
**2018 National Health Interview Survey (NHIS)
Public Use Data Release**



Paradata File Description

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

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

June 2019

Table of Contents

NCHS Website and NHIS Electronic Mail List	3
Introduction.....	4
Sample Design, Weighting, and Variance Estimation	5
Sample Design.....	5
Weighting.....	5
Variance Estimation	5
General Information about the 2018 Data.....	7
Information about the 2018 Paradata File Documentation.....	8
Paradata File Variables	10
Measures of Time	10
Measures of Contactability	11
Contact Strategies.....	12
Measures of Cooperation	13
Partials/Breakoffs.....	14
Mode Measures	14
Case-Level Information	14
Guidelines for Citation of Data Source	16
Appendix I. Linking the Paradata File to Public Use Health Data Files.....	17
Appendix II. Changes From the 2017 Paradata File.....	19

NCHS Website and NHIS Electronic Mail List

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

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

The website features links to downloadable public use data and documentation for the 2018 NHIS, as well as important information about any modifications or updates to the data and/or documentation. Published reports from previous years' surveys are also available, as are updates about future surveys and datasets.

For more information about the 2018 NHIS health data files, please see the 2018 NHIS *Survey Description* and supporting documentation for that release at the following site:

- ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2018/srvydesc.pdf.

Data users are encouraged to join the NHIS Listserv, an electronic mail list. The Listserv is made up of over 3,000 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 website, 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.

Introduction

This document describes the 2018 National Health Interview Survey (NHIS) Paradata File, a complementary file to the 2018 NHIS public use health data files.

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

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

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

The Paradata File is on a case (family) level, where one record represents one case. Unlike the NHIS public use health data release, which contains information on fully complete and sufficiently complete interviewed cases only, the Paradata File also contains data on other types of cases, including cases that were ultimately refusals, insufficient partials, and other types of nonresponse. These are referred to as *Type A* cases. In addition, data are provided on cases that were deemed out of scope, such as families with Armed Forces-only adults, and cases that were screened out by race/ethnicity. The out-of-scope cases are referred to as *Type B* outcomes in this document. A more detailed explanation of the outcome codes on the Paradata File can be found in the section “General Information about the 2018 Data” later in this document.

The Paradata File is intended as both a stand-alone data file and one whose fully complete and sufficiently complete cases can be linked with the NHIS regular public use health data file. More information about linking the regular health files and Paradata File can be found in Appendix I.

Sample Design, Weighting, and Variance Estimation

Sample Design

The sample design for the 2018 Paradata File is the same as that for the regular 2018 NHIS public use data files. This information can be found in the *Survey Description* accompanying the regular 2018 NHIS health data files release. This document is available at the following location:

- ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2018/srvydesc.pdf.

Weighting

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

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

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

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

NOTE: If data from the Paradata File are used in support of analyses of the 2018 NHIS health data files, the corresponding weights provided on the health data file should be used. For example, a data user may be interested in exploring the determinants of health care access and utilization among sample adults and would like to control for mode of interview. The necessary mode measures located on the Paradata File would be merged with the Sample Adult File, and the final sample adult weight (WTFA_SA) would be used in the subsequent analysis.

Variance Estimation

The data collected in the NHIS are obtained through a complex sample design involving stratification, clustering, and multistage sampling. While standard software packages may provide unbiased point estimates for commonly computed first-order statistics like means or regression coefficients, the computed estimated standard errors of the estimates may be too small if the software package does not account for the complex sample design. Also, standard packages may produce hypothesis test results (such as p values) that are

incorrect. Therefore, it is recommended that users of the NHIS Paradata File utilize computer software that provides the capability of variance estimation and hypothesis testing for complex sample designs. Analysts at NCHS generally use the software package SUDAAN® (RTI International) with Taylor series linearization methods to produce standard error estimates and hypothesis test results (such as p values). SUDAAN code and a description of its use to compute standard errors of means, percentages and totals with the 2018 NHIS data files can be found at the following location:

- ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2018/srvydesc.pdf.

General Information about the 2018 Data

The Paradata File contains 48,118 records. Of these records, 47,373 cases are considered “in scope,” i.e., cases that are eligible for interview. A case with an outcome code of 201 is a fully completed interview, in which the household composition, family, sample adult, and sample child (if a child was present) modules were completed. A case with an outcome code of 203 is a sufficient partial interview, meaning that at least a sufficient portion of the family module was completed. Cases with an outcome code of 213, 215, 216, 217, 218, or 219 are in-scope cases that did not result in either a fully complete or sufficiently complete interview. These include refusal cases (218) and insufficient partial cases (215) where the family module was started, but not completed to a sufficient degree. These are referred to as *Type A* or *nonresponding* cases. In addition, 745 cases are considered “out of scope” and thus ineligible for interview (also called Type B cases). Type B cases are families comprised of military-only adults, families whose usual residence is elsewhere, and families screened out by race/ethnicity. *Type B* cases are included in the data file because, although they are not counted in the NHIS response rate, they represent households where some contact was made with person(s) at that address in order to determine eligibility for the interview. Table 1 shows the frequency distribution of interview outcome codes for the Paradata File.

