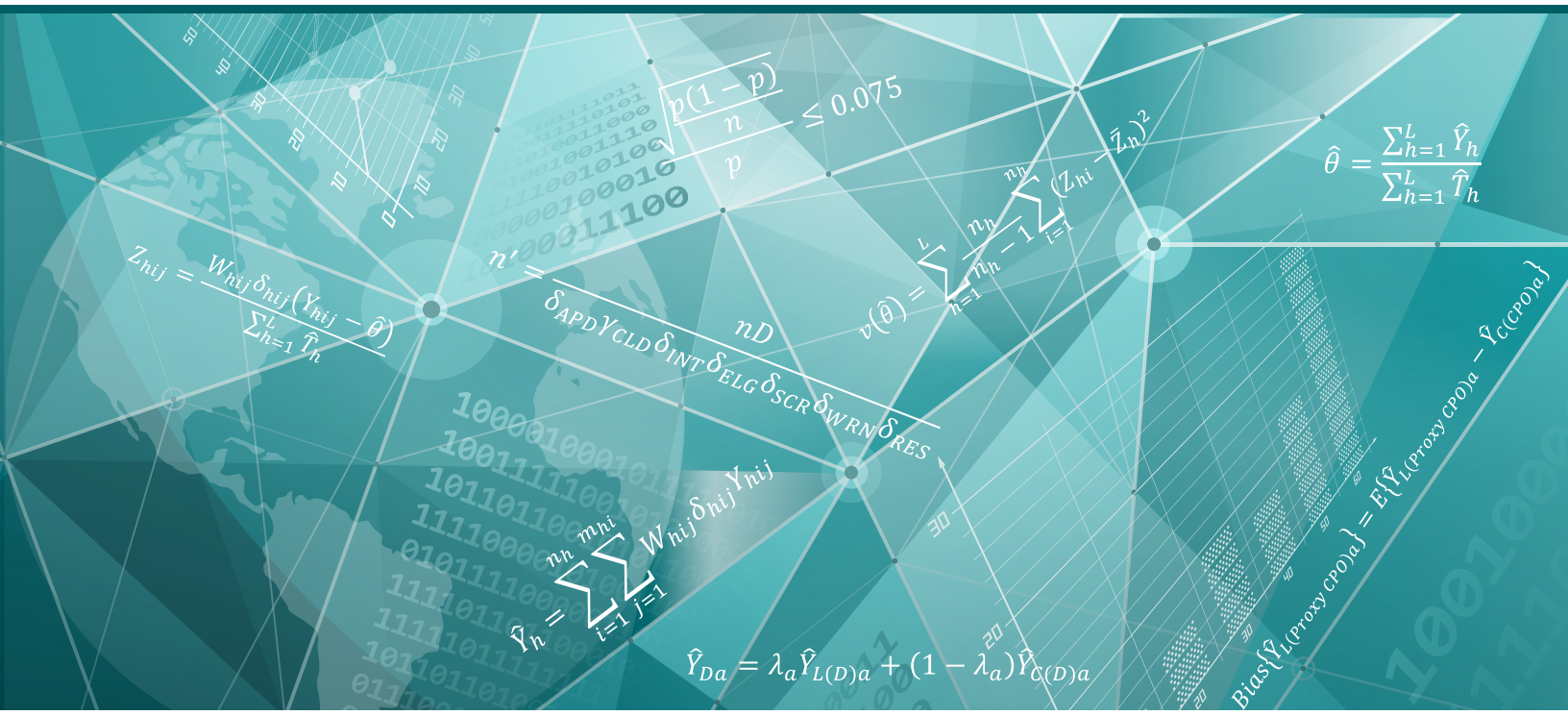
NATIONAL CENTER FOR HEALTH STATISTICS

Vital and Health Statistics

Series 2, Number 191

June 2022



Sample Design and Estimation Structures for the National Health Interview Survey, 2016–2025

Data Evaluation and Methods Research

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics

NCHS reports can be downloaded from: https://www.cdc.gov/nchs/products/index.htm.

Suggested citation

Moriarity C, Parsons VL, Jonas K, Schar BG, Bose J, Bramlett MD. Sample design and estimation structures for the National Health Interview Survey, 2016–2025. National Center for Health Statistics. Vital Health Stat 2(191). 2022. DOI: https://dx.doi.org/10.15620/cdc:115394.

For sale by the U.S. Government Publishing Office
Superintendent of Documents
Mail Stop: SSOP
Washington, DC 20401–0001
Printed on acid-free paper.

NATIONAL CENTER FOR HEALTH STATISTICS

Vital and Health Statistics

Series 2, Number 191 June 2022

Sample Design and Estimation Structures for the National Health Interview Survey, 2016–2025

Data Evaluation and Methods Research

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics

Hyattsville, Maryland
June 2022

Contents

Acknowledgments

The authors thank Rosemary Byrne, Aliza Kwiat, Alejandro Seguro, and Will Waldron from the Census Bureau, and Stephen Blumberg, Morgan Earp, Aaron Maitland, and Jennifer Parker from the National Center for Health Statistics (NCHS) for their contributions to this report. This report was edited and produced by the NCHS Office of Information Services, Information Design and Publishing Staff: Jen Hurlburt edited the report and typesetting was done by Ebony Davis.

Sample Design and Estimation Structures for the National Health Interview Survey, 2016–2025

by Chris Moriarity, Ph.D., and Van L. Parsons, Ph.D., National Center for Health Statistics; Kimball Jonas, Ph.D., and Bryan G. Schar, M.S., U.S. Census Bureau; and Jonaki Bose, M.Sc., and Matthew D. Bramlett, Ph.D., National Center for Health Statistics

Abstract

Background

The National Health Interview Survey (NHIS) is one of the major surveys sponsored by the Centers for Disease Control and Prevention's National Center for Health Statistics (NCHS). NHIS conducts household interviews throughout the United States to collect health-related information from the U.S. civilian noninstitutionalized population. After the 2010 decennial census, NCHS' Division of Research and Methodology and the Division of Health Interview Statistics collaborated with the U.S. Census Bureau (through an interagency agreement that includes the collection of data in the field) on the NHIS sample redesign.

The 2016–2025 NHIS sample design uses cost-effective complex sampling techniques including stratification, clustering, and differential sampling rates to achieve several objectives, such as transitioning away from a fully enumerated sample list frame and allowing for flexible annual sample sizes. This report describes these methods.

Objectives

This report presents operating characteristics of the NHIS 2016–2025 sample design. The general sampling structure is presented, along with a discussion of weighting and variance estimation techniques primarily for 2016–2018. This report is organized into four major sections. The first section presents a general overview of NHIS and its sample design. The second section describes the redesign process, updates for 2016–2025, and includes general frame and sample design considerations. The third section provides a more detailed description of the sample design and how the sample was selected. The last two sections present a description of the estimators used in NHIS for analyzing and summarizing survey results. Documentation for subsequent changes to the sampling and weighting procedures is available on the NCHS website as separate reports and through each year's survey description document. This report is intended for general users of NHIS data.

Keywords: sampling • weighting • nonresponse adjustment • variance estimation

Overview of the National Health Interview Survey

The National Health Interview Survey (NHIS) is the nation's primary source of general health information for the resident civilian noninstitutionalized population. NHIS is administered by the Centers for Disease Control and Prevention's (CDC) National Center for Health Statistics (NCHS). Following NCHS specifications, the U.S. Census Bureau, through an interagency agreement, participates in the planning, including sampling and sampling frame construction, and collection of data for NHIS. NHIS has continuously collected data since July 1957.

Data from NHIS provide estimates of health, health care utilization and access, and health-related behaviors for the U.S. resident civilian noninstitutionalized population. Summary reports and reports on special topics are prepared by the NCHS Division of Health Interview Statistics (DHIS) for publication in NCHS publications, journals, and elsewhere. The NHIS Early Release Program, begun in 2001, provides several online analytic reports during each data collection year. Further information on the Early Release Program is available from: https://www.cdc.gov/nchs/nhis/releases.htm.

NCHS makes Early Release data and in-house analytic files available through the NCHS Research Data Center. Additionally, NCHS releases NHIS public-use microdata files annually at: https://www.cdc.gov/nchs/nhis.htm.

NHIS public-use data files are a subset of the in-house analytic files, and they include variables that have been recoded or otherwise modified to minimize disclosure risks (that is, identification of individual survey participants). Variables identifying specific geographic locations smaller than census regions are withheld from the public-use files to protect participant confidentiality. Other variables are recoded to provide additional protection of participant confidentiality. NCHS also includes variance estimation information (pseudo-stratum, pseudo-primary sampling unit [PSU]) for the public-use microdata files to allow data users to compute direct estimates of sampling errors that are consistent with the complex survey design of NHIS. Since 1980, public-use microdata files contain variance estimation information. NCHS provides auxiliary files with variance estimation information that can be linked to pre-1980 files. NCHS also provides guidance on how to estimate variances for users of software such as SUDAAN, Stata, SAS survey procedures, R, and SPSS.

Sample Design and Basic Subsamples

NHIS is based on a stratified, multistage sample design. A new sample design is implemented after each decennial census. The NHIS sample design implemented at the beginning of 2016 is referred to in this report as the "2016–2025 NHIS," because the initial sampling steps for this sample design allowed for a 10-year design period and up to 2 additional years as a contingency. The current sample design can end before 2025 if the next sample design is ready to be implemented sooner, or it can end as late as 2027. Unless otherwise noted, this report documents the sample design as initially implemented in 2016; aspects of the design may be modified over time, and major changes, if any, will be described in subsequent reports. The 2016–2025 sample redesign is independent of the 2019 questionnaire redesign. Information on the 2019 questionnaire redesign is available from: https://www.cdc.gov/nchs/nhis/2019_quest_redesign.htm.

The 2016–2025 NHIS has been designed to produce estimates at the national level, as well as estimates by the four census regions, and by metropolitan status within the four census regions. Although the 2016–2025 survey draws samples from all of the states and the District of Columbia, it is not designed to produce precise state-level estimates for every state annually.

For the 2016–2025 survey design, the initial NHIS sample was partitioned into a number of subsamples. First, the sample was partitioned into annual subsamples that are assigned for data collection each year from 2016 through 2027. The NHIS annual sample was further assigned to four nationally representative calendar quarters. As of 2011, the smallest time interval for assignment that makes up a representative probability sample changed from 1 week to 1 month, and the definition of "calendar quarter" changed from 13 weeks to 3 months.

Assignment of the NHIS sample to monthly subsamples has operational and administrative benefits because it provides a continuous workload for the field staff, which ensures workforce stability and avoids seasonal effects of data collected during only part of the year. Having nationally representative monthly samples also allows for the production of estimates for time periods smaller than a complete year as long as sample sizes are large enough to meet precision requirements.

