Overview of the 2016-2025 National Health Interview Survey Sample Design

Chris Moriarity (NCHS), Van Parsons (NCHS), Kim Jonas (Census)
National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782 USA
U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233 USA

Abstract
A new sample design was implemented for the National Health Interview Survey (NHIS) at the beginning of 2016. The new NHIS sample design contains several new features. One new feature is increased flexibility to implement changes in the sample size and/or sample allocation. Another new feature is a different source of most of the sample addresses, relative to the 1985-2015 survey period.

Key Words: Sample Survey

1. Introduction

The National Health Interview Survey (NHIS) is a 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 and is administered by the National Center for Health Statistics (NCHS). The current NHIS sample design was implemented in 2016 and is anticipated to be in place through the end of 2025. For 2016-2018, the first three completed years of the survey within the current sample design, the initial sample of addresses selected for the NHIS was in the range 57,500-59,000, not accounting for sample reductions or augmentations. The final number of interviews is less, due to vacant units, out-of-scope units, and nonresponse. As of 2019, a new questionnaire was implemented for the survey, with a projected sample yield of 27,000 sample adult and 9,000 sample child interviews if there are no sample reductions or augmentations. Sample sizes can increase or decrease appreciably, according to the availability of funding. Sample augmentation due to increased funding occurred in 2016, 2018, and 2019. 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.

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 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

Two major changes were implemented at the beginning of the current NHIS sample design in 2016. The first was a major change in the main source of sample addresses, and the second was increased flexibility to accommodate changes in annual sample sizes. This paper describes both of the changes.

2. Continuing Features of the 2016 NHIS Sample Design

The NHIS has always been a household survey with interviews conducted by interviewers visiting the household's address, and will continue to be so in the foreseeable future. The sample distribution for a household survey with personal visit interviews usually has some level of geographic clustering, to reduce interviewer travel. This has been, and will continue to be, a feature of the NHIS sample design.

The precision of national-level annual estimates has been and remains a high priority for the NHIS sample design. To meet other geographical estimation objectives, while still maximizing precision for national-level estimates, the NHIS "base" sample (i.e., the sample with no reduction/augmentation) usually has been allocated proportional to population size within each state. The current base sample allocation is close to being proportional to state population size. A small amount of undersampling was done in the most populous states to increase the sample sizes in the least populous states to a projected 250 completed interviews annually. (For convenience, any future reference in this paper to "states" is a reference to the fifty states and the District of Columbia.)

3. New Features of the 2016 NHIS Sample Design

Sample Frame Change: The NHIS sample that is selected for interviewing is a sample of addresses. The NHIS sample frame of addresses is a list of addresses within the geographic areas selected into sample. NHIS sample designs from 1985 to 2015 used field listing to develop most of the sample frame of addresses. This was feasible because NCHS shared the cost of field listing with other federal agencies that sponsor demographic surveys (e.g., The Current Population Survey) conducted by the Census Bureau. This was not feasible for the 2016 NHIS sample design because the other federal agencies that sponsor demographic surveys conducted by the Census Bureau decided to discontinue using field listing nationwide. In 2016 and beyond, field listing is only a small part of the NHIS sample frame development process. Instead, in most geographic areas, NCHS is using addresses purchased from a commercial vendor, Marketing Systems Group (MSG).

Flexibility to accommodate changes in annual sample sizes: A main goal of the 2016 NHIS sample design planning process was to have flexibility to increase or decrease annual sample sizes, and/or flexibility to shift annual sample sizes on a state-by-state level. In order to maintain year-to-year stability while allowing for the option to shift annual sample sizes on a state-by-state level, the "base" NHIS

sample, which as mentioned above is the anticipated annual sample size if there are no sample reductions or augmentations, consists of two parts. One part (~70% of the total) will remain the same from year to year, with state-level sample allocation proportional to state population size. The other part (~30% of the total) will be allowed to change on a state-by-state level from year to year during the sample design period, if directed by NCHS leadership. The 70/30 split provides a large degree of stability, a feature of all previous NHIS sample designs.

The next two sections describe in more detail how the flexibility feature was implemented.

4. NHIS's Conceptual Sampling Structure

The conceptual structure of the NHIS survey sample is delineated in advance for the entire planned ten-year sample design period, plus a contingency of two additional years. The conceptual structure is then filled with actual sample addresses a few months prior to interviewing in a given month. The conceptual structure allows the Census Bureau to do advance planning for logistics such as hiring and assignment of survey interviewers.

