+ years is given twice the chance of being selected as the sample adult as any other adult in the family. This new procedure was implemented to increase the proportion of sample adults who are black, Asian, or Hispanic, and aged 65+ years.

The hierarchy of sampling allows the creation of household- and person-level base weights. Each base weight is the product of the inverses of the probability of selection at each sampling stage. Roughly speaking, the base weight is the number of population units a sampled unit represents. Under ideal sampling conditions, and if 100% response occurred, a base-weighted sample total will be an unbiased estimator for the true total in the target population. In practice, however, the base weights are adjusted for non-response, and ratio-adjusted to create final sampling weights. The final person-level weights are adjusted according to a quarterly poststratification by age/sex/race/ethnicity classes based on population estimates produced by the U.S. Bureau of the Census that also are used for Current Population Survey weight adjustments. Most other weights receive some form of ratio adjustment as well.

Internally, NCHS uses the design and weighting information to formulate appropriate variance estimates for NHIS statistics. While recognizing the need to provide accurate information, NCHS also must adhere to the Public Health Service Act (Section 308(d)), which forbids the disclosure of any information that may compromise the confidentiality promised to its survey respondents. Consequently, much of the NHIS design information cannot be publicly released, and other data are either suppressed or recoded to insure confidentiality. In order to satisfy this disclosure constraint, many of the original design strata, substrata, PSUs, and SSUs are masked for public release by applying techniques to cluster, collapse, mix, and partition the original design variables. Through this process the original NHIS design variables are transformed into public use design variables (i.e., STRAT\_P and PSU\_P). Data users who want access to internal NCHS data have the option of accessing data through the NCHS Research Data Center. For further information, refer to <http://www.cdc.gov/nchs/r&d/rdc.htm>.

### **Design Information Available on the NHIS Public Use Data**

The 2007 Household, Person, Sample Adult, and Sample Child public use files contain the design variables necessary for variance estimation; Table 1 provides a summary of the Person file variables. The stratum and PSU variable names are the same in the other files, but the weight variable has a different name.

### Appendix III, Table 1. Variables Used for Variance Estimation, 2007 NHIS Person File

Variable Name	Variable Label
STRAT_P	Stratum for variance estimation
PSU_P	PSU for variance estimation
WTFA	Weight - Final annual Person weight

As discussed above, in order to mask true geographical locations, the STRAT\_P and PSU\_P levels are pseudo-levels or simplified versions of the true NHIS sample design variables. Analysts are cautioned that these simplified design structures do not support geographical analyses below the Census Region level.

**CAUTION.** The STRAT\_P and PSU\_P values for 2007 are based on the new NHIS sample design, and have no connection with the STRATUM and PSU values for 2005 and earlier years. Refer to the final section of this appendix for variance estimation guidance for pooled analyses of adjacent years of the NHIS, including pooling 2006 and 2007 data with data for 2005 and earlier years.

#### Variance Estimation Method for Public Use Data

The method described below is applicable to the 2007 NHIS Household, Person, Sample Adult, Adult Complementary and Alternative Medicine, Sample Child, and Child Complementary and Alternative Medicine public use data files. See the note about a variance estimation issue for the Sample Child and the Child Complementary and Alternative Medicine public use data files at the end of this section.

The limited public release design information requires a mathematical simplification that the PSUs be treated as if they were sampled with replacement (WR). This public use method tends to provide slightly more conservative (larger) standard errors than the variance estimation method that is applied internally by analysts at NCHS. The public use method is robust when analyzing subsetted or subgroup data (see the section "Subsetted Data Analysis" below).

The simplified design structure can be specified with the following statements in SUDAAN for the Person file:

```
PROC <DESCRIPT, CROSSTAB, ...> ... DESIGN = WR ;  
  NEST STRAT_P PSU_P ;  
  WEIGHT WTFA ;
```

Note that SUDAAN requires that the input file be sorted by the variables listed on the NEST statement (i.e., STRAT\_P and PSU\_P). Design statements for other data files should use the appropriate weight variables found on these files.

Corresponding statements for other software packages are as follows:

**Stata svy:**

*SVYSET [PWEIGHT=WTFA],STRATA(STRAT\_P)PSU(PSU\_P)*

*SVY: MEAN* <name of variable to be analyzed for average>

Or

*SVY: PROPORTION* <name of variable to be analyzed for percentage/proportion>

**SPSS csdescriptives (for averages) or cstabulate (for percentages/proportions):**

One needs first to define a "plan file" with information about the weight and variance estimation, e.g.:

*CSPLAN ANALYSIS*

*/PLAN FILE="< file name >"*

*/PLANVARS ANALYSISWEIGHT=WTFA*

*/DESIGN STRATA=STRAT\_P CLUSTER=PSU\_P*

*/ESTIMATOR TYPE=WR.*

And then refer to the plan file when using csdescriptives or cstabulate, e.g.:

*CSDESCRIPTIVES*

*/PLAN FILE="< file name >"*

*/SUMMARY VARIABLES =<name of variable to be analyzed>*

*/MEAN.*

*CSTABULATE*

*/PLAN FILE="< file name >"*

*/TABLES VARIABLES =<name of variable to be analyzed>*

*/CELLS TABLEPCT.*

**SAS proc surveymeans (for averages) or surveyfreq (for percentages/proportions) :**

*PROC SURVEYMEANS;*

*STRATA STRAT\_P;*

*CLUSTER PSU\_P;*

*WEIGHT WTFA;*

*VAR <name of variable to be analyzed>;*

*RUN;*

*PROC SURVEYFREQ;*

*STRATA STRAT\_P;*

*CLUSTER PSU\_P;*

```
WEIGHT WTFA;  
TABLES <name of variable to be analyzed>;  
RUN;
```