Starting with the 1985–1994 survey design, the annual NHIS samples have been partitioned into four annual panels (subsamples), each having about the same number of sample households and conceptually similar statistical features. Each panel can be considered a nationally representative probability sample of the U.S. population. ("Panels" and "calendar quarters" are not the same; the set of four panels is a partition of the annual survey samples that encompasses the four calendar quarters.) The panels have several realized and anticipated uses, and they are a mechanism for NHIS to provide nonoverlapping samples for reuse as sampling frames for other studies. A portion of the 1996 Medical Expenditure Panel Survey (MEPS), sponsored by the Agency for Healthcare Research and Quality (AHRQ), used a portion of the 1995 NHIS sample as a sampling frame. Later years of the noninstitutional component of MEPS have used part of the previous year's NHIS sample (usually two of the four panels) as a sampling frame. For example, the noninstitutional component of the 2017 MEPS sample is based on a portion of the 2016 NHIS sample. Additionally, panels provide a mechanism to make large cuts in the survey sample size if, for example, sufficient funds are not available to support full-scale data collection. Therefore, after the 2016–2025 survey design sample was selected, the new sample was partitioned into a new set of four panels.

Sample Redesign

Since it began in July 1957, the NHIS sample has been redesigned after each decennial census of the population to accommodate changes in survey requirements, the population, and its distribution (1–7).

The primary features of the 2016–2025 NHIS sample design implemented in January 2016 are:

- *Use of a unit or area sampling frame*—Most Census surveys (for example, the Decennial Census and the American Community Survey [ACS]), as well as most of the other sponsored national demographic surveys conducted by the Census Bureau (for example, Current Population Survey and National Crime Victimization Survey) use a sampling frame called the Master Address File (MAF). The MAF is created and maintained by the Census Bureau, and is based on an area frame where field staff list all the households in an area. However, Census Bureau confidentiality constraints do not permit the release of MAF addresses outside the Census Bureau. NCHS requires

the release of the sample frame so that follow-up surveys like MEPS are feasible. Therefore, the Census Bureau independently develops a separate sampling frame for NHIS that permits NCHS to receive the survey sample addresses from the Census Bureau. NCHS can then use the sample addresses for additional data collections and provide addresses to AHRQ for MEPS. Previous NHIS sampling frames were based on a separate non-MAF area frame that NCHS and other survey sponsors jointly funded. However, the 2016–2025 NHIS sample frame is based primarily on a vendor-supplied address list referred to as the unit frame. The NHIS vendor-based sample frame is supplemented by an area frame in areas where the unit frame does not have high enough coverage rates (that is, does not include all households).

- *State stratification*—The sampling strata are defined within each state (that is, they do not cross state boundaries).
- *Flexibility to accommodate variations in the sample design composition*—The sample design implemented in 2016 incorporates flexibility that allows for variations in the sample design composition (for example, in response to supplemental funding being available for sample augmentation). This flexibility was achieved by first selecting a very large nationally representative sample, (a super sample) and then determining the order in which an additional preselected sample could be included (or removed) and still preserve the representativeness in the augmented sample.

An overview of sample and frame considerations for the NHIS 2016–2025 sample design follows in the section "Redesigning the 2016–2025 NHIS Sample Based on 2010 Census Data." Additional detail on the NHIS 2016–2025 sample design is included in the section "2016–2025 NHIS Sample Design." The sections on weighting and variance estimation provide a detailed description of estimation procedures for the NHIS 2016–2025 sample design. A glossary is included in the Appendix.

Redesigning the 2016–2025 NHIS Sample Based on 2010 Census Data

Objectives

The objectives for the NHIS sample redesign are:

- Continue producing descriptive statistics and monitoring trends of health and health-related factors for the civilian noninstitutionalized population of the United States.
- Implement a sample design for 2016–2025 with minimal additional annual costs.
- Transition away from the all-area frame concept used since 1985, due to the cost of developing and maintaining a national area frame.
- Incorporate more flexibility in changing the annual sample size or the allocation of the annual sample size than was possible in previous sample designs.

These broad objectives were used as the primary criteria for redesigning the 2016–2025 NHIS sample design. The 2016–2025 sample redesign is separate from the 2019 questionnaire redesign.

Based on the objectives, two major issues were considered: 1) identifying more cost-efficient methods for developing a sampling frame with minimal losses in coverage; and 2) developing a flexible sample design where the sample could be increased or decreased relatively easily while maintaining representativeness.

Frame Considerations

As mentioned earlier, Census Bureau surveys use the MAF as a household survey sampling frame, while survey sponsors jointly sponsored an alternate area-based sampling frame that permitted them access to sample addresses. However, during the redesign of the demographic surveys' sample based on the 2010 census data, the sponsors of the non-NCHS Census Bureau demographic household surveys transitioned away from the development of a separate area frame, relying instead on the MAF as the sample address source because it was more cost efficient. Obtaining new growth and sample updates from MAF processing was much more efficient than the previous designs that included costly area and permit frame listings. A permit frame is a portion of the NHIS sample frame from 1985 through 2015 consisting of residential building permits.

Previously, NCHS shared area frame infrastructure development costs with other federal agencies; however, given their move away from sponsoring the alternate area frame concept, NCHS had to either entirely fund the separate area frame or find other solutions. NCHS explored other options for obtaining the NHIS sample addresses for the 2016–2025 sample frame instead of solely funding and maintaining the area frame development infrastructure.

NCHS first sought access to the United States Postal Service's (USPS) Delivery Sequence File (DSF) via the Census Bureau for use as the primary source of sample addresses for the 2016–2025 NHIS. This was within the scope of Section 412 of Title 39 USC, passed after the 1990 census, which states in part: "The Postal Service shall provide...address information...as may be determined...to be appropriate for any census or survey being conducted by the Bureau of the Census." However, due to time constraints, NCHS used another primary source of sample addresses for NHIS.

The Census Bureau contracted with the Marketing Systems Group (MSG) in late 2013. MSG is a commercial information reseller that provides frames and samples for surveys. Under the contract, MSG was to provide the Census Bureau with data comprised of all the addresses, and selected information

associated with these addresses, based on USPS' Address Management System (AMS) database for all 50 states and the District of Columbia. The USPS AMS database includes all of USPS' mailing addresses. MSG also agreed to provide semiannual updates of the address data to keep the NHIS address-based frame (the unit frame) current. Each address has a unique ID across all deliveries.

The MSG file includes both addresses and information about the addresses. Examples of the information provided include:

- Census Bureau geocodes (state Federal Information Processing Standards [FIPS] codes, county FIPS codes, Census Bureau tracts, Census Bureau blocks)
- Residential or business address classification information
- Ways to identify post office (PO) boxes, highway contract routes and box numbers, rural routes and box numbers, and general delivery addresses
- Whether an address is a No-Stat address (see definition in Glossary)
- Whether an address is a PO box throwback address (see definition in Glossary)
- Whether an address is a drop point (see definition in Glossary)
- Whether an address has been vacant

In addition, the data provided indicate if each address in the file is of sufficient or insufficient quality, which informs whether the address is included on the NHIS unit frame. The criteria for sufficient quality include all of the following:

- The address is for habitable living quarters.
- The living quarters are likely to be the residence for a member or members of the civilian noninstitutionalized population.
- The address has been geocoded to at least the state and county level.
- Field staff must be able to uniquely identify and find the living quarters associated with the address.

In 2013 and 2014, the Census Bureau evaluated the MSG address file in these areas:

- Coverage: Where was the coverage of valid NHIS living quarters high enough to allow the use of the MSG address file as the source of sample, and, on the other hand, what areas still required listings to be used as the source of sample?
- Filtering: Was there a possibility of developing an algorithm for filtering MSG addresses (beyond the initial filtering applied by MSG) that would accept as many valid addresses as possible as potential sample addresses, while excluding as many invalid addresses as possible, to reach an overall result of maximum coverage of the target population with minimal incorrect inclusions?
- Geocoding: What was the quality of the geocodes on the MSG address file? Accurate geocodes help interviewers find the sample addresses and place the address in the correct Census Bureau tabulation block for sampling purposes.
- Updating: How to incorporate the updated MSG address files into the NHIS unit frame and how to sample from the new addresses identified in the updates.
- College population: How to capture information about people living in college housing, a unique and important subpopulation of NHIS.

These evaluations are further described in the following sections.

Coverage

Vendors of address-based frames, such as MSG, periodically receive updated information from the USPS Delivery Sequence Files. This is the primary source of updates to vendor-provided frames as well as the Census Bureau's MAF. However, the MAF also contains additional information from previous censuses that canvassed the entire country and updates from local government partnerships. These sources of information on the MAF provide a much broader coverage of noncity type addresses, group quarters, and other locations where a household is found but their mailing address is elsewhere. For this reason, the MAF-based surveys currently monitor coverage but do not require area listing, whereas household surveys using the vendor files traditionally supplement counties with insufficient coverage with an area frame where field staff list the households in the county to create a sampling frame for that county.

In 2014, the Census Bureau linked an October 2013 version of the MSG address file provided for research purposes to a data set of valid addresses from the July 2013 MAF. Both record linkage results and a comparison of aggregate counts were evaluated. The record linkage results are theoretically more precise than a comparison of aggregate counts, which can mask omission and duplication errors. The proportion of MSG addresses that matched to valid MAF housing units (HUs) was used to determine coverage adequacy. Valid MAF addresses are those that have passed a filter for use in current demographic surveys. The record linkage results allowed the Census Bureau and NCHS to make a well-informed decision based on where the MSG address file's coverage was considered to be of sufficient quality for use as the only source of the NHIS sample.

Whether or not to use the MSG address file as the only source of sample was decided at the county level because this level of geography stays relatively stable over time. About one-half of all U.S. counties, containing about 85% of the 2010 census HUs, were determined to have sufficient coverage in the MSG address file. If the coverage rate of a county was 85% or greater, then an MSG-based unit frame was used as the only source of sample for that county. The 85% cutoff

was used because the 2006 Office of Management and Budget's Standards and Guidelines for Statistical Surveys, Guideline 2.1.3 states, "Coverage rates in excess of 95% overall and for each major stratum are desirable. If coverage rates fall below 85%, conduct an evaluation of the potential bias" (8). Group quarters on the MAF were not considered in this analysis.

A coverage cutoff of 90% for inclusion in the NHIS unit frame was also considered. However, research indicated that the corresponding amount of listings would be beyond the amount budgeted for area listings by NCHS.

For those counties where the coverage of the MSG address file was classified as too low to use it as the only source of sample, an area frame for the county was created from field listings. These area frames (also referred to as the area frame) were used as the only source of sample for these counties. If it were later determined that the coverage of the MSG address data had substantially improved in an area frame county, then MSG updates would be used instead of additional field listings for that county.

