The 2006 National Health Interview Survey (NHIS) Paradata File: Overview and Applications

Beth L. Taylor
National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782

Abstract

In January 2008, the National Health Interview Survey released its first annual Paradata File, containing information about the data collection process for the 2006 NHIS. Analyses of paradata can be used to explore case characteristics and allow for a better understanding of NHIS respondents and the interviewer–respondent dynamic. This discussion will provide an overview of the 2006 NHIS Paradata file, including the scope of the cases included on the file and a summary of major variables related to interviewer strategies, measures of contactability and cooperation, measures of time, mode of interview, and reasons for partial interviews and break-offs. Applications of the data as a stand-alone file and as a file linked to the 2006 NHIS health data files will also be discussed.

Key Words: Paradata, cooperation, contact attempt, mode, partial

1. Introduction

In January 2008 the National Center for Health Statistics (NCHS), part of the Centers for Disease Control and Prevention (CDC), released the 2006 National Health Interview Survey (NHIS) Paradata File, a complementary file to the 2006 NHIS public use health data files that were released in June 2007.

The NHIS is a multi-purpose health survey and is the principal source of national estimates 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. NHIS data are collected through computer-assisted personal interviewing (CAPI) using government-issued laptops. The NHIS is an in-person interview with some telephone follow-up permitted to complete the case when necessary. Up to about 35,000 families containing about 87,500 persons are interviewed annually. The NHIS consists of four core modules: Household, Family, Sample Child and Sample Adult. The Household and Family sections are completed by knowledgeable adults, and one randomly selected adult and child (if a child is present) are selected within each family to receive more detailed health questions. The NHIS has been conducted continuously since its establishment in 1957, and public use microdata files are released on an annual basis.

The NHIS Paradata File does not contain health-related data, but rather contains paradata, which are data about the survey process. The majority of NHIS paradata are recorded by the Census interviewer during and immediately after the NHIS interview. The 2006 Paradata File has dual uses: it can be analyzed as a stand-alone file on interview characteristics, or linked with the 2006 health data files. Although the 2006 NHIS Paradata File was released approximately 6 months after the 2006 NHIS health data files, for the 2007 and subsequent data year releases, paradata and health data files will be released simultaneously.

2. Sources of NHIS Paradata

2.1 The Contact History Instrument (CHI)

NHIS paradata come from a number of sources, the largest of which is the Contact History Instrument (CHI). The CHI is a supplement piece to the NHIS that collects data about each contact attempt. A contact attempt occurs each time the interviewer tries to make contact with the sample household, whether driving by the house to see if someone is home, knocking at the door, or attempting to call the household via telephone. Whether or not the effort results in a contact, the interviewer will record information about the attempt in the CHI. The CHI was developed primarily by the Census Bureau and was incorporated into the NHIS reengineering from a DOS-based (CASES) to Windows-based (Blaise) data collection software in 2004. The CHI is standardized for use on other Census Bureau Surveys.

Specific information collected in the CHI includes whether the contact attempt was made in person or over the phone, and whether the attempt resulted in a contact or noncontact. If the attempt resulted in a contact, interviewers are asked for a description of the contact attempt, any type of reluctance shown by the respondent (busy, not interested, etc.), and any strategies employed by the interviewer to complete the case. The noncontact path includes a description of the noncontact (no one home, repeated knocks, etc.), and any strategies used to gain future participation (left note/appointment card, etc.).

Data from the CHI are output into two files - a case-level (family) file, which has summary data on the number of contacts, noncontacts, measures of reluctance, and strategies used for a single case, as well as an attempt-level (visits) file, which collects the date and time of each contact attempt and a description of the outcome for a particular visit. For the 2006 Paradata File release, the data are primarily from the case-level file.

2.2 Other Sources of NHIS Paradata

Another source of NHIS paradata comes from the Front/Back sections of the instrument. The Front section asks a series of questions to determine whether the address is eligible for interview. Information is collected on in-scope but nonresponding cases, such as refusals and cases where household members are temporarily absent as well as cases where the dwelling falls out of scope (demolished, vacant, etc.).