**R (including the "survey" add-on package):**  
(note: R syntax is case-sensitive)

```
# load survey package  
require(survey)  
# create data frame with NHIS design information, using existing data frame of NHIS data  
nhissvy <- svydesign(id=~psu_p, strata=~strat_p,  
                  nest = TRUE,  
                  weights=~wtfa,  
                  data=< existing data frame name>)  
svymean(~<name of variable to be analyzed>, design=nhissvy)
```

note: svymean will produce proportions for "factor variables". Consult the R documentation (<http://cran.r-project.org/manuals.html>) for details.

### **VPLX:**

In the CREATE step, include the following statements:

```
STRATUM   STRAT_P  
CLUSTER   PSU_P  
WEIGHT    WTFA
```

Then specify the variable to be analyzed in the DISPLAY step:

```
LIST      MEAN(<name of variable to be analyzed>)
```

VPLX can produce percentages by including a CAT statement in the CREATE step. Consult the VPLX documentation (<http://www.census.gov/sdms/www/vdoc.html>) for details.

**CAUTION.** A rule of thumb to calculate the number of degrees of freedom to associate with a standard error is the quantity *number of PSUs - number of strata*. Typically, this rule is applied to a design with two PSUs per stratum and when the variance components by stratum are roughly the same magnitude. The applicability of this rule depends upon the variable of interest and its interaction with the design structure (for additional information, see Chapter 5 of Korn and Graubard 1999). The number of degrees of freedom is used to determine the *t*-statistic, its associated percentage points, p-values, standard error, and confidence intervals. As the number of degrees of freedom becomes large, the distribution of the *t*-statistic approaches the standard normal distribution. For example, with 120 degrees of freedom, the 97.5 percentage point of the  $t_{120}$  distribution is 1.980, while the 97.5 percentage point of the standard normal distribution is 1.960. If a variable of interest is distributed across most of the NHIS PSUs, a normal distribution

assumption may be adequate for analysis since the number of degrees of freedom would be large. The user should consult a mathematical statistician for further discussion.

**CAUTION.** The 2007 Sample Child public use files have two records from a singleton PSU, i.e., only one PSU was drawn---see Appendix III Conceptual NHIS design for 2007 for discussion of primary sampling units. These same two respondents are also in the Child Complementary and Alternative Medicine supplement public use file. Contemporary software for complex survey design variance estimation such as SAS, SPSS and SUDAAN may estimate variances that are marginally different from estimates obtained from including dummy records for the two singleton PSU respondents. Users of some software (e.g., Stata 9) may get missing values for variance estimates. Users of SUDAAN should specify the MISSUNIT option in the NEST statement to prevent errors that stop processing of data. A mini-file with two dummy records for the singleton PSUs, assigning a record weight equal to 1 to each, will be released in August 2008 along with instructions and documentation describing the potential impact of singleton PSUs on variance estimation. Users encountering software problems arising from singleton PSUs can append the dummy record file to the public use files.

### Subsetting Data Analysis

Frequently, studies using NHIS data are restricted to specific population subgroups, e.g., persons aged 65 and older. Some users delete all records outside of the domain of interest (e.g., persons aged less than 65 years) in order to work with smaller data files and run computer jobs more quickly. This procedure of keeping only selected records (and list-wise deleting other records) is called subsetting the data. With a subsetting dataset that is appropriately weighted, correct point estimates (e.g., estimates of population subgroup means) can be produced. **However, in general, software packages that correctly analyze complex survey data cannot compute accurate standard errors for subsetting data.** When complex survey data are subsetting, oftentimes the sample design structure is compromised because the complete design information is not available; subsetting data deletes important design information needed for variance estimation. Note that SUDAAN has a SUBPOPN option that allows the targeting of a subpopulation while using the full (unsubsetting) data file containing the design information for the entire sample. (See a SUDAAN manual for more information.) **NCHS recommends that subpopulation analyses be carried out using the full data file and the SUBPOPN option in SUDAAN, or an equivalent procedure (see below) with another complex design variance estimation software package.**

**Strategy 1 (recommended)** Use the SUBPOPN statement with the method described above for the full Person file dataset:

```
PROC ...    DESIGN = WR ;
NEST STRAT_P PSU_P ;
WEIGHT     WTFA ;
SUBGROUP   (variable names);
LEVELS ... ;
SUBPOPN    RACRECI3=2 & SEX=2 / NAME="Analysis of African American
women;"
```

Using the full dataset with the SUBPOPN statement in this example would constrain this analysis to African American women only (RACRECI3 = 2 for black and SEX = 2 for female). Use of the SUBPOPN statement is equivalent to subsetting the dataset, except that any resulting variance estimates are based on the full design structure for the complete dataset.