Filtering

Census Bureau staff conducted research to explore an alternative to the sufficient or insufficient filter MSG provides as specified in the contract. This revised filter resulted in fewer erroneous inclusions and exclusions on the unit frame than the filter provided by MSG. The data set of valid addresses from the July 2013 MAF was used as the gold standard to compare these two filters. The research also indicated that the inclusion of a subset of the USPS No-Stat addresses in the MSG address file resulted in substantial coverage gains (the filter provided by MSG excluded all No-Stat addresses). As a result, a separate filter was defined for the No-Stat addresses. This process is described as follows:

An initial filtering step removes several types of addresses that were either out of scope to NHIS or would not allow for in-person interviews of the household associated with the address:

- Addresses without a house number or street name
- Addresses that are PO boxes, rural routes and box numbers, or highway contract routes and box numbers
- Addresses for government buildings and addresses assigned unique ZIP Codes (Unique ZIP Codes are used to service large nonresidential organizations, and thus do not include the NHIS target population.)

For records that pass the initial filtering step, a second filter, with different criteria for No-Stat and non-No-Stat addresses, is applied. Any address that meets any of the criteria passes through the filter. The secondary filtering criteria for non-No-Stat addresses identify if a record is:

- A residential address and either a street or a high-rise record type
- A primarily residential address combined with a business, and a drop point in a rural route or highway delivery route
- A nondrop point address, regardless of residential or business usage, that is a street or a high-rise record type
- A primary business address with some residential mail activity that is not a drop point, and is a street record type, and mail is delivered at the curb, to a mail receptacle in a cluster box, or to a centralized unit (see definition in Glossary)

For No-Stat addresses that pass the initial filtering step, the subsequent filtering criteria for inclusion in the sample frame are:

- Residential address or a primarily residential address with some business mail activity, and
- Either a street or a high-rise record type, and is either:
  - a PO box throwback address, or
  - not a PO box throwback address and the geocode precision field provided by MSG is at the address or ZIP + 4 centroid level.

The final filters purposely allowed a higher number of potentially erroneous inclusions as the price of obtaining potentially higher coverage gain of the target population than if only the MSG filter was used. While the final filter did a more effective job of filtering out some potentially erroneous exclusions than the MSG filter, allowing some of the No-Stat addresses to be included could result in a trade-off of a higher overall number of potentially erroneous inclusions.

Geocoding

The Census Bureau defines geocoding as the process of assigning census block codes to the addresses in an address list or address-based sampling frame. Assigning correct block codes is important for several reasons:

- Block information can be important in locating an address in the field for personal interview.
- Block summary data from the decennial census or ACS can be linked to geocoded HUs in the sampling frame for efficient sorting of the universe for sampling.
- The block code is the gateway to other census geography; if the block for an address is known, its place, county subdivision, urban or rural status, urban area, principal city status, etc. can be determined. This information is usually not available for addresses that are not geocoded.
- In a sample design such as NHIS that includes an area block listing component, placing addresses in an incorrect block can lead to coverage error.

The 2010 census tabulation geography was used for the 2016–2025 NHIS sample design. The geocodes retained reflect where the unit was tabulated in the 2010 census. For

any new HUs, the geocodes reflect the 2010 tabulation block (and state, county, and tract) that the HU currently occupies.

While rare, local boundaries may change or HUs can be assigned to different geography between decennial censuses. For instance, in 2002, Broomfield County, Colo. was newly created from portions of four previously existing counties. The HUs that became part of Broomfield County were tabulated in the 2000 census in one of those other four counties.

Therefore, an HU has two sets of geocodes: 1) the tabulation geography geocode (where it was counted in the previous census) and 2) the current geography geocode. For the NHIS sample design, these two sets of geocodes are referred to as the geocode at the time of sampling and the geocode at the time of interviewing. For weighting, sorting, and tabulations, the geocode at the time of sampling is used. For the address information provided to the NHIS interviewer, the geocode at the time of interview (the most recent information) is used.

Geocoding each address on the NHIS frames to the correct Census 2010 tabulation block is attempted. Different processes are used for the unit frame and the area frame geocoding.

Unit frame geocoding

For the unit frame, MSG provides Census Bureau geocodes for each address in their file. These include: state FIPS code, county FIPS code, Census Bureau 2010 tabulation tract, and Census Bureau 2010 tabulation block. MSG obtains its geocodes from publicly available Census Bureau Topologically Integrated Geographic Encoding and Referencing (TIGER)/ Line Shapefiles in addition to information bought from the GPS navigation technology company TomTom. TIGER/ Line Shapefiles contain geographic features such as roads, railroads, rivers, lakes, political boundaries, and census statistical boundaries covering all areas of the United States.

However, when MSG assigns geocodes to addresses, sometimes all that is provided are different levels of ZIP centroids for addresses. Given that these geocodes could sometimes be quite distant from the actual block of the HU (for example, if a ZIP centroid is used), the Census Bureau's Geography Division (GEO) regeocodes the MSG addresses. This allows for a more reliable assignment of the HUs to the same geocodes as used by other surveys using tabulation geography (such as the ACS). This standardization allows for more confidence when comparing block counts across MSG to ACS or other MAF-based surveys.

GEO geocoded all valid addresses in the first production file MSG delivered to the Census Bureau in 2015. For each MSG update file received since then, only addresses flagged by MSG as additions or changes are sent to GEO to be geocoded.

MSG provides an indicator for the precision of their GPS coordinates. This is also used to determine the precision of the geocodes. If a GPS coordinate for a given unit is based on the centroid of the ZIP code for that address, it is inferred that the block code for this address is also imprecise (that is, that the block code determined for this address represents not necessarily the actual block this unit is found in but rather the block containing the ZIP centroid). When GEO assigns geocodes, other address information may be used to assign a more accurate block geocode than MSG had assigned.

The possible geocode precision levels for an address on the unit frame, listed from most accurate to least, are:

- Address range-based – Geocoded by GEO
- Derived from MSG GPS coordinates – Geocoded by GEO
- Address-level GPS coordinates – Geocoded by MSG
- ZIP + 4 Centroid GPS coordinates – Geocoded by MSG
- ZIP + 2 Centroid GPS coordinates – Geocoded by MSG
- ZIP Centroid GPS coordinates – Geocoded by MSG

Area frame geocoding

Field staff who list (or canvas) a block obtain geocode information for units on the area frame. It is assumed that these are the best possible geocodes that can be obtained for an address. Geocodes obtained for addresses on the area frame are only updated when a block needs to be relisted for sampling purposes; this is done if it has been more than 3 years since the block was last listed.

Having the most accurate geocodes for an address on the NHIS Frame is critical for two reasons:

- If an address is in the incorrect block, it could be sampled when it should be or when it should not; and
- NHIS interviewers use the geocode information to help find the addresses they have been assigned.

Updating or New Growth

Every January and July, the Census Bureau receives an updated address file from MSG. Included on this file is a flag that has these values:

- N: There have been no changes to the address since the last update.
- C: There have been significant changes to the address since the last update.
- A: This is the first time an address has appeared on the MSG database.
- D: The address was deleted from the MSG database since the last update.
- G: Minor changes to the latitude and longitude of the address since the last update.

First, this file is examined for any errors or unusual changes, and it is used after it has been determined to be of acceptable quality. Then, the update file is filtered so that only records that are valid on the NHIS unit frame remain. For these valid addresses, this happens:

- Addresses flagged as A or C are sent to GEO to be geocoded. The results are then used to add new records (A) to the database or to modify existing records (C).
- Addresses flagged as D are flagged as invalid on the database. There is no system currently in place to reactivate deleted records.
- No action is taken for records flagged as N.

With every MSG delivery, these flags are reset. Those records that were marked as deleted records in previous deliveries no longer exist in future MSG update files. The process above is repeated semiannually for each update delivery of the MSG file. Addresses first appearing in MSG deliveries after the original January 2015 delivery are called new growth, as they reflect the increase in the housing supply from the January 2015 baseline in areas where NHIS uses the unit frame. For this reason, new growth is a small but continually increasing fraction of the unit frame. Area frame blocks that have been previously listed are relisted if it has been more than 3 years since they have been updated. These blocks are listed between 3 to 6 months before the interview month. The relisting picks up any new growth up to that point in area frame blocks in the sample.

College Population

People living in college housing are of special interest to NHIS because of their unique sociodemographic and health characteristics. Because of the poor coverage of group quarters such as college dormitories on the MSG address frame, an initial effort to capture this subpopulation using a separate sampling frame of college dormitory residents was made.

This college housing frame was developed in 2014 using information from the Integrated Postsecondary Education Data System (IPEDS), which is administered by the National Center for Education Statistics. IPEDS is a system of interrelated surveys conducted annually by the U.S. Department of Education's National Center for Education Statistics that gathers information from every college, university, and technical and vocational school that participates in the federal student financial aid programs and includes data on enrollments, program completions, graduation rates, faculty and staff, finances, institutional prices, and student financial aid.

The use of this separate college dormitory sampling frame was discontinued at the end of 2017 due to low response rates and cost considerations (9). Beginning in 2018, the NHIS questionnaire was modified to enumerate people who live in on-campus college housing. These people are now captured at the living quarters where they live while not attending college. A weakness of this new method is that it does not capture foreign college students or students whose only place of residence is on-campus housing.

Sample Design Considerations

Flexibility

One of the goals of the sample redesign was to increase flexibility to allow for changes in sample sizes and sample allocations (10). A limitation of the 2006–2015 NHIS sample design became apparent when funding for substantial sample augmentation was provided after the passage of the Patient Protection and Affordable Care Act in 2010. The sample was not designed to accommodate rapid large scale increases in sample sizes, and the final augmentation was implemented using various methods, which introduced additional variation to the sampling weights, which increased standard errors.

An early decision was to consider the annual base (or core) sample (the expected sample size if no sample cuts or augmentations occur) of about 35,000 completed household interviews to be partitioned into two parts: a stable sample of size 25,000, allocated proportional to state population size, and the remaining 10,000 sample, to be more flexible. This would permit an efficient reduction in the sample should it be needed. This partition would also help to ensure year-to-year stability in sample estimates from the survey data by ensuring a minimum national sample of 25,000 per year.

The ability to expand the sample as needed was built into the sample design. Instead of only selecting the core sample, an initial larger sample (or super sample) was selected (7). Parts of the sample were designated as the core sample, and the remaining parts of the sample were classified so that representativeness would be maintained when sample size or sample allocation changes were made.

The NHIS state allocations were intended to be proportional to state population sizes. However, NCHS decided in late 2014 on a national sample allocation that was slightly different from allocating a sample proportional to state population size. The sample allocation chosen allocated an additional sample to the 10 least populous states and the District of Columbia to yield about 250 completed household interviews annually in these areas. The sample allocation in several of the most populous states was decreased to keep the overall core sample size the same.

After all of the major design features had been established, the specific parameters for the NHIS design were determined. This process is described in the following section.

2016–2025 NHIS Sample Design

The current NHIS sample design was implemented in 2016 and is scheduled to be used through 2025. NHIS is redesigned each decade to align the sample design with demographic shifts in the U.S. population between two decennial censuses. This also creates an opportunity for further refinements based on new or modified objectives.

The basic design objectives for the 2016–2025 design are similar to those for the 2006–2015 NHIS (1). One important difference is the discontinuation of oversampling of Black, Hispanic, and Asian people.

As in most previous designs, all 50 states and the District of Columbia were sampled (from here on, the District of Columbia is referred to as a state for sampling purposes).