The conceptual sample structure planning process considered the time length (10 plus 2 years), interviewer workload (~100 sample addresses annually), and a doubling factor as a contingency for a possible future reinstatement of oversampling race and ethnicity groups such as black persons, Hispanic persons, and/or Asian persons, which was discontinued in 2016. The combination of these factors results in the conceptual sampling process selecting groups of approximately 2500 geographically clustered addresses ("address groups") into sample. Each sampled address group is later partitioned into pieces that are then assigned to individual years of the sample design period.

Within each state, the counties (or county equivalents) were grouped into geographic areas consisting of one or more counties. Each county is assigned to exactly one geographic area. Almost without exception, the counties in the multi-county geographic areas are contiguous. In some states, the geographic areas were divided into two groups, roughly along urban/rural lines. For states with two groups, the geographic areas in the groups were designated as "Type A" (~urban)/"Type B" (~rural). For the remaining states, the geographic areas in each were all designated as either Type A or Type B.

Each geographic area was assigned a measure of size: the count of 2010 Census housing units. Using this measure of size, an integer number of address groups were defined within each geographic area.

5. Implementing the Sample Flexibility Feature

The first step to implement the flexibility feature was the selection of a very large initial conceptual sample, called the "super sample", in each of the states. The super sample contains many more address groups than what would be needed for a typical NHIS sample size during the previous history of the survey. The size of the super sample was large enough to accommodate any possible sample augmentation/expansion in any state, relative to the historic NHIS sample allocations since the inception of the survey.

The super sample was selected independently within each state. For states with two groups of geographic areas, the sampling was done independently within each group.

The sampling departed from the historic method of designating the geographic areas as primary sampling units (PSU), selecting a sample of PSUs, then selecting address groups within each of the sampled PSUs. Instead, within each state, the conceptual address groups were sorted geographically across the collection of geographic areas, and a systematic sample was selected. In states with Type A and Type B areas, the process occurred separately within each area. The locations of the sampled address groups determined which geographic areas were part of the super sample.

The second step to implement the flexibility feature was to form a nested sequence of subsamples of the super sample by the assignment of entry orders. The entry orders are numbers 1, 2, … that define a nested sequence of subsamples of the super sample in each state/geographic area type. An entry order sequence was assigned within each state/geographic area type. Once it was determined how much sample was to be allocated in a given state/geographic area type for a given year, entry orders 1, 2, … were taken into the final annual sample until the total number of addresses in the subsample met the target allocation.

The departure from the traditional method of selecting PSUs/selecting sample within PSUs provides the desired flexibility, allowing changes in sample sizes that can be implemented in a manner that retains stability in the sampling weights.

NCHS began the sampling process by selecting the super sample in states with Type B areas and assigning entry orders. NCHS and the Census Bureau worked together to develop the specifications for selecting the super sample in states with Type A areas. The Census Bureau then selected the Type A super sample, and later assigned entry orders within that part of the super sample.

6. Enhancements of the Entry Order Algorithm

The algorithm originally used to assign entry orders was designed for use within a state, for a group of geographic areas. The entry order sequences govern how the sample increases or decreases within each state/geographic area type, but they do not by themselves govern how a sample size increase/decrease is to be

implemented across the entire NHIS sample. Research to generalize the entry order concept began in 2017.

The first generalization was to reformulate the algorithm to be applied across the 50 U.S. states, and examine the result of running the algorithm 435 times using 2010 Census population counts at the state level. The result was identical to the current U.S. House of Representatives apportionment to the 50 states. This indicates that the algorithm could be a suitable replacement for Hill's Method that does not lead to the paradoxes of historic apportionment methods (see, e.g., documentation available at the Mathematical Association of America website, https://www.maa.org/press/periodicals/convergence/apportioning-representatives-in-the-united-states-congress-paradoxes-of-apportionment).

The second generalization was to reformulate the algorithm to be applied across the entire United States (50 states and the District of Columbia), and take account of different sample correspondences in different states. First, there is some state-to-state variability in the average size of the individual annual address groups in sample. Second, in states with both Type A and Type B units, NCHS and the Census Bureau has created a correspondence between the samples in the two groups. If there is a change in the sample allocation to these states, there is increase/decrease in sample in both Type A and Type B areas, using the correspondence code, to keep the state-level base weights constant as part of the change.