**Strategy 2 (not recommended, except when Strategy 1 is infeasible)** Use the MISSUNIT option on the NEST statement with the method described above for subsetted data:

```
NEST        STRAT_P PSU_P / MISSUNIT ;
```

In a WR design, when some PSUs are removed from the database through the listwise deletion of records outside the population of interest, leaving only one PSU in one or more strata, the MISSUNIT option in SUDAAN "fixes" the estimation to avoid errors due to the presence of strata with only one PSU. In the special case of a WR design with exactly two PSUs per stratum, using the MISSUNIT option with subsetted data gives the same variance estimate as using Strategy 1. However, except for this special case, there is no guarantee that the variance estimates obtained by this method are equivalent to those obtained using Strategy 1. Other calculations, such as those for design effects, degrees of freedom, standardization, etc., may need to be carried out differently. Users are responsible for verifying the correctness of their results based on subsetted data.

Implementing Strategy 1 in other software packages can be accomplished as follows:

**Stata svy:**

Add SUBPOP to the SVY statement, e.g.:

```
SVY,SUBPOP( RACRECI3==2 & SEX==2 ): MEAN <name of variable to be analyzed>
```

**SPSS cdescriptives or cstabulate:**

One must first define an indicator variable, e.g.:

```
DO IF (RACRECI3 EQ 2 AND SEX EQ 2).  
  COMPUTE SUBGRP=1.  
ELSE.  
  COMPUTE SUBGRP=0.  
END IF.
```

And then refer to the indicator variable in csdescriptives or cstabulate, e.g.:

```
CSDSCRIPTIVES (or CSTABULATE)  
/SUBPOP TABLE=SUBGRP
```

It is **very important** that the indicator variable be defined for all data records, otherwise an invalid result can occur.

### **SAS proc surveymeans or surveyfreq:**

One must first define an indicator variable, e.g.:

```
IF RACRECI3=2 & SEX=2   THEN SUBGRP=1;  
                        ELSE SUBGRP=0;
```

And then refer to the indicator variable in proc surveymeans using the DOMAIN statement, e.g.:

```
PROC SURVEYMEANS;  
DOMAIN SUBGRP;
```

Proc surveyfreq does not have a DOMAIN statement. Instead, include the indicator variable in the TABLES specification:

```
PROC SURVEYFREQ;  
TABLES SUBGRP*<name of variable to be analyzed>;
```

As with SPSS, it is **very important** that the indicator variable is defined for all data records; otherwise an invalid result can occur.

### **R (including the "survey" add-on package):**

After applying the svydesign function to a data frame that contains the entire NHIS sample file being analyzed, specify the criteria that define the subgroup of interest in the subset function and apply the function to the R "object" created by the svydesign function to create a new R object. Note that R is very "feisty" when testing for equality, hence the syntax that follows specifies the subgroup of interest without using an equality test.

```
# subset for racreci3=2 & sex=2 without using equal signs
subgrp <- subset(nhissvy,racreci3>1 & racreci3<3 & sex>1)
svymean(~<name of variable to be analyzed>,design=subgr
```

### **VPLX:**

In the CREATE step, define one or more CLASS variables that can be used to specify the criteria that define the subgroup of interest.

```
COPY RACRECI3 INTO RACECAT
COPY SEX INTO SEXCAT
CLASS RACECAT (1/2/3-HIGH)
CLASS SEXCAT (1/2)
```

The second category of RACECAT, crossed with the second category of SEXCAT, defines the subgroup of interest.

Then, specify the variable to be analyzed in the DISPLAY step, and specify the subgroup of interest as well:

```
LIST MEAN(<name of variable to be analyzed>)/CLASS RACECAT(2)*SEXCAT(2)
```

Note that the specification of RACECAT(2) and SEXCAT(2) is to the second category of each variable, which happens to be the value "2" in both cases in this example. Specification of RACECAT(3) would include all values of RACRECI3 of 3 and higher ("3-HIGH").

### **Variance Estimation for Pooled Analyses of Adjacent Years of the NHIS**

Adjacent years of NHIS data sometimes are combined for a pooled analysis, e.g., 2005 and 2006, or 2002-2004. A pooled analysis might be done, for example, to increase the sample size for some small population. An estimate from a pooled analysis can be interpreted to be an estimate for the midpoint of or the "average" over the time interval of the pooled data.

See Appendix VI, "Combining Years of Data" section, for an example SAS program that combines 2004 and 2005 NHIS data, and an example program that forms a combined 2004-2007 NHIS dataset.

The sampling weights for pooled data should be adjusted; otherwise, estimates of totals will be too high. For example, the estimated total U.S. civilian noninstitutionalized population from two years of pooled data, using unadjusted weights, would be about twice as large as it should be. A simple, valid weight adjustment procedure that NCHS recommends is to divide each sample weight in the pooled dataset by the number of years that are being pooled; e.g., divide by 2 when two years of data are combined, divide by 3 when three years of data are combined, etc. A sophisticated user may want to consider an alternative weight adjustment method that would minimize the variance of a particular estimate; however, in general, if the sample sizes are similar

in the data years being combined, the simple procedure and the sophisticated alternative would give a similar adjustment.

Variance estimation for pooled analyses falls into one or more of the following three classifications:

#1. The years being pooled fall within the same sample design period with the same public use design variables, and no changes were made to the design variables within the years being pooled.

#2. The years being pooled fall into different sample design periods (e.g., design periods 1963-1972, 1973-1984, 1985-1994, 1995-2005, 2006 and later years).