NHIS has a stratified sampling plan. Because data are collected via face-to-face interviewing, geographic clustering is needed to control the cost of field operations. The sampling consists of the selection of address clusters that are within well-defined geographic areas that do not cross state boundaries. These geographic areas are defined as counties or groups of counties (similar to PSUs in previous NHIS sample designs) that are almost always contiguous. (The term county includes county equivalents such as parishes in Louisiana and independent cities in Virginia, Maryland, Missouri, and Nevada.) The 2013 metropolitan statistical area (MSA) definitions, which are at the county level for the entire United States, were used to define some geographic areas. More information on the definition of MSAs is available from: https://www.census.gov/programs-surveys/metro-micro.html.

Within each state, each county was assigned to a single geographic area. Each geographic area was assigned a measure of size (MOS), which is defined as the number of 2010 census HUs within that geographic area. Within each state, the geographic areas were assigned to either one or two strata. If two strata, the stratifications were defined by a general population size and urban or rural classification, labeled Type A and Type B, respectively. If a given state had only one stratum, the label could be either Type A or Type B, depending on the state's population size and urban or rural characteristics.

Instead of using a traditional sample frame structure with a listing of the sampling units (for example, the actual HUs), the redesigned NHIS sample started with address clusters based on 2010 census housing counts. These expected address clusters were defined within each geographic area, based on interviewer workload, the number of years in the sample design period, and the contingency to implement oversampling of specific population subgroups should it be needed in the future. The annual interviewer workload was set equal to about 100 addresses. A 10-year period was assumed in the sample design, and a reserve contingency of 2 years was planned for. A doubling factor was employed for any potential oversampling contingency. The product of these numbers is 100 • 12 • 2 = 2,400 addresses. The conceptual address clusters were taken to be a value around 2,400, with some state-to-state variation in the cluster size.

The address clusters were ordered across all the geographic areas in a given single-stratum state or within-state stratum. A large systematic sample (referred to as the super sample) of address clusters was selected at a uniform sampling rate within each state or stratum combination. The locations of the sampled address clusters determined which geographic areas had sample areas in the super sample. In the states with Type A geographic areas, all the Type A geographic areas had one or more address clusters in the super sample. In the states with Type B geographic areas, usually only some of the geographic areas had one or more address clusters in the super sample.

Geographic Area Definition and Stratification

The three tasks that led to the geographic area definition and selection were: 1) forming a county-based partition of the United States into a sampling universe of geographic areas, 2) stratifying the universe of geographic areas, and 3) choosing a sampling procedure. Given that the 2016–2025 NHIS design objectives were somewhat consistent with those of the 2006–2015 NHIS design, the existing 2006–2015 defined universe of NHIS PSUs offered a reasonable starting point for creating geographic areas. Therefore, the 2016–2025 geographic areas were created by fine-tuning the 2006–2015 NHIS PSUs. This saved substantial resources.

Some additional points regarding geographic area definitions include:

- Geographic areas are single or combined counties (or equivalents), almost always contiguous, as defined by the Census Bureau's 2010 FIPS definitions. As of April 2010, the total number of counties was 3,143.
- The component counties in multicounty geographic areas do not cross state boundaries.
- The component counties within each geographic area have the 2013 MSA status as defined by the Office of Management and Budget in February 2013 (https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/bulletins/2013/b13-01.pdf).
- For the 2016–2025 defined geographic areas, most non-MSA geographic areas correspond to the PSU definition that was used during the 2006–2015 design.
- MSA PSUs from the 2006–2015 design were subject to change due to area growth in surrounding counties. In many such cases, a 2016–2025 defined geographic area is the 2006–2015 defined PSU plus the expansion.
- A few relatively small-area and small-population counties were combined with an adjacent geographic area.

Geographic area stratification

As described previously, given the focus on implementation flexibility, the geographic area stratification was coarse, with no more than two strata of geographic areas within each state. The strata were roughly defined based on population size and urban or rural classification.

Systematic Sampling Within States

Systematic sampling is used to select clusters to cover the expected 10-year life of the current NHIS sample design. To facilitate Census Bureau field operation planning, the locations of the majority of the sample were determined before the first year of the sample design period and known for all successive years of the sample design period. The systematic sampling operations used for the 2016–2025 NHIS are very similar to those used for the previous NHIS designs in the Type A strata. In the Type B strata, the population tends to be less dense. In these situations, the annual sample generally needs to be about 100 sampled addresses per geographic area containing sample addresses. For the Type B strata, the geographic areas are ordered (implicit stratification) by MSA or non-MSA status and total MOS (in 2010 Census HU counts). For systematic sampling with an address sampling interval (SI), the expected number of sampled K-sized cluster units within a geographic area is

$$\frac{\text{MOS}_K\,(\text{geo area})}{SI_K}$$

where MOS_K is the number of *K*-sized clusters in the geographic area, and SI_K is the *K*-cluster sampling interval. (Both are defined by the address count divided by *K*.)

In the Type B strata, state-specific SIs were chosen to obtain samples of clusters of approximate size containing 2,400 addresses to cover up to 12 years of data collection and to have additional sample clusters in reserve for possible future needs. The actual HU addresses are not sampled at this selection stage. Conceptually, a state's Type B stratum is considered as a set of clusters of approximate size containing 2,400 addresses each that cover the entire B stratum. The sampling identifies the geographic area of each sample cluster but not the actual addresses within a cluster to be used for the annual surveys. The actual addresses are identified by the Census Bureau a few months before when interviewing is scheduled to occur.

In the Type A strata, there is more flexibility as to cluster size and proximity of projected HUs within. Smaller clusters can be combined to form larger clusters of approximate size containing 2,400 addresses. Techniques similar to those used as for the Type B strata are used to create a super sample.

The conceptual sampling employs the concept of a "measure" (different from MOS), which is a cluster of addresses with expected size four. Before the conceptual sampling, each census block is partitioned by the Census Bureau into an integer number of measures. About 25 measures are needed to achieve 100 sampled addresses annually and roughly 600 measures are needed to achieve 2,400 sampled addresses to cover 12 years of data collection.

The transition from the conceptual sample to the actual sample consists of assigning addresses to the measures. In the area frame, after the area frame listing results come back from the field, the Census Bureau implements an operation called the "sampling of listings" that assigns the addresses in the listing to measures. In the unit frame, a similar operation occurs using MSG addresses that have passed the filtering criteria.

The Type B and Type A super samples are treated as systematic cluster samples from sampling universes of equal MOS-sized clusters. The super sample units can be treated as having equal probabilities of selection within their respective strata. The super sample clusters in each stratum were assigned a sequential index of integers, starting with 1 (referred to as entry orders from here on) to define the order in which a subset of the super sample would be used for a specified sample allocation to the stratum.

For a state with both Type B and Type A strata, if n is the specified sample allocation to the state, and n_B and n_A are the numbers of sample addresses to be allocated to the Type B and Type A strata, respectively, then, $n = n_B + n_A$, and n_B and n_A are determined by the equation

$$\frac{\text{MOS(type-A stratum)}}{\text{MOS(type-B stratum)}} = \frac{n_A}{n_B}$$

This method of suballocation to n_B and n_A allows the same weight to be assigned to all sample addresses in the state because it enables equal probabilities of selection between Type A and Type B strata. In each state with both Type B and Type A strata, each Type A cluster was partitioned by the Census Bureau into three pieces of expected equal size, and a correspondence was established between the Type B clusters and the Type A cluster parts to allow for closer agreement to the specified n_B and n_A values. This correspondence also governs how Type B and Type A sample addresses move in or out of the sample concurrently if the state's sample allocation changes.

These procedures are for sampling geographical clusters that remain available over the duration of the current sample design. Within each cluster, the Census Bureau allocates the parts of the cluster to an annual survey year and month. This allocation attempts to minimize year-to-year intraclass cluster correlation and control field costs related to interviewer travel. Conceptually, the annual samples within clusters are treated as equal probability in nature. This procedure is implemented on all super sample clusters.

It should be noted that the super sample methodology discussed so far uses counties, census blocks, and expected measures of size as the fundamental components of sample construction. The actual physical HU addresses are designated by the Census Bureau either from a vendor or

census area list of the designated annual areas relatively close to when the addresses are scheduled for interview, as described previously.

Sampling Structures of the 2016–2025 NHIS Sample Design

In this section, the NHIS sample is presented in a mathematically structured form; these structures explain the methodologies used. There are a number of components to the sampling design.

Universe

Geographic clustering of the NHIS sample is essential to control the cost of field operations because the data are collected largely via face-to-face interviewing. Geographic clustering helps limit the amount of interviewer travel and thus the cost of collecting data.

The nation can be thought of as partitioned based on geographically clustered addresses in which the civilian noninstitutionalized population may live. The NHIS sample design follows a cost-efficient sampling strategy to define and select sample address clusters to represent the nation over a possible 12-year sample design lifecycle. The following list presents the geographical units that are the building blocks of the 2016–2025 sampling system.

- Nation: Partitioned by the 50 States and the District of Columbia. These are the primary strata.
- State: Partitioned into individual counties or aggregate adjacent counties, referred to as geographic areas. Each of these areas has sufficient population to be considered suitable to support the workload of at least one census interviewer over the sample design period.
- State strata: The geographic areas were partitioned into at most two groups, or strata, within each state. The partition is based on geographic area population and metropolitan status and each is referred to as a Type A or Type B stratum. The Type A strata tend to cover major metropolitan areas.
- State Type B stratum: Implicitly stratified by its sampling universe geographic areas.
- State Type A stratum: Implicitly and explicitly stratified by its sampling universe geographic areas.
- Individual geographic area: Structured to allow sampling of address clusters suitable for interviewer workloads. These clusters project an expected size K (about 100) addresses per year plus an extra K addresses for possible annual oversampling contingencies. Further, each annual sample is extended to include addresses found nearby to cover a possible 12-year survey cycle period. In the Type B geographic areas, the total expected count value, $K \bullet 2 \bullet 12$ (annual, annual extra, 12-year), may be about 2,400 addresses, while in Type A geographic areas this value may be about 800 addresses.
- Extended workload cluster: The large $K \bullet 2 \bullet 12$ expected sized cluster is treated as the fundamental unit to be sampled and serves as a source of interviews over the life of the survey. The addresses within the large cluster are not necessarily compact with respect to geography. For example, in a geographic area containing more than one county, the addresses in some of the large clusters within the geographic area may be in more than one county.

This sampling universe partition structure serves as a conceptual universe framework for developing a sampling methodology that considers both survey objectives and field operations. This conceptual universe can be thought of as a collection of strata where each stratum consists of a collection of extended workload clusters and each cluster has the same MOS in expected number of addresses. The cluster sizes are uniform within strata, but can vary between strata. Such a framework fits well with systematic sampling, which is a preferred operational sampling method in large scale sampling operations.

Sampling the universe

With a conceptual universe structured as above, a description of the sampling follows.