Table 1. Frequency distribution of outcome codes for 2018 National Health Interview Survey cases

Outcome code	Number of families
In-scope cases	
Interview cases	
201 – Fully complete case	24,756
203 - Sufficient partial case	5,669
Type A cases (nonresponding)	
213 - Language problem	195
215 - Insufficient partial case	1,438
216 – No one home, repeated calls	2,734
217 – Temporarily absent, no follow-up	283
218 – Refusal case	10,905
219 – Other Type A	1,393
Out-of-scope cases	
299 - Type B: Occupied entirely by Armed Forces adults or occupied entirely by persons with usual residence elsewhere,	745
Total	48,118

Information about the 2018 Paradata File Documentation

Datasets and related documentation for the Paradata File are available on the NHIS website, <http://www.cdc.gov/nchs/nhis.htm>. The website provides the Paradata File Description Document and supporting documentation. Information about the general 2018 NHIS data release is also available here.

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

- Variable Summary Report
- Variable Layout Report
- ASCII Data
- Variable Frequency Report
- Sample Program Statements

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

Recodes — A *recode* is a variable derived from the reordering, collapsing, or verbatim coding of another variable, such as the recode UNABLE1R found in the Paradata File. Alternatively, a recode may be constructed from two or more variables, such as the point in the interview period when the household composition section was started (STRTPNT). If a particular variable was used in making recode variables, then those recode variables are listed as a cross reference.

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

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

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.

Universe — The *universe* refers to those respondents eligible to be asked a given question. For example, the universe for many variables on the Paradata file is “All cases.” This means that all cases should have a valid value for the variable.

The Variable Frequency Report provides the frequencies, percents, and the frequency missing (not-in-universe) for each variable. Within the NHIS, the same codes are used across all files to designate “Refused” and “Don’t know” responses: refusals are coded as “7” (with leading 9’s to the length of the field, as in 7, 97, 997, etc.), while “don’t know” responses are “9” (again, with leading 9’s to the length of the field, such as 9, 99, 999, etc.). A code of “8” is used to indicate “Not ascertained” responses, which typically occur when an in-the-universe respondent had a blank field or the field contained an impossible code. There are few variables on the Paradata File that contain values of “Refused” or “Don’t know” because most variables are self-reported by the interviewer, and thus a value should be assigned to all cases. One item to note is that there are 407 cases that do not have a CHI record, because the interviewer did not record information for any of their contact attempts. This, however, represents a small proportion of total CHI cases.

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

Paradata File Variables

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

Measures of Time

A series of time measures can be found on the Paradata File. The first set of time measures convey information on the start dates of main sections of the survey instrument and the completion date of the case. For confidentiality purposes, actual dates are not provided. Instead, original date variables have been recoded into categorical measures that anchor the case within the monthly NHIS interview period.¹ For example, STRTPNT indicates that the household composition section of the survey instrument was started in the early, middle, or late portion of the monthly interview period. Similarly, ENDPNT indicates that the case was completed in the early, middle, or late portion of the monthly interview period. Among several possible uses, measures of this sort allow researchers to explore whether the quality of data (e.g., item nonresponse) collected varies by when in the interview period the case was started and completed.

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

- STRTPNT indicates that the household composition section of the instrument was started early in the interview period, in the middle of the interview period, or late in the interview period.
- FMSTRPNT indicates that the family module was started early in the interview period, in the middle of the interview period, or late in the interview period.
- SCSTRPNT indicates that the sample child module, if there was a sample child, was started early in the interview period, in the middle of the interview period, or late in the interview period.
- SASTRPNT indicates that the sample adult module was started early in the interview period, in the middle of the interview period, or late in the interview period.

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

¹In 2011, the NHIS transitioned to a monthly interview sample. An interviewer received a new assignment of sample cases on the first of the month and had until the end of the month to complete the assignments. Prior to 2011, there was a 17-day interview period to complete the assignment.

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

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

The following are descriptions of the time of day measures:

- HHC_TOD indicates that the household composition section was started in the morning, afternoon, or evening.
- FAM_TOD indicates that the family module was started in the morning, afternoon, or evening.
- SC_TOD indicates that the sample child module, if there was a sample child, was started in the morning, afternoon, or evening.
- SA_TOD indicates that the sample adult module was started in the morning, afternoon, or evening.