#3. The years being pooled fall within the same sample design period, and there were changes to the public use design variables (e.g., from 1995-1996 to 1997-2005).

For #1, the sample has been drawn from the same geographic areas (same sample design), and the definitions of the variables used for public use variance estimation have not changed within the time period being analyzed. A valid method for variance estimation is to treat the pooled data like one year of data with a very large sample size. It is not correct to treat the different data years as being statistically independent, because the samples for the different years were drawn from the same geographic areas (i.e., same PSUs, nearby SSUs). Treating different data years as being statistically independent generally will lead to standard error estimates that are too small, and standard error estimates of contrasts (differences) between years would tend to be too large if the yearly estimates are positively correlated.

For #2, the different sample design periods should be treated as statistically independent. If there are multiple years of data being used for one or both design periods, each group should be treated in a similar manner as described in #1, assuming that the design variables within each group were unchanged. For example, if 1992-1995 NHIS data were pooled, the #1 procedure applies for the 1992-1994 data, and that aggregate is treated as being statistically independent from the 1995 data.

Note that it may be necessary to create new design variables to carry out this type of analysis. For example, consider an analysis of 1992-1995 NHIS data. The design variables have different names in the two sample design periods, and the stratum identifiers have different lengths. Referring to the first method described in "Variance Estimation for Person Data Using SUDAAN and the National Health Interview Survey (NHIS) Public-Use Person Data Files, 1987-94", currently available online at <http://www.cdc.gov/nchs/about/major/nhis/sudaan.htm>, the (Method 1) design variables for the 1992-1994 data are CSTRATUM (stratum), CPSU (PSU), and WTF (weight), while they are STRATUM, PSU, and WFTA, respectively, for the 1995 data. Suppose the names of the new design variables are NSTRATUM (stratum), NPSU (PSU), and NWT (weight). One method to create values for NSTRATUM that are of consistent length and take account of the different sample design periods is to do the following: for the 1992-1994 data, where the CSTRATUM values are 1, 2, ..., 62, first change these to 001, 002, ..., 062 (consistent length with STRATUM), and then do something to make them distinct from the STRATUM

values, such as put a "1" in front: 1001, 1002, ..., 1062. For the 1995 data, where the STRATUM values are 1, 2, ..., 339, first change these to 001, 002, ..., 339, and then do something to make them distinct from the CSTRATUM values, such as put a "2" in front: 2001, 2002, ..., 2339. NPSU can be set equal to CPSU for the 1992-1994 data, and equal to PSU for the 1995 data, as both CPSU and PSU are of length one. NWT can be set equal to WTF/4 for the 1992-1994 data, and to WFTA/4 for the 1995 data.

For #3, no entirely satisfactory approach is available. Grouping of years should be done over the periods where the same public use design variables are present (i.e., like #1). Then, for combining across years where there were changes to the public use design variables, the only option is to carry out an analysis as if the data years were statistically independent. For example, if 1995-1999 NHIS data were pooled, the #1 procedure applies for 1995-1996, and 1997-1999; then, the only alternative is to treat these two groups as statistically independent. The resulting standard error estimates may be too small, and standard error estimates of contrasts between years might be too large if the yearly estimates are positively correlated.

### References

- Cochran, W.G. *Sampling techniques* (3rd edition), John Wiley & Sons. 1977.
- Korn, E.L., and Graubard, B.I. *Analysis of Health Surveys*, John Wiley & Sons. 1999.
- National Center for Health Statistics. *National Health Interview Survey: Research for the 1995-2004 redesign*, Vital and Health Statistics, Series 2, No. 126. 1999.
- National Center for Health Statistics. *Design and Estimation for the National Health Interview Survey, 1995-2004*, Vital and Health Statistics, Series 2, No. 130. 2000.
- Research Triangle Institute. *SUDAAN Language Manual; Release 9.0*, Research Triangle Institute, Research Triangle Park, NC. 2004.

## **Appendix IV**

### **A Preliminary Evaluation and Recommendations for use of the Mental Health Indicator (MHI) in the NHIS for Children Aged 2 to 3 years**

This is based on a report by Thomas M. Achenbach, Ph.D., which was submitted to the Division of Health Interview Statistics on May 10, 1999.

#### **Introduction**

The NHIS mental health indicators MHIBOY2 and MHIGRL2 are located in the Child Health Status (CHS) section of the survey, and are based on items from the Child Behavior Checklist (CBCL) that were identified by Dr. Thomas Achenbach as providing the best discrimination between demographically similar children referred for mental health services versus nonreferred (Achenbach and Edelbrock, 1983). To take account of gender and age differences in the discriminative power of particular items, the items were selected separately for each gender and age group. From the original ten items identified in Dr. Achenbach's 1995 analyses, the NHIS elected to include only 4 items (per gender). These include whether male sample children (aged 2-3 years) had been uncooperative, had trouble sleeping, had speech problems, or had been unhappy or depressed in the past 2 months, and whether female sample children (aged 2-3 years) had temper tantrums, had speech problems, had been nervous or high-strung, or had been unhappy or depressed in the past 2 months. Response categories included "Not true," "Sometimes true," or "Often true" (as well as "Refused" and "Don't know"). These items are also located in the CHS section (see CHS.321\_01-04.000 and CHS.361\_01-04.000).