- National sampling interval: First, an address-based, annual sampling interval for the nation, SI_0^a, that will yield about 35,000 completed interviews with states being sampled approximately proportional to size is identified. This sampling interval value would be used if the entire nation was considered as the only stratum and no clustering was needed.
- State sampling interval: As state estimation was an important consideration during the initial planning, some state-level tuning of the national sampling interval was needed. First, the state Type B clusters tend to be geographically large with significant field operational constraints imposed on the minimal size of annual samples. Fractional cluster sizes are problematic because they could lead to incomplete interviewer workloads or increased interviewer travel. To ensure that an integral number of clusters would be selected in a state's Type B strata, national sampling intervals and workload sizes are slightly adjusted to force integral sized Type B workloads (at least with the MOS being used). More precisely, if N_B is the MOS of a Type B stratum and K_B is the annual address cluster size, the national sampling interval, SI_0^a, is state-adjusted by a multiplicative factor to define a modified sampling interval, SI_{state}^a, so that

$$N_B = SI_{state}^a \bullet S_B \bullet K_B^{'}$$

where S_B is an integer, then the sample consists of S_B clusters of size $K_B^{'}$. The value $S_B \bullet K_B^{'}$ will be approximately

$$\frac{N_B}{SI_{state}^a}$$

(If modified, the value K_B' will replace K_B in the extended workload cluster size discussed earlier.)

This same Type B sampling interval will also be used on the Type A strata in states with both Type A and Type B strata. If N_A is the MOS of a Type A stratum and K_A is the annual cluster size in addresses, then

$$N_A \approx SI^a_{state} \bullet S_A \bullet K_A$$

Here, the ratio

$$\frac{N_A}{SI^a_{state}}$$

will usually not be an integer; for survey operations the value S_A will be mapped to an integer value.

As workload efficiency was an important component of the sampling methodology, a planned cluster size relation, $K_A < K_B$, would suggest that the Type B strata, frequently covering the nonmetro areas within a state, will have a few annual clusters of large size, while the Type A strata, covering the larger metropolitan areas, would have many annual clusters of smaller size.

- Super sample: To select a larger state sample than specified by a national design that will yield about 35,000 completed interviews with states being sampled approximately proportional to size, the state sampling interval specified above can be reduced. Rather than using SI^a_{state} as a sampling interval, a reduced sampling interval

  $$\frac{SI^a_{state}}{R_B}$$

  is used, with $R_B > S_B$, R_B chosen to be an integer that makes the value of the product $R_B \bullet S_B$ 30 or greater for most strata. Here, the population or sample relation can be expressed as

  $$N_B = \frac{SI^a_{state}}{R_B} R_B \bullet S_B \bullet K_B$$

  then the sample consists of $R_B \bullet S_B$ clusters of size K_B. This set of clusters will be referred to as the super sample. The Type A strata super sample is structured in a similar way.

- Extended workload clusters: In the discussion of the super sample concept above, the value of K_B can be replaced with the size of the extended workload cluster, $[12 \bullet 2 \bullet K_B]$, then the Type B population is now counted as

  $$\frac{N_B}{12 \bullet 2 \bullet K_B}$$

  units of size $[12 \bullet 2 \bullet K_B]$ so the super sample consists of $R_B \bullet S_B$ clusters of size $[12 \bullet 2 \bullet K_B]$.

- Subsampling the super sample clusters: In a Type B stratum, the use of the systematic sampling interval

  $$\frac{SI^a_{state}}{R_B}$$

  yields a total of $R_B \bullet S_B$ potential sample clusters. Any equal probability method to select S_B clusters from the $R_B \bullet S_B$ cluster total will provide an equal probability sample.

- Allocation of the extended workload cluster to annual workload clusters: The construction of the $[12 \bullet 2 \bullet K_B]$-sized cluster allows for a partition into 12 components each of size $[2 \bullet K_B]$ in such a way that the K_B units used for the annual sample provide an unbiased estimator of the cluster total.

- Sampling weights. The systematic sampling and requirement of equal sized clusters leads to multiplicative weighting factors depending on the cluster sizes. In general, if N is a population size with size measured in addresses, K is the uniform cluster size and n is the number of clusters sampled, then the sampling weight, wt_{samp}, satisfies the expression $N = n \bullet K \bullet wt_{samp}$ and each of the $n \bullet K$ units is given a sampling weight wt_{samp}.

  Two important cases:

  Super sample: Each of the $R \bullet S \bullet K$ total sample cases gets a weight

  $$\frac{SI^a_{state}}{R}$$

  Annual sample: Take one of the R clusters: each of the $S \bullet K$ sample cases gets a weight SI^a_{state}.

- Assigning entry orders to the super sample within a state: The super sample defines clusters in both the Type A and Type B state strata. Only a subsample is used for the typical design year (for example, only S_B designated clusters from available $R_B \bullet S_B$ total sample are used for the typical annual sample) and the rest are kept in reserve. For each state Type A or Type B stratum, a super sample entry-order sequence for the sample extended workload clusters is established.

  For the Type B super sample, a value of $R_B \bullet S_B$ total super sample clusters was targeted. Here, the original systematic sequence $1,2,\ldots,R_B \bullet S_B$ was assigned to an R_B by S_B grid, columns filled first, and then the rows and columns were randomly shuffled. Each resulting row consisting of S_B components was considered a systematic subsample of size S_B. Entry orders of sample would sequentially move across the rows, and then at the last column go to the next row and start again. This shuffling of the original systematic sample attempted to keep the components geographically dispersed while maintaining an equal probability sample. Some small population states, originally planned to have two sample clusters in the core sample for a national sample allocation proportional to state population size, had three sample clusters in the core sample for the

national sample allocation that was implemented. If any of these small states had a Type B geographic area with a much larger population than the remaining Type B areas in the state, the larger area could have had two sample clusters in the core sample that was implemented.

The initial Type A super sample ordering was not carried out in the intended way in some areas due to a combination of miscommunication between NCHS and the Census Bureau and a programming error. NCHS and the Census Bureau reordered the Type A super sample in these areas in mid-2017; this went into effect at the beginning of 2018 (11). In some states, changes in the location of part of the Type A annual sample as of the beginning of 2018 were made, due to the reordering.

- Balancing the samples to achieve equal probability of selection within states: The sampling plan requires that each state use an equal probability of selection method (EPSEM) (12) constant over the entire state. If both Type A and Type B super samples exist, then the entry orders of Type A and Type B need to be coordinated to satisfy the condition:

  If n_A and n_B are the numbers of expected addresses in the Type A and Type B state strata, respectively, and N_A and N_B are the measures of total size of the Type A and Type B state strata, respectively, then for an EPSEM sample, the relation

$$\frac{n_A}{n_B} = \frac{N_A}{N_B}$$

  must hold. An approximate correspondence was established between the Type A and Type B entry orders to achieve this. In the event of sample size change in a state with both Type A and Type B areas, the change is implemented as a combination of Type A sample change and Type B sample change in a way that preserves the EPSEM relationship in the state.

- Active and reserve samples: The super sample was planned with the mid- and smaller- sized population states in mind. The required cost efficiency often results in few sample clusters in some states, making clustering effects more pronounced in design-based state data analyses. A structured method of adding sample via the entry-order sequence is a manageable way to add sample to a state.

Conceptual design to actual implementation

The conceptual universe partition and sampling mechanisms discussed above are ideal structures that provide a framework for planning, but from past experience with previous NHIS sample redesigns it was anticipated that many departures from the ideal framework would occur for the 2016 design implementation.

- The MOS used is an expectation of the number of addresses in a cluster, not the true number. The true value can change over time. The major planning for the sampling operation was done by 2015 to identify sample clusters and project a static size that would be reasonable over the life of the survey. At any point in time, the number of actual eligible occupied HUs in an annual cluster can vary. The 2010 census HU counts available at different geographical levels were the source of MOS parameters.
- Forming an integral number of clusters of a fixed size within a geographic area is rarely achievable. There are many operational constraints and special situations considered when forming clusters. Over the many years that they have conducted NHIS, the Census Bureau has developed many contingency plans for difficult sampling situations. The ideal partition and sampling plans just discussed can be considered reasonable structures for the implemented design.
- The structuring of the universe and the sampling mechanisms conceptually involves many moving parts. Actual implementation of design strategies requires several iterations and sampling parameter changes to achieve a final result.
- Systematic sampling is commonly used in complex multistage survey designs but has the drawback that an unbiased estimator for the variance of a linear estimator does not exist. For NHIS, the sampled clusters will be treated as having been sampled with replacement. This simplified design structure is widely used for government surveys.

Sampling of People Within Households

One sample adult and one sample child (if children under age 18 are present) are selected from each family for administration of a large portion of the NHIS adult and child interviews. For 2016–2018, all adults (aged 18 and over) were assigned the same sampling probability, except for any Black, Hispanic, or Asian people aged 65 and over who were given twice the chance of selection compared with the rest of the adults in the family. Using this sampling algorithm, one sample adult is selected by the interviewer's computer using an automated sampling system. The child is selected at random, and no differential sampling probabilities are applied to the children. Beginning in 2019, no differential sampling probabilities are applied to the adults, and the sampling of sample adults and sample children changed from the family level to the household (HU or sample address) level.

Logistics and Special Sampling Scenarios

Panels

Since 1985, the annual core NHIS sample has been partitioned into four nationally representative subsamples, referred to as panels. Each sample address is assigned to only one panel. The four panels have close to identical marginal sampling properties; in particular, each can produce

unbiased population estimates. Each panel has about the same interviewer workload. Each Type B geographic area is assigned to only one panel in most cases.

The primary objective for creating NHIS panels is to have a contingency to handle potential budget cuts. Large sample reductions can be made, if needed, by dropping one or more panels, either on a quarterly basis or for the entire year's data collection. This is much more practical and cost efficient than dropping months of interviewing or dropping randomly chosen finer-level sample units. The secondary objective was to provide a subsample that could be used as a sampling frame for any smaller survey linked to NHIS. For instance, the MEPS sampling frame is derived from two NHIS panels. In MEPS, the term panel is used as a longitudinal data descriptor rather than for referencing a subdesign, as in NHIS.

For the design of the panels, address clusters were systematically assigned to panels after the original systematic sampling was done. Large Type A geographic areas usually have representation in two, three, or all four panels. Some special situations caused the sample in a few smaller geographic areas to be partitioned into pieces and assigned to multiple panels.

Interviewer assignments

The NHIS sample in each geographical area is partitioned into geographically based address clusters that provide for a monthly interviewer workload of about 8–12 sample addresses per month each year over the life of the 2016–2025 design. The workload in each cluster during a given year is usually contained within a single county, or two contiguous counties, making the annual workload in a cluster manageable for a single interviewer both from a workload and a geographical standpoint.

Group quarters

Noninstitutionalized nonmilitary group quarters such as college dormitories house people eligible for inclusion in NHIS. The within-group quarters sampling procedures are similar to those used in the rest of the sample. Before the first interviews at a group-quarters address can be conducted, an interviewer visits the group quarters to establish a list of eligible units (for example, rooms, beds, or people). A systematic sampling pattern is applied to the listing to identify the people to be interviewed. In the 2010 census, less than 2% of the target population lived in group quarters.