Measures of Contactability

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

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

- In-person noncontact measures: NCTP01R through NCTPR99R:

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

- Telephone noncontact measures: NCTL01R through NCTEL99R:

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

For each contact attempt that an interviewer makes on a sample household, she/he must report one of the following three outcomes: contact with a sample unit member, contact with a non-sample unit member, and noncontact. There are two sets of CHI-based noncontact measures based on mode of attempt.

Four measures are included under this conceptual grouping: CTSTAT1, CTSTAT2, CTSTAT3 and UNABLE1R.

- CTSTAT1, CTSTAT2, and CTSTAT3 provide the number of contact attempts that resulted in contact with sample unit members, the number of contact attempts that resulted in contact with non-sample unit members, and the number of contact attempts that resulted in noncontact.
- UNABLE1R provides the number of times contact was made with a sample unit member but the interviewer was unable to conduct or complete an interview because the “eligible person was not available.” This measure captures situations of sub-level noncontact -contact was made at the household level, but the household member needed to complete a particular section of the interview was not available. This largely pertains to the sample adult, followed by the sample child respondent.

There are several points to remember when working with these measures.

- As noted previously, CHI entries were not made for 407 cases.
- The CHI-based measures presented above were created by summing a particular entry across all contact attempts and then top-coding values greater than one. For example, if two personal visit contact attempts for a household resulted in contact with a sample unit member, but “eligible person was not available” was entered for both attempts, the summary CHI measure (UNABLE1R) would have a value of 1.
- These measures are based on interviewer self-reports. A corresponding amount of recall error is expected, especially when CHI entries were not made at the time of the contact attempt.

Contact Strategies

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

- Advance letter given (STRAT01R): Interviewers are instructed to ask the household member if she/he received the advance letter (mailed from the regional office several weeks before the interview period). If the household member did not receive the letter, the interviewer will provide the household member with a copy of the letter. The letter can also be left in situations of noncontact.
- Scheduled appointment (STRAT02R): If the respondent is unavailable at the time of contact, the interviewer may choose to schedule an appointment some other time during the interview week.
- Left note/appointment card (STRAT03R): If no contact is made at the time the interviewer visits the household, she/he may leave a personalized note or an appointment card.
- Left promotional packet/informational brochure (STRAT04R): If the householder is not available at the time of visit, the interviewer may give the householder informational materials to review before an

interview is conducted. These materials may include brochures about how the NHIS data are used, a ruler, magnet, or other information related to the survey.

- Called household (STRAT05R): The interviewer may choose to call the household to make contact. This is usually done after a personal visit has been attempted.
- Left message on answering machine (STRAT06R): Interviewers may choose to leave a message on an answering machine for the householder.
- Waited for respondent (STRAT11R): The interviewer may wait near the residence in the hopes that the interviewer will see the householder arrive.
- Checked with neighbors (STRAT12R): The interviewer may try to learn more about the best time to reach the householder by asking neighbors about the householder's schedule.
- Contacted other family members (STRAT13R): If the interviewer has difficulty contacting a respondent, she/he may choose to contact a relative of the respondent in order to help to locate the respondent.
- Contact property manager/doorman (STRAT14R): The interviewer may need to contact the property manager or doorman if the community where the household is located has limited access.
- None (STRAT98R): No strategies were used on this contact attempt.
- Other-specify (STRAT99R): Some other strategy not listed was used on the contact attempt.

Measures of Cooperation

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

- Language concerns (LANG1R–LANG5R): These dichotomous coded measures (0=never entered, 1=entered at 1 or more contacts) describe contact outcomes in contact attempts, both in person and over the telephone. Some examples include “no household member able to translate” and “unable to find translator.”
- Respondent reluctance (RELC01R–RELUC99R): These dichotomous coded measures (0=never entered, 1=entered at 1 or more contacts) represent contact outcomes in contact attempts, both in person and over the telephone. Some examples include “too busy” and “interview takes too much time.”
- Unable to complete interview (UNABLE2R–UNABL99R): These dichotomous coded measures (0=never entered, 1=entered at 1 or more contacts) represent contact outcomes in contact attempts, both in person and over the telephone. Some examples include “inconvenient time” and “health problem.” These variables could also be conceptually grouped with the partial/break-off variables from the file.

Partials/Breakoffs

A series of questions to ascertain reasons for partially completed interviews (interviews that are not entirely complete) appear in the Back section of the instrument. PARWHY is a variable that differentiates between the interview being stopped in progress (break-off) and being unable to reach a particular respondent or obtain some information on follow-up (partial, other than break-off). For cases marked as a break-off in PARWHY, BRKWHER is a variable that determines the main section of the survey in which the break-off occurred. BRKRES1 allows interviewers to select one main reason that the interview stopped in progress from a list of ten answer categories. Similarly, NCOMRES allows interviewers to record the main reason that the interview is not complete.