It is essential to note that such a small set of items cannot be used to evaluate individual children for clinical or other purposes. Even for use as a mental health indicator in large surveys such as the NHIS, very small sets of items can serve only as approximate indicators of needs for mental health services. Multiple items tapping each of several specific areas of functioning would be needed to identify specific disorders, such as Attention Deficit Hyperactivity Disorder (ADHD), Depression, Conduct Disorder, and Somatization Disorder. (Note: The items for children ages 4 to 17 were replaced in the 2001 NHIS with a different instrument, the Strengths and Difficulties Questionnaire (SDQ). The SDQ is described in Appendix V.)

It should also be noted that different cut points on the distributions of item scores may be needed for different purposes. For example, a very low cut point may be useful if the goal is to identify every possible case for which mental health services might be considered. However, very low cut points result in relatively high false positive rates, i.e., the inclusion of substantial numbers of healthy individuals among those identified as potentially needing services. Conversely, higher cut points may yield greater overall accuracy in classifying potential cases versus noncases, but at the cost of missing more cases potentially needing services.

## **Data Analyses**

Dr. Achenbach specified and reviewed data analyses that were done at NCHS. These included tabulations of specific responses to each behavioral/emotional problem item; tabulations of relations between total problem scores and classification of children as deviant versus nondeviant on the basis of external criteria (e.g., parents ever being told by health professionals that their child had ADHD, mental retardation, other developmental delay, autism, down syndrome, or a learning disability; parents having talked to mental health professionals about their child in the preceding 12 months; or parents needing mental health services for their child but being unable to afford it); and Relative Operating Characteristic (ROC) analyses of cut points on the total problem scores. Because each behavioral/emotional problem item was scored “0” (not true of the child), “1” (somewhat or sometimes true), or “2” (very true or often true), total scores across the 4 items for each gender/age group could range from “0” to “8.” Dr. Achenbach examined the results and recommended changes and additions to the analyses.

Based on the analyses to date, Dr. Achenbach makes the following recommendations for boys and girls ages 2-3. Total scores on the 8 problem items are useful for quantitative analyses in relation to other variables. However, categorical mental health indicators should not be derived from specific cut points on the total scores for the 4 behavioral/emotional problem items on the basis of NHIS data for ages 2-3 for the following reasons:

The total number of children classified as deviant according to external criteria (e.g., parents being told their child had ADHD; talking to mental health professionals about their child) was too small to provide a sound basis for establishing cut points;

Many disorders relevant to defining criterion groups (e.g., ADHD) are not identified as early as age 2-3;

The rates of referral for mental health services and other possible indicators of deviance are much lower at ages 2-3 than at older ages.

## Appendix V

### The Short Strengths and Difficulties Questionnaire (SDQ)

In the NHIS questions CMB.020\_01.000 to CMB.030\_00.000 make up a brief version of the SDQ. The questions are derived from the parent version of the long Strengths and Difficulties Questionnaire Extended (SDQ), developed and copyrighted by Dr. Robert Goodman, Institute of Psychiatry, London, England. Questions from the SDQ are used in the NHIS with Dr. Goodman's permission. The short SDQ, constructed to save time and space in the questionnaire, was added for children aged 4-17 years as a part of a collaborative agreement between NCHS and the National Institute of Mental Health (NIMH) of the National Institutes of Health (NIH). The long SDQ consists of 25 scale items. Detailed information on the SDQ can be found in Appendix 5 of the Dataset Documentation for the 2004 NHIS and on the SDQ web site at <http://www.sdqinfo.com>.

The items in the short SDQ correlate to the subscales in the long SDQ as follows:

CMB.020\_01.000 *Generally obedient*, correlates 0.69 with the long SDQ conduct score.  
CMB.020\_02.000 *Many worries...*, correlates 0.71 with the long SDQ emotion score.  
CMB.020\_03.000 *Often unhappy...*, correlates 0.64 with the long SDQ emotion score.  
CMB.020\_04.000 *Gets along better...*, correlates 0.69 with the long SDQ peer problems score.  
CMB.020\_05.000 *Sees tasks through*, correlates 0.72 with the long SDQ hyperactivity-inattention score.

In order to score the short SDQ the response for each item in CMB.020 is assigned a value from 0 – 2 based on the scale below, then all values are summed to produce a total score. A total score from 1 to 5 correlates 0.84 with the long SDQ total difficulties score.

#### Scoring of the Short SDQ

Response	<u>Not true</u>	<u>Somewhat</u>	<u>Definitely</u>
Value	<u>0</u>	<u>1</u>	<u>2</u>

CMB.030 is taken from a set of SDQ extended or impact questions which measure the impact of the child's difficulty on various aspects of his/her life. CMB.030 correlates 0.62 with the SDQ impact score from the extended SDQ questions. (See Appendix V of the Dataset Documentation for the 2004 NHIS and/or the SDQ Web site at <http://www.sdqinfo.com>)

### **Additional References on the SDQ and/or its use in the NHIS**

Bourdon KH, Goodman R, Rae D, Simpson G, Koretz D. The Strengths and Difficulties Questionnaire: U. S. Normative Data and Psychometric Properties. *Journal of American Academy of Child and Adolescent Psychiatry*, 44(6): 557-564. 2005.

Goodman, R. The Strengths and Difficulties Questionnaire: A Research Note. *Journal of Child Psychology and Psychiatry*. 38: 581-586. 1997.