In 2016–2017, a separate sampling frame for sampling college dormitory residents was used, due to concerns of insufficient coverage of this subgroup of the target population on the MSG address file. Use of this separate sampling frame was discontinued at the end of 2017, due to low response rates. Beginning in 2018, the NHIS interview questionnaire was modified in an attempt to enumerate college students at the residence they lived at before entering or when not at college.

Estimation Structures for 2016–2025 NHIS

NHIS is designed to make inferences about the civilian noninstitutionalized population of the United States. Although a general description of the NHIS sample design is presented in the section "2016–2025 NHIS Sample Design" of this report, this section focuses on the design and estimation structures. The NHIS program at NCHS is focused on making design-based estimates about the health of people in the target population. This is accomplished by inflating the responses of each surveyed person in NHIS, referred to as an elementary unit, by a national weight factor that permits an approximately unbiased design-based estimator of any U.S. target population total. With this weight, an unbiased estimator, $\hat{X}$, for any given true population characteristic total, X, can be expressed as a weighted sum over all elementary units:

$$\hat{X} = \sum_{u} W_f(u)x(u) \qquad [1]$$

where

u indexes the elementary units of NHIS,

$x(u)$ is the characteristic or response for unit u, and

$w_f(u)$ is the final national weight for unit u.

This estimator is used to generate NHIS estimates of population totals, as well as numerators and denominators of percentages and rates that appear in official publications. The final national weight is provided on NHIS public-use microdata files, which allow users to directly create estimators of the form in equation 1. In the following sections, the technical aspects of the procedures used to create weights and estimate the variances of NHIS estimators are discussed.

Complex estimation techniques are required for NHIS because the survey is based on a stratified probability sample. The true sampling distributions of any survey implementing complex clustering structures, implicit stratification, and systematic sampling tend to be mathematically intractable; this means it is impossible to write a mathematical expression that correctly describes the sampling distribution. For this reason, the NHIS design is conceptualized in a somewhat simplified framework, to provide a tractable model that still captures the most important design features.

The number of eligible households in an address cluster is treated as a random variable. An HU may be classified as eligible or ineligible (for example, vacant dwellings or no civilian members). Nationally, about 20% of all addresses yield an ineligible classification. The annual number of people who complete NHIS interviews also is a random variable. As a consequence, the year-to-year NHIS sample counts of households and people vary.

NCHS applies three broad estimation criteria when deciding on estimation strategies to use for NHIS data:

1. The basic estimation methods are design-based for finite populations. That is, the randomness of the data is a result of sampling finite universes having no imposed distributional assumptions. This is in contrast to a model-based approach, where the data usually have imposed distributional assumptions. The design-based methods may be thought of as nonparametric and robust.
2. The design-based methods should be practical and permit approximately unbiased estimators of population totals.
3. The design-based methods should permit practical variance estimation strategies to assess the stability of the estimator.

To satisfy these criteria, NCHS, as well as many other large government surveys, has been using standard, accepted design-based methods discussed in such classic references as Kish (12), Cochran (13), and Hansen, Hurwitz, and Madow (14).

Creating Respondent Weights

The following procedures were used to compute the final weights to accompany the public data release for survey years 2016–2018. Weighting procedures were updated for survey year 2019 and beyond in response to a 2019 questionnaire redesign, and the changes to the 2019 procedures are discussed in a separate document (15). The changes included nonresponse adjustments based on propensity models and a more extensive raking calibration method.

The NHIS estimator of a characteristic total, as presented in equation 1, uses methodology based on the features of a complex probability sample to define a national weight, W_f, for each responding unit, which is the product of up to four weighting factors:

- Inverse of the probability of selection
- First-stage household nonresponse adjustment
- Second-stage household nonresponse adjustment
- Calibration to independent population estimates (poststratification or raking)

When the analytic unit is a person, all four weighting factors contribute to the individual's final weight. Because the NHIS ratio adjustment is based on person-level characteristics, only the first three weighting factors are used to create the household weight.

NCHS creates weights designed to be nationally representative and to sum to the population totals for each calendar quarter of the NHIS sample, using information provided by the U.S. Census Bureau. These weights permit national population estimates to be made for each quarter. Resulting quarterly weights are divided by four to create annual weights, which are used when making national population estimates from a calendar year of the NHIS sample.

Base weight

The overall probability that a unit—whether cluster, household, or person—is in the sample is the product of the conditional selection probabilities. This basic inflation weight is defined as:

$$W_I(u) = \frac{1}{\text{Prob(unit u is in sample)}}$$

Generally, based on probabilistic sampling, unit u represents $W_I(u)$ population units.

For the 2016–2025 NHIS, the first (and only) stage of sample address selection is a sample of address clusters within a geographic area.

Initially, each HU in a cluster has its basic inflation weight $W_I(u)$ equal to the sampling interval used within the state as specified for the annual survey.

Infrequently, this base sampling weight, W_I, will be modified. During the sampling of listed units from the area frame, if a block is found to have a much larger number of addresses compared with the block's 2010 census HU count that would result in measures being assigned more than eight addresses, subsampling occurs to ensure measures are assigned no more than eight addresses. If the subsample consists of less than one-quarter of all addresses, then the conditional probability of selection will be truncated by the Census Bureau at one-quarter. This rarely occurs, and the biases introduced by such a modification should be small. There is no modification of the base sampling weight in the unit frame.

In an ideal, hypothetical sampling situation having no nonsampling error components (for example, frame problems, nonresponse, or interviewer effects), equation 1 with W_I substituted for W_f becomes

$$\hat{X}_1 = \sum_u W_I(u)x(u) \qquad [2]$$

which will provide an unbiased estimator for the true population total X. Such an estimator is referred to as a base weight Horvitz–Thompson estimator.

Factors contributing to year-to-year fluctuations in weights

Complex sampling implemented by specified guidelines is difficult to achieve in the real world. Below are certain special situations that may occur when conducting NHIS and

that may have an influence on weighting procedures for a given year:

- Initial start-up problems during the early years of the sample design period may result in minor deviations from the sampling plan.
- NHIS budgetary changes may result in additional subsampling to reduce the sample (see subsection "Panels" in section "2016–2025 NHIS Sample Design"), or supplemental funding may result in sample augmentation. For example, augmentation occurred for the 2016, 2018, and 2019 NHIS.
- Phase-in of new field operations may modify the sample.
- Unexpected onetime events may alter the design.

First-stage household nonresponse adjustment for 2016–2018 weights

During the first year of the current NHIS design, 2016, the household nonresponse rate was about 32%. This form of nonresponse exerts a downward bias on an estimator of total, such as in equation 2; as a result, a weighting adjustment for household nonresponse was implemented. To account for the nonresponse, a second weight factor, the first-stage household nonresponse adjustment, is applied.

The segments for the first-stage nonresponse adjustment are geographical segments defined using variables created by Census Bureau staff when they partitioned the sampled address clusters into pieces for assignment to year, and within year, to month.

The standard household nonresponse adjustment inflates the sampling weights for all responding households within a segment to make up for the nonresponding households within the same segment. However, this adjustment does not address the situation where a segment has 100% nonresponse. Usually, a small number of segments in a quarter have 100% nonresponse.

In the 1985–1994 NHIS design, when all eligible households in a segment were sampled with certainty (that is, no screening occurred), NCHS used the nonresponse adjustment

$$W_{NR,1} = \frac{\sum \text{all eligible households in segment}}{\sum \text{all responding households in segment}}$$

This unweighted sum is usually equivalent to the sum obtained using the base weights because the base weight is constant within a segment except for the rare event in which the NHIS interviewer discovers three or more additional units at a sample address.

In the current sample design, in the absence of screening (that is, using additional criteria to determine whether an eligible person from within the household will be sampled) the above equation once again is the procedure for the first-stage nonresponse adjustment except when the above-described rare event occurs; if it does, the subsampling factor is applied to the sample units at the subsampled address.

If the ratio of eligible to responding households is greater than two, the first-stage nonresponse adjustment is truncated to two.

Second-stage household nonresponse adjustment for 2016–2018 weights

The second-stage household nonresponse adjustment occurs in geographic areas where segments with 100% nonresponse occur, where the first-stage nonresponse adjustment is truncated to two, or both, at a higher geographic level (for example, the geographic area) than the segment used in the first-stage nonresponse adjustment.

Within the higher geographic level where the second-stage household nonresponse adjustment is to be applied, two sums are accumulated:

SUM 1: For responding households, the sum of the product of the factors (base weight) • (subsampling factor, if applicable) • (first-stage nonresponse adjustment factor).

SUM 2: For responding households, the sum of the product of the factors (base weight) • (subsampling factor, if applicable) • (untruncated first-stage nonresponse adjustment factor); and for nonresponding households in segments with 100% nonresponse, the sum of the product of the factors (base weight) • (subsampling factor, if applicable).

The second-stage household nonresponse adjustment factor is

$$W_{NR,2} = \frac{\text{SUM 2}}{\text{SUM 1}}$$

If the factor is more than 1.5, it is truncated to 1.5. At this point, the base inflation weight, W_I, in equation 2 is inflated by the product $W_{NR,1} \bullet W_{NR,2}$ to define a final household-level weight. An estimator of a population total based on nonresponse adjustment is defined as

$$\hat{X}'_2 = \sum_u W_I(u) \bullet W_{NR,1} \bullet W_{NR,2} \bullet x(u) \qquad [3]$$

The household-level weights are the starting point for the set of weights described below.

2016–2018 person-level weights

Person-level weights are created from the household-level weights using a ratio adjustment and are intended for use when analyzing data about all the household members. Statistical sampling theory has shown that in many situations, the estimators obtained using a ratio estimation procedure often have smaller mean squared error (MSE) than the base weight estimators expressed by equation 2. More precisely, if $\hat{X}$ and $\hat{Y}$ are base weight estimators of two population

characteristic totals, X and Y, respectively, and if the "true" total Y is known, then the ratio estimator

$$\tilde{X} = \frac{\hat{X}}{\hat{Y}} Y$$

for X will have smaller MSE than the estimator $\hat{X}$ when there is a high positive correlation between $\hat{X}$ and $\hat{Y}$, and the sample size is large.

The ratio adjustment also is used to help correct survey bias due to systematic undercoverage. An observed survey estimator of the form in equation 3 may be larger or smaller than the true value just by chance alone. The U.S. Census Bureau has identified some populations as hard to sample. For example, historically, survey undercoverage of the young Black male population has occurred in NHIS, and the estimator of equation 3 may be negatively biased in estimating young Black male population characteristic totals. Such a bias due to undercoverage is often reduced by using ratio adjustments.