The Back section of the NHIS is completed at the conclusion of the interview. It contains a series of questions for the interviewer about the case, including mode of interview (in-person visit vs. phone interview), the assessed cooperativeness of the respondent, and reasons for partial interviews or break-offs in cases where the interview could not be fully completed. Unlike the CHI, which as mentioned previously is standardized for use on multiple Census Bureau Surveys, the Front/Back questions are tailored to the NHIS.

Other sources of NHIS paradata include a time file, which contains the total interview time in addition to module and within-module section times, as well as audit trail (keystroke) data, which include item times, the dates and times when sections of the instrument were started and completed, and function key use. For the 2006 data release, much of these data have been released in aggregate form, due to confidentiality constraints.

3. General Information About the Paradata File

There are 125 variables in the 2006 Paradata File. The 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 contains data on several outcomes from 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, paradata are provided for 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. As mentioned previously, the Paradata File is intended as both a stand-alone data file and one that can be linked with the NHIS public use health data files for the fully complete and sufficiently complete interviews. It is important to note, however, that the Paradata File has a slightly larger number of interviewed cases than the health files because the Paradata File represents field data before cleaning, so that there will not be a complete 1:1 match when linking the files.

The 2006 Paradata File contains 44,264 records. Of these records, 34,270 cases are considered “in scope,” i.e., cases that are eligible for interview. A case with an outcome code of 201 is a fully completed interview, in which the household composition, family, sample adult, and sample child (if a child was present) modules were completed. A case with an outcome code of 203 is a sufficient partial interview, meaning that at least a sufficient portion of the Family Module was completed. Cases with outcome codes of 213, 215, 216, 217, 218, or 219 are in-scope cases that did not result in either a fully complete or sufficiently complete interview. These include refusal cases (218) and insufficient partial cases (215) where the Family Module was started, but not completed to a sufficient degree. These are referred to as *Type A* or *nonresponding* cases. In addition, 9,994 cases are considered “out of scope” (299) and thus ineligible for interview (these are 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 Paradata File because, although they are not counted in the NHIS response rate, they represent households where some contact was made with person(s) at that address in order to determine eligibility for the interview. Table 1 shows the frequency distribution of interview outcome codes for the Paradata File.

Table 1: Frequency Distribution of Cases by Outcome Code

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

4. Paradata File Documentation

Datasets and related documentation for the 2006 NHIS Paradata File and the general 2006 data release are free to download and publicly available on the NHIS Website, **http://www.cdc.gov/nchs/about/major/nhis/nhis_2006_data_release.htm**. The Paradata File includes the supporting documentation listed below.

- Paradata File Description Document
- Variable Summary Report
- Variable Layout Report
- ASCII data set
- Sample SAS program
- Variable Frequency Report

The *Paradata File Description Document* contains information about the sample design and variance estimation measures for the file. It details the 125 variables on the file by grouping them conceptually into measures of time, contactability, cooperation, contact strategies, variables related to partial and break-off interviews, mode measures, and case-level information such as geographical region, case disposition codes, interview quarter, etc. This document also has information and a sample program to link the Paradata File with the corresponding public use health data files. Annual changes to variables will be listed in this document in subsequent years to aid trend analysis.

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. The *Variable Layout Report* is a more detailed document, which includes the variable universe, source of the variable (such as from the CHI instrument), and, where applicable, lists the questions and response codes (many of the variables in the Paradata File, such as the time-related variables, are collected behind the scenes and thus do not have actual questions in the instrument). The file layout is in searchable PDF

[1] Prior to interviewing, one part of the NHIS sample is assigned to be "screened". In that 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, the interview continues regardless of racial/ethnic household composition.

format (as is all text documentation), so that "Keywords" listed for each variable can be used to search for topics of interest.

The Paradata File contains several recodes. A recode is a variable derived from the reordering, collapsing, or verbatim coding of another variable, such as the item (HOWLNGWK - weeks without telephone service) found in the Paradata File. Alternatively, a recode may be constructed from two or more variables, such as the point in the 17-day interview period when the case was started (STRTPNT). The Paradata File contains a number of recodes related to the date and time variables, including the points in the interview period when each core module was started and the times of day when the modules were 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".

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 a respondent is asked any questions pertaining to time - for example, how long the family has been without phone service - the answer is typically obtained in two parts; the respondent provides the number of time units, followed by the type of time unit. During the course of data editing, this information is standardized into a single appropriate time unit. Examples of this in the Paradata File include RH1LNGDY, and RH2LNGDY (days without telephone service).