Goodman, R. The Strengths and Difficulties Questionnaire as a Guide to Child Psychiatric Caseness and Consequent Burden. *Journal of Child Psychology and Psychiatry*. 40 (5): 791-799. 1999.

Goodman, R. and Scott S. Comparing the Strengths and Difficulties Questionnaire and the Child Behavior Checklist: Is small beautiful? *Journal of Abnormal Child Psychology*, 27(1): 7-24. 1999.

Simpson GA, Bloom B, Cohen R, Blumberg S, Bourdon K. U. S. Children with Emotional and Behavioral Difficulties: Data from the 2001, 2002, and 2003 National Health Interview Surveys. *Advance data from vital and health statistics no. 360*. Hyattsville, MD. National Center for Health Statistics. 2005.

## Appendix VI

### Transition to the 2000 Census-based Weights

For the NHIS sample design that was implemented in 1995, between 1995 and 2002, the weights for the NHIS data were derived from 1990 census-based postcensal population estimates. Beginning with the 2003 data, the NHIS has made the transition to weights derived from the 2000 census-based population estimates. The new population estimates no longer contain any adjustment for under-enumeration of the population. The NHIS sample weights were calibrated to 2000 census-based totals for sex, age, and race/ethnicity of the U.S. civilian noninstitutionalized population.

During the preparation of the 2003 Summary Health Statistics reports, the impact of this transition was assessed for the 2002 NHIS by comparing estimates for selected health characteristics using the 1990 census-based weights with those using the 2000 census-based weights. The effect of new population controls on survey estimates differed by type of health characteristic. The person health estimates and sample adult health estimates were more affected than sample child estimates. The percent of health estimates expressed as percentages and rates with significant differences were 0.27% for person estimates, 0.27% for sample adult estimates, and 0.0% percent for sample child estimates. The percent of health estimates expressed as frequencies with significant differences were 13% for person estimates, 16% for sample adult estimates, and 1% for sample child estimates (Lynch and Parsons, 2004).

The impact of this transition was also assessed for the Early Release estimates from the 2000-2002 NHIS by comparing estimates using the 1990 census-based weights with those using the 2000 census-based weights. The changes for all selected measures are at most 1 percentage point. Results of these findings are presented in tables II and III at:  
[http://www.cdc.gov/nchs/data/nhis/earlyrelease/200409\\_app.pdf](http://www.cdc.gov/nchs/data/nhis/earlyrelease/200409_app.pdf).

### References

Lynch, C. and Parsons, V. The Impact of 2000 Census Based Population Controls on Health Estimates in the National Health Interview Survey, *2004 Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM]*, Alexandria, VA: American Statistical Association. 2004.

## Appendix VII Merging Data Files and Combining Years of Data in the NHIS

NHIS data files can be merged within years as well as combined across years. The purpose of merging data *within* a particular data year is to incorporate variables from different data files when respondents are common to both files, thereby increasing the number of variables available for analysis for a given individual. In contrast, the purpose behind combining NHIS data files *across* survey years is to combine respondents from different data years while retaining variables common to both files, thereby increasing the number of respondents (as long as the same variables are found in both files) and the precision of estimates.

### Merging Data Files

It is important to note that for the 2005 data year and beyond, some frequently used variables are repeated on various data files; therefore, merging of files may be required less often than for the 2004 data year files. However, each data file contains household, family, and person record identifiers that make merging the data files possible, if needed. Once the data files are sorted by record identifiers common to each file, merging is straightforward. Below is an example of a SAS program that will merge data files within an NHIS data year. Using the household, family, and person record identifiers (HHX, FMX and FPX, respectively), this program merges data from the 2005 Household, Family, Person, and Sample Child data files. **Variable names may change from one year to another. Users are advised to check variable names and update computer programs when changes occur.**

```
/* Merge the 2005 Household file and the 2005 Family file. */
```

```
/* Create a Household dataset with selected variables and sorted by HHX.*/
```

```
DATA HH (KEEP=HHX REGION); /* HH is a SAS dataset; the KEEP statement retains only the  
listed variables for processing. */
```

```
SET NHIS2005.HOUSEHLD; /*The SET statement reads data from the 2005 Household file. */
```

```
PROC SORT DATA=HH; /* Sort by HHX, the household identifier. */
```

```
BY HHX;
```

```
RUN;
```

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

```
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 NHIS2005.FAMILYXX; /*The SET statement reads data from the 2005 Family file. */
```

```
PROC SORT DATA=FM; /* Sort by HHX, the household identifier. */
```

```
BY HHX;
```

```
RUN;
```

```
DATA HHFM; /* New combined dataset called HHFM */
```

```
MERGE FM (IN=FROMFM) HH ; /* Merge the newly created FM and HH files, using an IN  
statement.*/
```

```
BY HHX;
```

```

IF FROMFM = 1; /* The combined dataset HHFM will contain only those records that are in the
Family file; the Household file's REGION variable will be appended to these records. */
PROC SORT DATA=HHFM; /* 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), thereby eliminating household file records corresponding to nonresponse cases (no family/person records are available for nonresponse cases). For more information on IN statements in SAS, consult Delwiche and Slaughter (1998).