Ratio adjustment is applied at the person-level, which can introduce variation in the person-level weights within a given sampled household. The previous three components of the weights—the inverse of the probability of selection and the first-stage and second-stage household nonresponse adjustments—are equal for all people in a given sampled household. This weight ensures that NHIS estimates for 100 age-sex-race and ethnicity classes of the civilian noninstitutionalized population of the United States (Table) agree with independently determined population controls prepared by the Census Bureau. Furthermore, these independent controls are the same controls used for the Current Population Survey. Thus, national population estimates for any combination of the age-sex-race and ethnicity groups from the two surveys are about the same, which may enhance the comparability of estimates from the two surveys.

The independent controls until 2021 are based on the U.S. Census 2010 estimates. Starting around 2022, the independent control totals will be derived from the 2020 U.S. Census data.

Each month, the Census Bureau produces national estimates for the 100 age-sex-race and ethnicity classes. Although NHIS is conducted monthly, the poststratification adjustment is computed only for NHIS quarterly estimates. NHIS quarters and the dates of the population estimates used as the controls are:

NHIS quarter	Population estimate
January–March	February 1
April–June	May 1
July–September	August 1
October–December	November 1

For each NHIS quarter, 100 age-sex-race and ethnicity adjustment weights are computed, for a total of 400 adjustment weights annually. If a represents one of the 100 age-sex-race and ethnicity classes, $Y(a)$ represents the Census Bureau population estimate for class a, and $\hat{Y}_2'(a)$ represents the NHIS nonresponse-adjusted national total for class a as expressed in equation 3, that is,

$$\hat{Y}_2'(a) = \sum_u W_I(u)\, W_{NR,1}(u)\, W_{NR,2}(u)\, I_a(u)$$

where $I_a(u) = 1$ if person u is in class a, 0 otherwise, then the ratio adjustment for class a, $W_R(a)$, is defined as

$$W_R(a) = \frac{Y(a)}{\hat{Y}_2'(a)}$$

In implementing this ratio adjustment, NCHS generally requires each class a to contain at least 30 sample people. If a class contains too few sample people, that class will be pooled with an adjacent age class. Similarly, pooling occurs if a factor falls outside of the interval 0.7–2.0.

The final ratio adjusted national estimate, $\hat{X}$, of a population total, X, is defined by equation 1 with the weight W_f defined by the product of the four component weights:

$$W_f(u) = W_I(u) W_{NR,1}(u) W_{NR,2}(u) W_R(u)$$

Table. The 100 age, sex, and race and ethnicity classes used for poststratification: National Health Interview Survey, 2016

Hispanic male or female	Non-Hispanic Black male or female	Non-Hispanic Asian male or female	Non-Hispanic other male or female[1]
	Age group (years)		
Under 1	Under 1	Under 5	Under 1
1–4	1–4	5–17	1–4
5–9	5–9	18–24	5–9
10–14	10–14	25–44	10–14
15–17	15–17	45–64	15–17
18–19	18–19	65 and over	18–19
20–24	20–24	…	20–24
25–29	25–29	…	25–29
30–34	30–34	…	30–34
35–44	35–44	…	35–44
45–49	45–49	…	45–49
50–54	50–54	…	50–54
55–64	55–64	…	55–64
65 and over	65–74	…	65–74
…	75 and over	…	75 and over

… Category not applicable.
[1]Includes non-Hispanic non-Black and non-Hispanic non-Asian people.

SOURCE: National Center for Health Statistics, National Health Interview Survey, 2016.

Thus,

$$\hat{X} = \sum_u W_f(u) = \sum_u W_I(u) W_{NR,1}(u) W_{NR,2}(u) W_R(u) x(u) \qquad [4]$$

No person-level weight was created starting in 2019 due to the changes in the 2019 redesign questionnaire (14).

2016–2018 family, sample adult, and sample child weights

The previous discussion outlined the procedure for creating household-level and person-level weights for NHIS in years 2016–2018. Starting in 1997, other weights also were created for sample adult, sample child, and family-level files. The basic strategy for creating these other weights is very similar to the previous discussion. The nonresponse-adjusted household-level weight is the starting point for the creation of the sample adult and sample child weights.

The family weight (2016–2018) is intended to be used when making estimates at the family level. The NHIS family weight takes the value of the person weight for one of the people in the family. A person-level ratio adjustment is used as a proxy for the NHIS family-level ratio adjustment. Davis (16) has shown that the person weight with the smallest ratio adjustment within each family—that is, the smallest poststratification factor between the interim and final person weights within the family—provides a more accurate estimate of the total number of U.S. families than a weight that does not include a poststratification factor. The NHIS family weight follows this procedure. Starting in 2019, due to changes in the questionnaire, a family weight is no longer computed.

Creation of the sample adult weight and the sample child weight begins with the nonresponse-adjusted household weight. An inflation factor is applied to account for selection of the sample adult and sample child within the family (household as of 2019) (14), and then a poststratification occurs. The inflation factor is the inverse of the selection probability in that household. For example, in a household of two eligible adults and two eligible children, the sample adult has a selection probability of one-half, so their inflation factor is two. The process is similar for children.

Before the 2010 NHIS, any nonresponse bias to the sample adult core and sample child core was assumed to be accounted for through the poststratification process. Beginning with the 2010 NHIS and through 2018, NCHS added a nonresponse adjustment for the sample adult and sample child weights that uses methodology similar to that used for the geographic household first-stage nonresponse adjustment. The nonresponse adjustment is calculated after the inflation factor for selection of the sample adult and sample child is applied, and before poststratification.

Since 2010, there are separate poststratification steps for sample adult and sample child weights. The sample adult and sample child poststratification adjustment factors are modified (that is, collapsed with another class) if the factors fall outside the interval 0.7–3.0. The sample adult and sample child weights use a smaller set of poststrata than the 100 poststrata listed in the Table because of the subsampling processes for the sample adult and the sample child. A total of 52 sample adult poststrata and 30 sample child poststrata exist.

Adjacent years of NHIS data are often combined for pooled analysis (for example, 2016, 2017, and 2018) to increase the sample sizes for some small demographic or geographic subpopulations. The sampling weights for pooled data should be adjusted; otherwise, annualized estimates of totals will be too high. A valid weight adjustment procedure that NCHS recommends is to divide each sample weight in the pooled data set by the number of years that are being pooled: for example, divide by three when 3 years of data are combined. A Survey Description Document (SDD) that includes details specific to that year's data collection is provided for each year's data release. SDDs can be found at: https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm.

Variance Estimation

Most of the estimates produced by NCHS from NHIS are totals and ratios of totals, such as means and percentages. All such totals and ratios of totals are produced using the appropriate final weight described in the previous sections. These estimators are subject to both sampling and nonsampling errors. Nonsampling errors include measurement errors, coverage errors, and nonresponse bias, all of which are difficult to measure and quantify, but an effort is made to minimize such errors at each step of the NHIS operation. The sampling error, however, can be measured by the variance of the estimator.

Although equation 1 provides a functional form that permits simple computation of point estimates, the variances of such estimators are harder to compute. The functional form of a variance estimator depends on the nature of the sample design and methods used to adjust the weights. Some complexities in the NHIS survey design require special techniques:

- The sampling (for example, selection of one or more address clusters within a geographic area) is the result of a very complicated process involving features such as estimating measures of size and applying systematic sampling techniques. Even given the census information about the geographic area, defining a "user friendly" sampling mechanism that captures the system's true stochastic structure and can be implemented with a standard variance estimation procedure is extremely difficult.
- To protect survey respondent confidentiality, NCHS does not release design information that could be used to identify smaller geographical areas where NHIS was conducted. Small sample areas with rare socioeconomic

or demographic characteristics must not be explicitly or implicitly identifiable by design information.

- With weighting adjustments applied to the base weight, estimates of totals become nonlinear. This complicates the variance estimation procedure.
- In practice, data analysts who use NHIS data use large sample theory when making inferences about populations. Variance estimation procedures suitable for large subpopulations may be unstable for smaller subpopulations. NCHS uses stable, all-purpose variance estimation structures that should be easy to implement with existing computer software.
- As mentioned previously, adjacent years of NHIS data are often combined for pooled analysis (for example, 2016, 2017, and 2018) to increase the sample sizes for some small demographic or geographic subpopulations. Estimates produced from different years of data within the same sample design period are dependent (that is, the sample drawn for each year is not independent), while estimates produced from different years in different sample design periods are independent for variance estimation purposes. Further discussion of variance estimation for pooled analyses when the years fall into different sample design periods, or when changes occur to the public-use design variables, is available on the NHIS webpage.

Simplified Design Structures for Variance Estimation

Wolter (17) and Rust (18) comprehensively discuss design-based variance estimation for complex surveys. Of the available methods, the three most commonly used are Taylor series linearization, balanced repeated replication, and the jackknife. Software for analysis of complex surveys includes the R survey package, SAS survey procedures, SPSS, Stata, SUDAAN, and VPLX.

NHIS public-use microdata files currently contain design information suitable for the Taylor series linearization method.

In the following discussion, a simplified design structure that allows design-based variance estimation for NHIS is detailed.

Conceptual NHIS Variance Estimation Structure

The NHIS conceptual variance estimation structure takes into account the sampling stratification by state and where applicable, within state. The systematic sampling of address clusters is treated as simple random sampling with replacement of clusters, where the clusters are broader in the Type B areas than in the Type A areas.

A variance estimator for an estimator of the form

$$\hat{X} = \sum_{u=1}^{n} W_f \bullet x_u$$

can be defined as:

The n sampled units u are the components of C clusters that are found in H strata. These clusters and strata are state-based. For each stratum h, there will be C_h clusters, with each cluster c having weighted cluster total

$$\hat{X}_{hc} = \sum_{u \in c} W_f \bullet x_u$$

The estimator $\hat{X}$ can thus be expressed as

$$\hat{X} = \sum_{h=1}^{H} \sum_{c=1}^{C_h} \hat{X}_{hc}$$

Treating the clusters as being sampled independently between and within strata, a variance estimator of $\hat{X}$ is

$$\widehat{Var}(\hat{X}) = \sum_{h=1}^{H} C_h \sum_{c=1}^{C_h} (\hat{X}_{hc} - \bar{\hat{X}}_{h.})^2 / (C_h - 1)$$

where $\bar{\hat{X}}_{h.}$ is the sample mean of the C_h cluster totals within stratum h.

Estimating variances for poststratified totals and nonlinear statistics

The final national weight estimator $\hat{X}$ of equations 1 and 4 incorporates a poststratification adjustment. This is the form of the estimator presented in official NCHS publications and the one that most analysts use. This estimator is nonlinear because of the poststratification adjustment. A commonly used method for estimating the variance of a nonlinear statistic is to linearize the statistic using Taylor series methods.

