# Defining and Updating the National Health Interview Survey Variance Estimation Structure Over a Sample Design Period

Chris Moriarity, Van Parsons
National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782 USA

**Abstract**
The National Health Interview Survey (NHIS) is a multi-purpose health survey conducted by the National Center for Health Statistics (NCHS). Public-use microdata files and complex sample design variance estimation structures are available at the NCHS Internet web site. NCHS also has restricted microdata files and separate variance estimation structures for internal use. NCHS defined the variance estimation structures for the first year of a new sample design period, and then updates them annually to account for new sample areas. The updates are done in a way to ensure consistent definitions over a sample design period, which is important when conducting analyses of NHIS annual data pooled across years. We describe our methodology for defining the structures and updating them on an annual basis.

**Key Words:** Sample Survey

## 1. Introduction

The National Health Interview Survey (NHIS) is the principal source of information on the health of the civilian noninstitutionalized population of the U.S. It is a continuous survey that has been in operation since 1957. The current NHIS sample design was implemented in 2006 and will be in place through the end of 2015. During the current sample design period, NCHS anticipates obtaining completed interviews at approximately 35,000 living quarters (households and noninstitutional group quarters such as college dormitories) each calendar year if there are no sample reductions or augmentations. All persons residing at a sampled address are covered by the NHIS interview, yielding a sample of approximately 87,500 persons each year if there are no sample reductions or augmentations. Sample sizes can increase or decrease appreciably, according to the availability of funding. Each interview is conducted via a personal visit to the living quarters by an employee of the U.S. Census Bureau, which is the data collection agent for the NHIS.

The NHIS sample consists of clusters of living quarters (addresses) chosen within a first-stage sample of U.S. counties. The NHIS sample design is complex, with multiple stages of sampling. The first stage of sampling is selection of primary sampling units (PSU), which are one or more contiguous U.S. counties. In the current NHIS sample design, PSU boundaries almost always lie within a single state. The PSUs are assigned to sampling strata, and then one or two PSUs are selected from each sampling stratum. The sampling strata generally were formed within state, and in more populous states there were additional stratum definition characteristics such as metropolitan status, etc. Once the sample PSUs are selected for a given sample design period, they delineate the geographic areas where the NHIS sample addresses will be selected for the entire sample design period. There are additional stages of sampling within each PSU, yielding a sample of addresses after the final stage of sampling. This sampling method is used to control the costs related to personal visit interviewing. The cost and time required to conduct personal visit interviews in a simple random sample of U.S. living quarters would be prohibitive, due to the amount of travel that would be required.

The current sample design of the NHIS (2006-2015) is based on Census 2000 information. The current sample design is very similar to the previous sample design, which was in effect from 1995 to 2005 and was based on 1990 Census information. For example, one similarity is that in both the previous and current sample design, the annual NHIS sample includes addresses in all 50 states and the District of Columbia. The NHIS has undergone changes in the sample design at intervals of approximately ten year duration, using information from the previous decennial census.

Additional information about the NHIS is available online at the NHIS home page, http://www.cdc.gov/nchs/nhis.htm. The reference section of this paper includes publications that describe the NHIS sample designs all the way back to when the survey began in 1957. All NCHS publications that describe the historic NHIS sample designs are available online at:

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

The NHIS has a complex sample design. Specialized software and methods are required to produce valid variance estimates from NHIS data. This paper describes how NCHS defined variance estimation structures for the current NHIS sample design, and our procedure for updating the structures on an annual basis.

## 2. Overview of the Current NHIS Variance Estimation Structures

NCHS creates two variance estimation structures for the current NHIS. One is provided to the public for use with analyses of NHIS public-use microdata files ("public structure"), and the other is used within NCHS for internal analyses of NHIS data ("inhouse structure"). Both structures consist of a Pseudo-Stratum variable and a Pseudo-PSU variable.

NCHS provides online guidance for the use of the public structure for a variety of complex survey variance estimation software - SUDAAN, the SAS survey procedures, Stata, SPSS, VPLX, and the R "survey" add-on package at: http://www.cdc.gov/nchs/nhis/methods.htm

Parsons and Moriarity (2007) provide a comprehensive description of the public structure features. The public structure always has at least 2 Pseudo-PSUs within each Pseudo-Stratum. The public structure is related to the actual sample design structure (the actual sampling strata and PSUs), but is constructed in a way to reduce disclosure risk.

The inhouse structure is similar to the public structure. It is more closely related to the actual sample design structure, and it provides for more statistically efficient variance estimation. Several comparisons (Parsons and Moriarity, 2007; Davis and Parsons, 2011) suggest that the inhouse structure and the public structure generally yield consistent variance estimates. The inhouse structure has more robust performance for internal analyses that cannot be done with the public use microdata files, e.g, state-level analyses. (State identifiers are not available on NHIS public use microdata files.)

## 3. Defining the Current NHIS Variance Estimation Structures

The initial definition of the NHIS variance estimation structures for a sample design period corresponds to the sample areas for the first year of the sample design period, e.g., 2006 in the current sample design period. The process for defining the 2006 structure follows.