```

/* Merge the 2005 Person file and the combined 2005 Family/Household file. */

```

```

/* Create a Person file with selected variables. */

```

```

DATA PR (KEEP=HHX FMX FPX SEX AGE_P WTFA STRATUM PSU); /* PR is a SAS
dataset; the KEEP statement retains only the listed variables for processing. */
SET NHIS2005.PERSONSX; /*The SET statement reads data from the 2005 Person file. */
PROC SORT DATA=PR; /* Sort by HHX and FMX, the household and family identifiers. */
BY HHX FMX;
RUN;

```

```

DATA PRHHFM; /* Combined Person, Family, and Household dataset called PRHHFM*/
MERGE PR HHFM (DROP=WTFA_FAM); /* Merge the newly created PR file and HHFM, the
combined Family/Household file, by the identifiers common to both files. At this point, users may
drop the Family file weight and retain only the Person file weight for person-level analyses.*/
BY HHX FMX;
PROC SORT DATA=PRHHFM; /* Sort by HHX, FMX, and FPX, the household, family, and
person identifiers. */
BY HHX FMX FPX;
RUN;

```

The above code will create a person-level file, copying the family/household information to each matching person record.

```

/* Merge the 2005 Sample Child file and the combined 2005 Person/Family/Household file. */

```

```

/* Create a Sample Child file with selected variables. */

```

```

DATA CH (KEEP=FPX HHX FMX CASHMEV PROBRX WTFA_SC); /* CH is a SAS
dataset; the KEEP statement retains only the listed variables for processing. */
SET NHIS2005.SAMCHILD; /*The SET statement reads data from the 2005 Sample Child file. */
PROC SORT DATA=CH; /* Sort by HHX, FMX, and FPX, the household, family, and person
identifiers. */
BY HHX FMX FPX;
RUN;

```

**DATA CHPRHHFM;** /\* Combined Sample Child, Person, Family, and Household dataset called CHPRHHFM\*/

**MERGE PRHHFM CH;** /\* Merge CH, the newly created Sample Child file, and PRHHFM, the combined Person/Family/Household file, by the identifiers common to both files.

**BY HHX FMX FPX;**

**RUN;**

In the code above, no IN statement was used in the MERGE statement, so the resulting file will have records for all persons, sample child or not. The sample child data items will be missing for persons who do not have a matching sample child record.

### Combining Years of Data

#### Important Note

**Variable names may change from one year to another. Users are advised to check variable names and where names differ, make certain it is appropriate to combine years of data for a given variable.**

As previously mentioned, the purpose of combining or concatenating years of data (in SAS terminology) is to increase the number of observations or respondents for the same number of variables, and thus increase the precision of estimates. It is possible to combine data from successive years of the National Health Interview Survey (NHIS) when the questions remain essentially the same over the years being combined.

Combining datasets from more than one year joins them one after the other (concatenates), as opposed to merging datasets. Analysts wishing to do both – merge data from multiple files within years and combine years of data – will need to first merge the data within each single year and then concatenate the files for the selected years of data (see the preceding section on Merging Data Files).

Weights will normally need to be adjusted when combining data years. For example, if two years of NHIS data are combined, the sum of the weights will be about twice the size of the civilian noninstitutionalized population of the United States. To achieve annualized results when two years of NHIS data are combined, one method for weight adjustment is to divide each weight by two before analyzing the data.

If data from the period 1997-2005 are combined or the 2006-2007 data are combined, the combined data are treated like a single year of data with a larger sample size for the purpose of variance estimation. If data from any year before 1997 are combined with data from 1997 and beyond, or data from 2005 or before are combined with data from 2006 and beyond, variance

estimation is more complicated. Refer to Appendix III for more information about variance estimation methods when combining datasets from more than one year.

The following is an example of a SAS program that will combine data files across NHIS data years. The program is written to concatenate the data from the Person files of the 2004 NHIS and the 2005 NHIS. This same program can be used to combine the 2006 and 2007 NHIS Person files after minor modifications (e.g., change "2004" and "04" to "2006" and "06", change "2005" and "05" to "2007" and "07", and change STRATUM PSU to STRAT\_P PSU\_P).

### Important Note

**The person identifier was called PX in the 2003 (and earlier) NHIS and FPX in the 2004 (and later) NHIS; users may find it necessary to create an FPX variable in the 2003 and earlier datasets (or, alternatively, a PX variable in the 2004 and later datasets) in order to make the data compatible for analyses.**

```
/*Combine data files from 2 different years. */
```

```
DATA PER_04; /* Create SAS dataset PER_04.*/
```

```
SET NHIS2004.PERSONSX /* The SET statement reads data from an existing SAS dataset, e.g.,  
the 2004 Person file */ (KEEP=HHX FMX FPX AGE_P SEX WTFA STRATUM PSU); /* The  
KEEP statement retains only the listed variables for processing. */
```

```
RUN;
```

```
PROC SORT DATA=PER_04; /* Sort SAS dataset PER_04. */
```

```
BY HHX FMX FPX;
```

```
RUN;
```

```
DATA PER_05; /* Create SAS dataset PER_05.*/
```

```
SET NHIS2005.PERSONSX /* The SET statement reads data from an existing SAS dataset, e.g.,  
the 2005 Person file */ (KEEP=HHX FMX FPX AGE_P SEX WTFA STRATUM PSU); /* The  
KEEP statement retains only the listed variables for processing. */
```