Mode Measures

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

Back Section Variables

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

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

CHI Variables

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

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

Case-Level Information

There is also a set of variables that give general case-level information about the record. For each case, the variable SRVY_YR (survey year) will have a value of “2018.” This variable is more useful when pooling years of

data for trend analysis. Interview quarter (INTV_QRT) and interview month (INTV_MON) indicate the general time when the case was in the field. There are 4 quarters and 3 months in each quarter.

The variable OUTCOME1 represents the disposition of the case. If the Paradata File is linked to the health data files, it is important to remember that there is not a 1:1 match between the two. The Paradata File represents field data before cleaning. There are some 201 and 203 outcomes (fully complete and sufficiently complete interviews) on the Paradata File that were dropped from the final 2018 public use health files because the record was not deemed adequate for data release (for example, there may have been too many “Don’t know” or “Refused” responses). The variables QCSASCFM and QCFAM flag records that have been removed from the health data files for quality reasons.

As mentioned before, the Paradata file also contains in-scope but not sufficiently complete cases (refusals, insufficient partial cases, etc.) and out-of-scope cases (usual residence elsewhere, screened out by race or ethnicity, etc).

Guidelines for Citation of Data Source

With the goal of mutual benefit, the National Center for Health Statistics (NCHS) requests that recipients of data files cooperate in certain actions related to their use. Any published material derived from the 2018 Paradata File should acknowledge NCHS as the original source. The suggested citation to appear at the bottom of all tables and graphs is as follows:

- Data Source: NCHS, National Health Interview Survey, 2018

In a bibliography, the suggested citation for this document should read:

- National Center for Health Statistics. Paradata File Description, National Health Interview Survey, 2018. Hyattsville, Maryland. 2018.

The suggested citation for 2018 NHIS survey data and other documentation should read:

- National Center for Health Statistics. National Health Interview Survey, 2018. Public-use data file and documentation. http://www.cdc.gov/nchs/nhis/quest_data_related_1997_forward. 2019.

The published material should also include a disclaimer that credits the author's analyses, interpretations, and conclusions 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 ensure 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/products/citations.htm>.

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

The variables HHX and FMX can be used to link the 2018 Paradata File with the 2018 NHIS health data files. HHX is a household identification number for each record in both files. FMX represents the family number of the case within the household. For approximately 98% of 201 or 203 cases (complete or sufficiently complete interviews), there is a single family in the household, thus the FMX value will be "01". For those cases with more than one family in the household, HHX will be the same, but the FMX will have values of "01", "02", etc.

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

For example, the 2018 Family Health File has 30,309 records. Merging on the variables HHX and FMX, 30,309 of the 48,118 records from the Paradata File will be pulled in to complete the merge. The final size for the merged file in this case should be 30,309 records. If making population inferences using this linked file, the final annual weight on the health data file (WTFA_FAM) should be used.

An example of SAS code to merge the 2018 Paradata File with the 2018 NHIS Family File is below:

```

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

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

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

SET NHIS.PARADATA;
/*The SET statement reads data from the 2018 Paradata File. */
PROC SORT DATA=PARADATA;
/* Sort by HHX, and FMX, the household and family identifier. */
BY HHX FMX; RUN;

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

DATA FM (KEEP=HHX FMX INCGRP RAT_CAT WTFA_FAM);
/* FM is a SAS dataset; the KEEP statement retains only the listed variables for processing. */
SET NHIS.FAMILYXX;
/*The SET statement reads data from the 2018 Family file. */
PROC SORT DATA=FM;
/* Sort by HHX and FMX the household and family identifier. */
BY HHX FMX; RUN;

DATA PARAFM;
/* New combined dataset called PARAFM */
MERGE FM (IN=FROMFM) PARADATA ;
/* Merge the newly created FM and PARADATA files, using an IN statement. */
BY HHX FMX; IF FROMFM = 1;
/* The combined dataset PARAFM will contain only those records that are in the Family file; the
Paradata file's MODE variable will be appended to these records. */

```

```
PROC SORT DATA= PARAFM;  
/* Sort by HHX and FMX, the household and family identifiers. */  
BY HHX FMX; RUN;
```

In the code above, the IN statement creates a temporary SAS variable (called FROMFM) that has a value of 1 if the dataset associated with the IN statement contributed to the current observation, or a value of 0 if it did not. The subsequent statement “IF FROMFM = 1” tells SAS to retain only those observations from the Family file (called FM).

Appendix II. Changes From the 2017 Paradata File

- There were no content changes to the 2018 Paradata File, compared to the 2017 file. A new redesigned NHIS questionnaire content and structure began fielding in January 2019, with a split-sample of this new survey also fielded in quarter 4 of the 2018 NHIS. The redesigned portion of the quarter 4 split-sample is not part of the 2018 public data release.

For additional information about the 2019 redesign, please visit:

https://www.cdc.gov/nchs/nhis/2019_quest_redesign.htm