Some details of the variance estimation structure definition methodology are not discussed here because of disclosure risk.

There are two types of NHIS PSUs: "self-representing" (SR), and "non-self-representing" (NSR). SR PSUs typically consist of one or more adjacent counties that define a metropolitan area within a state, or the portion of a metropolitan area within a state. The population size of an SR PSU is large enough that it is selected into sample with certainty. Conceptually, the first stage of sampling for an SR PSU is not at the PSU level, but within the PSU. NSR PSUs also typically consist of one or more adjacent counties within a state. The PSUs in this group have smaller population sizes than the SR PSUs. As part of preparing for the current NHIS sample design, NCHS assigned a collection of NSR PSUs within a given state to a given sampling stratum, and then a sample of one or two NSR PSUs was selected. Conceptually, the first stage of sampling for NSR PSUs is at the PSU level.

It is necessary to construct variance estimation structures that are distinct from the actual sample design structure for two main reasons:

● As described above, the actual sample design structure contains occurrences where only one PSU is sampled from a sampling stratum. The default action for certain variance estimation software is to return a variance estimate of zero for this situation. To obtain valid variance estimates, the structure must account for the sampling within SR PSUs, and include collapsed pseudo-strata for NSR sampling strata where only one NSR PSU was selected.

● Releasing too much detail of the actual sample design structure could increase disclosure risk.

The initial variance estimation structure definition process occurred in tandem with the creation of four NHIS "panels". The NHIS is a cross-sectional survey, not a longitudinal survey; for NHIS, the term "panel" refers to a subset of an annual NHIS sample that can be viewed as a probability subsample of the total NHIS sample for that year. NSR PSUs were assigned to one panel, and SR PSUs were assigned either to two or four panels, depending on the PSU population size. NCHS and the Census Bureau have worked together for the past several NHIS sample designs to define four NHIS panels that form a partition of each annual NHIS sample. This structure is useful in several ways. NCHS has determined that cutting one or more panels from the NHIS sample is the most efficient way to implement sample reductions when budget shortfalls occur. Another use of the NHIS panels is to delineate a probability subsample of the NHIS sample for other uses. For example, since 1996, the annual sample addresses for the Medical Expenditure Panel Survey, administered by the Agency for Healthcare Research and Quality, are a subset of the interviewed NHIS sample addresses from two NHIS panels of the previous year.

NCHS grouped the NHIS PSUs into approximately 100 "super strata" as part of the panel creation mechanism. The super strata were formed within Census Region by similarity of SR and NSR status, stratum size, geography, race and ethnicity characteristics, and metropolitan area status. The variance estimation structures also were created using the super strata. Consistent with the actual first stage of sampling, an inhouse pseudo-PSU generally was formed either from an entire NSR PSU, or from a portion of an SR PSU. The inhouse pseudo-strata were formed within the super strata. A range of 2 to 33 inhouse pseudo-strata were formed within each super stratum. For the public structure, some additional masking occurred for both pseudo-PSU and pseudo-stratum, and some structure coarsening was done so that the structure would still be robust if budget shortfalls in a given year resulted in sample cuts. The final public structure has 600 Pseudo-PSUs in 300 Pseudo-Strata, with two Pseudo-PSUs per Pseudo-Stratum. The final inhouse structure has a larger number of Pseudo-Strata (more than 500), with a variable number of Pseudo-PSUs per Pseudo-Stratum. We keep the public structure static during the sample redesign period; in some cases we allow the inhouse structure to expand as new areas enter the NHIS sample over the sample design period.

The Census Bureau supplied NCHS with a reference file in 2006 that contained much of the detailed annual sample information for the entire sample design period that began in 2006. This enabled NCHS to make assignment of annual variance structure variables to the majority of the NHIS sample in advance.

## 4. Updating the Current NHIS Variance Estimation Structures

The NHIS variance estimation structures require some updating on an annual basis. This is necessary because although the NHIS sample PSUs generally remain static, there are changes in the areas where each year's annual sample is located within each sample PSU. A given year's variance estimation structure corresponds to the areas that are in sample that year. Additionally, the "permit frame" portion of the NHIS sample typically contributes some new areas to the NHIS sample each year. (The "permit frame" is a small portion of the total NHIS sample that allows residences constructed after April 1, 2000 a chance to be selected into the NHIS sample.)

An important consideration with updating the variance estimation structure is to maintain consistency within a sample design period. Analysts commonly combine, or "pool", two or more annual NHIS microdata files to obtain larger sample sizes for analysis of questionnaire items that are consistent across the years for which the data are being combined. For variance estimation, a given geographic area within a given NHIS sample PSU should have the same set of Pseudo-Stratum and Pseudo-PSU codes assignments if it is present in more than one NHIS annual microdata file.

It is possible for an area to be in sample one year, not in sample for one or more subsequent years, and then back in sample again. The year-to-year update process must keep track of such changes to assure that the area receives consistent assignment when it is in sample, and to avoid reassigning that area's code to a different sample area during a year when the original area with the code is not in sample.

The advance assignment of annual variance structure variables to the reference file helped to assure consistent year-to-year assignments. What remains for each annual update is to make assignments of variance structure variables to new permit frame areas in a consistent way with previous assignments.