The second generalization was used to determine the stable part of the NHIS base sample (see Moriarity and Parsons (2015)), accounting for the variability described in the preceding paragraph. The algorithm output was modified slightly to decrease the variability of the base weights associated with a hypothetical base sample of this size.

The second generalization also was used to determine the location of the areas to be included in the 2019 sample augmentation, which was specified to be done proportional to state population. As with the determination of the stable part of the NHIS base sample, the algorithm output was modified slightly to decrease the variability of the base weights. After these steps were completed, the original entry order sequences were then used in the states/geographic area types that were to receive sample augmentation.

The implementation of the flexibility feature has been successful. It provides the capability of implementing sample size changes (with some lead time required) in a way that weight stability is maintained. This feature was used for the sample augmentations in 2016, 2018, and 2019.

The implementation of the flexibility feature has a cost, however; increased complexity. A number of new parameters are required to keep track of the various

pieces of the super sample, the part of the super sample that is the annual sample, etc.

7. Sample Frame Change

The sample frame change was a major departure from the methodology used to identify the NHIS sample addresses during the previous 30 years. A small number of the survey sample addresses are still developed using field listing, primarily in rural areas where the mailing addresses are not city-style addresses, i.e., house number/street name form.

The first step in the sample frame change process was to obtain one or more new sources of sample addresses for the NHIS. A contract was put into place between the Census Bureau and Marketing Systems Group (MSG) to achieve this.

The contract terms required MSG to provide to the Census Bureau a copy of their entire list of U.S. residential addresses. The Census Bureau and NCHS then worked together to select samples from the MSG list, maintaining the confidentiality of the NHIS sample addresses. The contract terms also required MSG to create and maintain an ID variable for each address, to be used as part of periodic update deliveries (e.g., new addresses, changes in existing addresses, deletion of addresses that were no longer residential addresses).

The contract terms specified that MSG mark each delivered residential address as "sufficient quality" or "insufficient quality". The criteria for sufficient quality included the specification that a given address was a city-style address (e.g., house number, street name) and that "the address is likely to be the residence for a member or members of the NHIS target population".

NCHS and the Census Bureau needed to decide where to use the MSG list as the NHIS address source, rather than listing. One possibility would have been to compare aggregate totals of MSG city-style addresses to a reference such as 2010 Census counts. A pitfall of such an approach would be that aggregate counts can mask errors of omission (addresses that exist that are not on the list) and errors of faulty inclusion (e.g., duplicate addresses, addresses with incorrect geocodes).

Census Bureau personnel were allowed to access the results of a record linkage at the Census Bureau between a version of the MSG list and the Census Bureau's Master Address File and share aggregate results with NCHS. This allowed a more precise determination of where to use the MSG list than comparison of aggregate totals.

The MSG list is the source of approximately 86% of the addresses in the NHIS base sample; the remaining addresses come from field listing.

The Census Bureau conducted research that indicated that the inclusion of a portion of the "no-stat" addresses (addresses currently not receiving mail delivery) on the MSG list into the pool of potential NHIS sample addresses resulted in substantial coverage improvement. The Census Bureau then conducted research to develop "filter" algorithms to be applied to the MSG list to determine which addresses were considered to be eligible to be selected as NHIS sample addresses. There were two filter algorithms developed; one for the "no-stat" addresses, one for the remaining addresses. The final filter algorithms gave some different outcomes for eligibility for inclusion in the survey than the initial eligibility criteria specified in the contract with MSG and applied by MSG to create an eligibility code prior to delivery. For example, MSG had marked all "no-stat" addresses to be "insufficient quality".

The Census Bureau and NCHS worked together to develop the sampling mechanism for the "new growth" addresses, i.e., the new addresses provided by MSG in the periodic updates. The new growth samples are selected from the geographic areas in a given state that already are in sample, at the same sampling rate as the existing sample in the state.

The sample frame change required considerable research to be carried out, and new systems to be created, after the contract was in place with MSG and the Census Bureau received the initial data delivery.

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.

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.

Moriarity C, Parsons V. 2016 Sample Redesign of the National Health Interview Survey. 2015 Proceedings of the Joint Statistical Meetings, 2444-2451.

Moriarity C, Parsons V. Nested Subsamples: a Method For Achieving Flexibility in Annual Sample Sizes For a Continuous Multiyear Survey. 2018 Proceedings of the Joint Statistical Meetings, 2185-2191.

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, 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.