```
RUN;
```

```
PROC SORT DATA=PER_05; /* Sort SAS dataset PER_05. */
```

```
BY HHX FMX FPX;
```

```
RUN;
```

```
DATA COMBO; /* New, combined SAS dataset */
```

```
SET PER_04 PER_05; /* Concatenate selected variables from 2004 and 2005 datasets. */
```

**WTFA\_2YR=WTFA/2; /\*Create a new weight by dividing the existing Person file weight (WTFA) by 2, the number of Person data files combined to create the data file called COMBO.\*/  
RUN;**

Now, suppose there exists a dataset "COMB0405" with the combined 2004 and 2005 Person files, and there exists a dataset "COMB0607" with the combined 2006 and 2007 Person files. As part of creating a dataset named "COMB0407" containing the combined 2004-2007 Person files, two issues need to be addressed:

1. Adjustment of weights
2. Formation of new variance estimation variables, because this combination goes across sample design periods

The weights in COMB0405 and COMB0607 should be divided by 2, so that the original weights have been divided by 4 (four years of data being combined). To avoid the possibility of errors, NCHS recommends that new names be used for the new variance estimation variables, e.g., NSTRATUM (stratum), NPSU (PSU). The PSU and PSU\_P values from COMB0405 and COMB0607 can be copied directly to NPSU. The NSTRATUM values need to be created in such a way to assure the values are distinct between 2004-2005 and 2006-2007. As STRATUM ranges from 1 to 339 and STRAT\_P ranges from 1 to 300, an appropriate method for creating the NSTRATUM values would be to add 1000 to the STRATUM values and 2000 to the STRAT\_P values.

*/\*Combine 2004-2005 data file with 2006-2007 data file \*/*

**DATA COMB0405;  
SET COMB0405;  
DROP STRATUM PSU;  
NSTRATUM=STRATUM+1000;  
NPSU=PSU;  
RUN;**

**DATA COMB0607;  
SET COMB0607;  
DROP STRAT\_P PSU\_P;  
NSTRATUM=STRAT\_P+2000;  
NPSU=PSU\_P;  
RUN;**

**DATA COMB0407;  
SET COMB0405 COMB0607;  
DROP WTFA\_2YR;  
WTFA\_4YR=WTFA\_2YR/2;  
RUN;**

## References

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

## Appendix VIII

### Core Changes/Additions/Deletions in 2007

A small number of changes were introduced to the Core sections of the 2007 NHIS, resulting in new, changed, or deleted variables (relative to 2006). A brief summary is provided below. Users are *strongly* encouraged to check the notes in the Variable Layout Report available on the NHIS Web site, <http://www.cdc.gov/nchs/nhis.htm>, as well as the relevant sections in this document, for more information.

#### Family File

<u>Variable Name</u>	<u>Brief Variable Description</u>	<u>Brief Description of Change</u>
INCGRP	Grouped family income, 1997-2006	Deleted variable
INCGRP2	Grouped family income, 2007	New variable
INCGRP3	Grouped family income, 2007 bridge variable	New variable
RAT_CAT	Poverty ratio, 1997-2006	Deleted variable
RAT_CAT2	Poverty ratio, 2007	New variable
RAT_CAT3	Poverty ratio, 2007, added detail	New variable
HOWLNGWK	Weeks w/o telephone service	Deleted variable
INSIDE	Working phone inside home	Deleted variable
NOSERV	No telephone service past 12 m.	Deleted variable
RNOSERV	No telephone service past 12 m.	Deleted variable
RNOSERVN	No telephone service past 12 m.	New wording; moved from REC to COV section
RH1LNGDY	Days w/o telephone service	Deleted variable
RH2LNGDY	Days w/o telephone service due to weather/natural disasters	Deleted variable
TELDIRC	Number listed in directory	Deleted variable
TELH_FLG	Recode home telephone number flag	Deleted variable
TELN_FLG	Recode telephone number here flag	New variable
VCELNUM	Is this your cell phone number	Deleted variable
RHLNGDY	Recode for time w/o telephone service	Change in question wording; moved from REC to COV section
TELCEL	Working cell phone in family	Deleted variable
TELCELN	Working cell phone in family	Moved from REC to COV section
WRKCEL	Number of working cell phones	Deleted variable

WRKCELN	Number of working cell phones	Moved from REC to COV section
CURWRK	Working phone inside home	Deleted variable
CURWRKN	Working phone inside home	Moved from REC to COV section
CELLOUT	Working cell during landline outage	New variable in COV section
PHONEUSE	Received calls cell/landline/both	New variable in COV section

### Family Core

#### Health Insurance Section (FHI)

<u>Variable Name</u>	<u>Brief Variable Description</u>	<u>Brief Description of Change</u>
HDHP1	High deductible health plan 1	New variable
HDHP2	High deductible health plan 2	New variable
HSAHRA1	Health savings account/Health reimbursement account 1	New variable
HSAHRA2	Health savings account/Health reimbursement account 2	New variable
FSA	Flexible spending accounts for medical expenses	New variable

### Sample Adult Core

#### Adult Conditions Section (ACN)

<u>Variable Name</u>	<u>Brief Variable Description</u>	<u>Brief Description of Change</u>
DIBPRE1	Diagnosed prediabetes, impaired fasting glucose, impaired glucose tolerance, borderline diabetes, or high blood sugar	New Variable
AHEARST	Hearing good, little trouble, lot of trouble, or deaf	Dropped; replaced by AHEARST1 (supplemental variable in 2007)