The Paradata File release also contains an *ASCII data set* and *sample SAS program*. The *Variable Frequency Report* provides the frequencies, percents, and the frequency missing (not-in-universe) for each variable. There are few variables in the Paradata File that contain values of "Refused" or "Don't know" because most items are self-reported by the interviewer. One item to note is that there are 988 cases that do not have a CHI record, because the interviewers did not record information for any of their contact attempts. This, however, represents a small proportion of total CHI entries (approximately 2%).

A few of the variables in the Paradata File are also found in the NHIS health data files. In particular, the items on telephone usage and outages are found on both files. They are included in the Paradata File because insufficient partial cases and other Type A and Type B cases that are not part of the NHIS health data files may contain this information and be of use to analysts interested in coverage issues.

5. Weighting and Variance Estimation

A weight variable (WTIA_PD) is included in the Paradata File. This weight reflects the probability of household selection and does not include nonresponse and post-stratification adjustments. It is the correct weight to use when conducting analysis of the Paradata File with the goal of making population inferences. If the analysis does not involve making population inferences, use of the paradata weight would not be required. When using the Paradata File to support analyses with the NHIS health data files, the weight from the health file should normally be used, as opposed to the paradata weight. For example, if merging the Paradata File and the Sample Adult file to determine if national health estimates differ by mode of interview, the Sample Adult weight WTFA_SA should be used in the analysis.

NHIS data are collected through a complex sample design involving stratification, clustering, and multistage sampling. It is strongly recommended that users of the NHIS Paradata File utilize computer software that provides the capability of variance estimation and hypothesis testing for complex sample designs. NCHS uses the software package SUDAAN (Research Triangle Institute, 2004) with Taylor series linearization methods for NHIS variance estimation.

6. Examples of Possible Research

6.1 Current Examples

Examples of analyses using the Paradata File as a stand-alone file include determining the average number of contact attempts to complete an interview, as well as the point in the interview period (Early, Middle, Late) and time of day (Morning, Afternoon, Evening) when the Sample Adult module was started. A data user could also investigate which strategies employed by the interviewer led to successful completion of the interview.

Examples of possible research using the Paradata File to support analysis of the health data files are determining the socio-demographic and economic characteristics of hard-to-contact families and modeling the impact of mode of interview on determinants of health. As noted above, the type of analysis conducted should determine if a weight is used and the type of weight employed.

6.2 Future Steps

As mentioned previously, the 2007 NHIS Paradata File was released in conjunction with the 2007 NHIS health microdata in June 2008. The 2007 Paradata File is similar to the 2006 Paradata File, with a few minor revisions to the questions regarding telephone ownership and usage. Although the revised 2007 items were given slightly different variable names, per NHIS protocol, data users interested in pooling data for trend analysis are advised to read the appendices of subsequent Paradata Description Documents to learn about additional annual changes. The 2007 Paradata File includes sample SPSS and Stata sample programs, in addition to the SAS sample program found in the initial Paradata File release.

For future releases, inclusion of additional data from the CHI visits file is planned, as well as possible audit trail data (function key use, toggling between language of interview, etc.). An expansion of the data released from the Front/Back sections and time files will also be considered. NCHS encourages feedback from data users about the current Paradata File as well as the type of paradata users would like to receive in forthcoming releases.

Acknowledgements

The author would like to thank the other members of the NHIS Paradata Committee - Pei-Lu Chiu, James Dahlhamer, Catherine Simile, and Barbara Stussman - for their guidance and support in the development of the 2006 and subsequent Paradata Files and for their ongoing analyses with NHIS paradata. Special thanks to Dr. Jane Gentleman for establishing the Paradata Committee and charging its members with the creation of the first publicly available Paradata File for the NHIS.

References

National Center for Health Statistics (2007). *Survey Description Document, National Health Interview Survey, 2006* . *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2006/srvydesc.pdf*

National Center for Health Statistics (2008). *Paradata File Description Document, National Health Interview Survey, 2006* . *http://www.cdc.gov/nchs/data/nhis/filedescriptiondoc.pdf*

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