## 5. Automating the Variance Estimation Structure Update Process

During the 1995-2005 NHIS sample design period, the annual update process was done manually. The manual process was tedious and time-consuming, and became more complicated in the later years of the sample design period.

We decided to try to automate the annual update process for the 2007 NHIS. To do this, we needed to assign consistent codes for areas in sample both in 2006 and 2007. We also needed to identify sample areas that were new in 2007 and assign codes to them that were consistent with the structure defined for 2006. It was a straightforward process to assign consistent codes for areas in sample both in 2006 and 2007. It also was a straightforward process to identify sample areas that were new in 2007, and check the reference file for assignments already done. The challenge for the new sample areas without assignments in the reference file was to create software that would recognize existing patterns of code assignments for a given Pseudo-Stratum, and then continue this code assignment pattern for new areas assigned to that Pseudo-Stratum. This needed to be done for Pseudo-Strata in both the public and inhouse structures. This challenge took time and effort to address; ultimately we were successful.

We carried out the 2008 NHIS annual update process using a minor modification of the software used for the 2007 update. A potential pitfall is that we did not refer to both the 2006 and 2007 assignments, only the 2007 assignments.

Our initial 2009 update continued what was done in 2007 and 2008; we referred back to only the 2008 assignments, not the 2006, 2007, and 2008 assignments. Fortunately, we noticed in our review that there were inconsistencies. Specifically, there were 2007 sample areas that were not in sample in 2008, but were in sample in 2009; not referencing all prior assignments had led to inconsistent assignments for 2007 and 2009.

We realized that each annual update needed to reference all previous assignments to assure that any sample area previously assigned codes was given the same codes if they reoccurred, and to assure that any new assignments that needed to be made accounted for all earlier assignments.

After completing the 2009 update, we reviewed the 2008 assignments and determined that we had not inadvertently introduced any inconsistencies.

Given that we discovered that our first code assignment for the 2009 NHIS sample had flaws, we decided to create independent verification software that compares the tentative public and inhouse structures generated for a given year to all previous years in the sample design period to check for errors. The verification process includes a step of partitioning the concatenated annual structures by all possible inclusion/exclusion combinations, followed by a reassembly step that checks for inconsistency in structure assignment codes. For example, the original verification software partitioned the concatenated 2006, 2007, 2008, and 2009 structures into all possible inclusion/exclusion combinations across those four years. Given that the number of possibilities for a sample area being in/out of sample over a period of years essentially doubles with each additional year, the independent verification software has become more complex as the sample design period has lengthened.

To elaborate: in 2006 an area is either in or out of sample. The situation of being out of sample is not of importance, so the number of possibilities is 2-1=1. For the combination of 2006 and 2007, the possibilities are: in sample both years, in sample only in 2006, in sample only in 2007, not in sample either year. Again, the situation of not in sample either year is not of importance, so the number of possibilities is $2^2$-1=3. For the combination of 2006-2008, the number of possibilities is $2^3$-1=7 (in all 3 years, in 2006 and 2007 only, in 2006 and 2008 only, in 2007 and 2008 only, in only 2006, in only 2007, in only 2008). Etc.

## 6. Conclusion

It is worthwhile to expend the necessary effort at the beginning of an NHIS sample design period to define the variance estimation structures for the entire sample design period. This approach enables us to create structures that we anticipate will have robust performance throughout the sample design period, even if budget shortfalls lead to sample cuts.

Automating the annual update process was a challenging task. There are several benefits of automating the process, such as: we are able to complete it more quickly; the process is more transparent; the process can be replicated by rerunning the update software. Modifying the independent verification software requires more effort in the later years of a sample design period, because the number of possibilities for how an area can be in/out of sample over the elapsed number of years in the sample design period essentially doubles each year.

## References

Botman S, Moore T, Moriarity C, and Parsons V. Design and Estimation for the National Health Interview Survey, 1995-2004. Vital Health Stat 2(130). 2000.

Davis K, Parsons V. Comparison of Variance Estimates in a National Health Survey. 2011 Proceedings of the Joint Statistical Meetings, 567-573.

Kovar M, Poe G. The National Health Interview Survey design, 1973–1984, and procedures, 1975–83. Vital Health Stat 1(18). 1985.

Massey J, Moore T, Parsons V, Tadros W. Design and estimation for the National Health Interview Survey, 1985–94. Vital Health Stat 2(110). 1989.

National Center for Health Statistics. The statistical design of the Health Household-Interview Survey. Health Statistics. PHS Pub. No. 584-A2. Public Health Service. Washington: U.S. Government Printing Office. 1958.

National Center for Health Statistics. Health Interview Survey procedure, 1957–1974. National Center for Health Statistics. Vital Health Stat 1(11). 1975.

Parsons V, Moriarity C. Review of NHIS Public-Design Structures. 2007 Proceedings of the Joint Statistical Meetings, 2903-2909.

Parsons V, Moriarity C, Jonas K, Moore T, Davis K, and Tompkins L. Design and Estimation for the National Health Interview Survey, 2006-2015. Vital Health Stat 2(165). 2014.