In practice, implementation of computer software packages based on linearization often requires treating the final weight, W_f, which may include a poststratification adjustment, as an inflation weight. For example, in the SUDAAN (19) version 11.0 software, regression statistics can be linearized but not with a simultaneous linearization of the poststratification weights. Thus, SUDAAN regression computations for variance assume that the final poststratification weight is an inflation weight. For estimated totals, this practice tends to lead to somewhat inflated variance estimators. For ratios of totals (for example, means or percentages), the impact varies. For many health variables, empirical evidence suggests that the inflation in the estimated standard errors of means may be of little practical importance. The treatment of the final weight, W_f, as an inflation weight may be reasonable if software limitations warrant such a simplification. Population domains that are aggregates of several component poststratification classes should be expected to have a greater variance reduction than population domains covered by few poststratification classes. In general, economic-type variables may have a

greater impact than health-type variables. For regression-type analysis, the inclusion of age-sex-race and ethnicity predictors tends to reduce the impact of treating the final weight as an inflation weight.

Public-use NHIS Data and Limitations on Design Structures

NCHS is required to prevent the disclosure of information that may compromise the confidentiality promised to survey respondents under the Confidential Information Protection and Statistical Efficiency Act, or CIPSEA, which refers to Title V of the E-Government Act of 2002, Public Law 107-347. To ensure confidentiality, some design information is not included in public-use data sets.

On the public-use data files, original state sampling strata and sampled address clusters may have been collapsed with others to avoid implicit or explicit geographical disclosure. The public-use variance estimation structures are designed to be robust to potential year-to-year sample allocation changes.

The techniques of stratum collapsing, stratum partitioning, and mixing are used to coarsen the Type A design structures with little anticipated bias, but at the expense of loss of degrees of freedom. These techniques are discussed in Parsons and Moriarity (20), Eltinge (21), and Parsons and Eltinge (22). The result is a design structure with an imposed two or more pseudo-PSUs per stratum and more than 300 nominal degrees of freedom.

The public-use variance estimator for an estimator of the form

$$\hat{X}_{hc} = \sum_{u \in c} W_f \bullet x_u$$

is structured very similarly to the in-house form, but with coarsened strata and clusters. The n sampled units u are collapsed or mixed into C clusters (referred to as public-use pseudo-PSUs), and these clusters are partitioned into H (public-use) pseudo-strata. These clusters and strata, while masking some of the true geographical clustering, capture much of the actual sampling geography. For each stratum h, there will be C_h clusters, with each cluster c having weighted cluster total

$$\hat{X}_{hc} = \sum_{u \in c} W_f \bullet x_u$$

The estimator $\hat{X}$ can thus be expressed as

$$\hat{X} = \sum_{h=1}^{H} \sum_{c=1}^{C_h} \hat{X}_{hc}$$

Treating the clusters as being sampled independently between and within strata, a variance estimator of $\hat{X}$ is

$$\widehat{Var}(\hat{X}) = \sum_{h=1}^{H} C_h \sum_{c=1}^{C_h} (\hat{X}_{hc} - \bar{\hat{X}}_{h.})^2 / (C_h - 1)$$

where $\bar{\hat{X}}_{h.}$ is the sample mean of the C_h cluster totals within stratum h.

Summary

The 2016–2025 NHIS sample design retained many features of previous NHIS sample designs such as in-person interviewing and a geographically clustered sample. It includes important new features such as a different source of most of the sample addresses and built-in flexibility to increase or decrease state-level sample sizes while maintaining the stability of the sampling weights.

References

1. Parsons VL, Moriarity C, Jonas K, Moore TF, Davis KE, Tompkins L. Design and estimation for the National Health Interview Survey, 2006–2015. National Center for Health Statistics. Vital Health Stat 2(165). 2014.
2. Botman S, Moore TF, Moriarity CL, Parsons VL. Design and estimation for the National Health Interview Survey, 1995–2004. National Center for Health Statistics. Vital Health Stat 2(130). 2000.
3. Massey JT, Moore TF, Parsons VL, Tadros W. Design and estimation for the National Health Interview Survey, 1985–1994. National Center for Health Statistics. Vital Health Stat 2(110). 1989.
4. Kovar MG, Poe GS. The National Health Interview Survey design, 1973–1984, and procedures, 1975–1983. National Center for Health Statistics. Vital Health Stat 1(18). 1985.
5. National Center for Health Statistics. Health Interview Survey procedure, 1957–1974. National Center for Health Statistics. Vital Health Stat 1(11). 1975.
6. National Center for Health Statistics. The statistical design of the Health Household-Interview Survey. Health Statistics. Public Health Service. PHS Pub. No. 584–A2. 1958.
7. Moriarity C, Parsons V, Jonas K. Overview of the 2016–2025 National Health Interview Survey sample design. In: Proceedings of the Joint Statistical Meetings, Survey Research Methods Section. Denver, CO: American Statistical Association. 2151–8. 2019.
8. Office of Management and Budget. Office of Management and Budget standards and guidelines for statistical surveys. 2006. Available from: https://obamawhitehouse.archives.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf.
9. Kuwik C, Moore B, Jonas K. Recommendation for the NHIS college housing frame, sample, and operations

(Doc. #2010-5.4-R-1 Version 1.0). U.S. Census Bureau's Demographic Statistical Methods Division. 2014.

10. Parsons V, Dienes E. State sampling allocation strategies for the 2016 redesigned National Health Interview Survey (NHIS). In: Proceedings of the Joint Statistical Meetings, Survey Research Methods Section. Seattle, WA: American Statistical Association. 2651–9. 2015.
11. Moriarity C, Parsons V. Nested subsamples: A method for achieving flexibility in annual sample sizes for a continuous multiyear survey. In: Proceedings of the Joint Statistical Meetings Survey Research Methods Section. Alexandria, VA: American Statistical Association. 2185–91. 2018.
12. Kish L. Survey sampling. New York, NY: Wiley. 1965.
13. Cochran W. Sampling techniques. New York, NY: Wiley. 1977.
14. Hansen M, Hurwitz W, Madow W. Sample survey methods and theory, vols I, II. New York, NY: Wiley. 1953.
15. Bramlett MD, Dahlhamer JM, Bose J, Blumberg SJ. New procedures for nonresponse adjustments to the 2019 National Health Interview Survey sampling weights. National Center for Health Statistics. 2020.
16. Davis KE. National Health Interview Survey family weighting research 2000. In: Proceedings of the Annual Meeting of the American Statistical Association. Alexandria, VA: American Statistical Association. 2001.
17. Wolter K. Introduction to variance estimation. 2nd ed. New York, NY: Springer-Verlag. 2003.
18. Rust K. Variance estimation for complex estimators in sample surveys. J Off Stat 1(4):381–97. 1985.
19. RTI International. SUDAAN language manual (Release 10.0) [computer software]. 2008.
20. Parsons V, Moriarity C. Review of NHIS public-design structures. In: Proceedings of the Joint Statistical Meetings. Alexandria, VA: American Statistical Association. 2903–9. 2007.
21. Eltinge J. Use of stratum mixing to reduce primary-unit-level identification risk in public-use survey datasets. In: Proceedings of the Federal Committee on Statistical Methodology. Washington, DC. 1999.
22. Parsons V, Eltinge J. Stratum partition, collapse and mixing in construction of balanced repeated replication variance estimators. In: Proceedings of the Joint Statistical Meetings. Alexandria, VA: American Statistical Association. 1999.

Appendix. Glossary

Area frame—A portion of the 2016–2025 National Health Interview Survey (NHIS) sample frame consisting of geographic areas where address listing operations are conducted in person by the Census Bureau to generate a list of eligible addresses from which NHIS sample cases are selected.

Block—A statistical area bounded by visible features, such as streets, roads, streams, and railroad tracks, and by nonvisible boundaries, such as selected property lines and city, township, school district, and county boundaries. A block is the smallest geographic unit used by the Census Bureau, and blocks nest within tracts.

Civilian noninstitutionalized population—People who currently live in one of the 50 states or the District of Columbia, who do not live in institutions (for example, penal and mental facilities or homes for older people) and who are not on active duty in the Armed Forces.

Delivery point type code—A code indicating the category of mail delivery point and its type of service. Examples are:

- Curbline—mail box is located at the curb
- Centralized box unit—mail box is located within a cluster box
- Central—mail box is located within a centralized unit

Delivery Sequence File (*DSF*)—A collection of data provided to the Census Bureau by the United States Postal Service (USPS) that contains address information, address-related information, and point of postal delivery information, including postal delivery codes, as may be determined by the Secretary of Commerce to be suitable for any census or survey being conducted by the Census Bureau.

Drop point—A single address that services multiple businesses or families. Examples include: a single box shared by more than one business or family, boarding or fraternity houses, and gated communities where mail for all homes is delivered to a gatehouse.

Erroneous exclusions—Units excluded from the frame that are considered valid for the purposes of the survey.

Erroneous inclusions—Units included on the frame that are not considered valid for the purposes of the survey.

Family—An individual or a group of two or more related people who are living together in the same household; for example, the reference person, their spouse, foster son, daughter, son-in-law, their children, and the wife's uncle. Unmarried cohabiting couples (same-sex and opposite-sex couples) are considered as belonging to the same family. Additional individuals, or groups of people living in the household who are related, but not related to the reference person, are considered to be separate families; for example, a lodger and their family, or a household employee and their spouse, or a single boarder with no one related living in the household. Hence, more than one family can live in a household, or a family can consist of only one person. Until the 2019 NHIS, each family was considered a separate case and interviewed separately. Starting from 2019, the entire household has been considered a single case with no distinction even if more than one family was living in a household.

General delivery—An alternate delivery service that lets customers with identification pick up mail at post offices. Provided mostly at offices without carrier delivery or for people who do not have a permanent address or who prefer not to use post office boxes.

Geocoding—The assignment of an address, structure, key geographic location, or business name to a location that is identified by one or more geographic codes. For living quarters, this usually requires identification of a specific block.

Group quarters—A type of living quarters where the residents share common facilities or receive authorized care (for example, dormitories, boarding houses, or convents). A group quarters unit does not meet the regular housing unit (HU) definition.

High-rise record type—May be used to identify a commercial building, apartment complex, high-rise, wing, or floor of a building, grouping of apartment mail boxes, or other physical location other than a street. A distinguishing feature of these types of records is that the record represents a single address rather than a range of addresses. If multiple records of this type for a given address exist, then there may be multiple secondary ranges or names to express different suites, room numbers, etc.

Highway contract routes—A route of travel served by a postal contractor to carry mail in bulk over highways between designated points. Highway contract routes usually do not deliver mail to individual customer addresses along the line of travel. Highway contract routes make up the largest single group of transportation services used by USPS and range from long-haul tractor trailers to box delivery routes. Some mailing addresses in rural areas consist of the combination of a highway contract route number and a box number.

Household—An entire group of people who live in one HU or one group quarters unit, made up of one or more families. It may include several people living together or one person living alone. A household includes the reference person and

any relatives living in the unit, and may also include roomers, live-in domestic workers, or other people not related to the reference person.

Housing unit (*HU*)—A group of rooms or a single room occupied or intended for occupancy as separate living quarters. An HU may be occupied by a family or one person, or by two or more unrelated people who share the living quarters. An HU does not need to be a structure; for example, trailers, tents, boats, trucks, buses, and caves, among others, may be considered HUs if they are used as separate living quarters.

Integrated Postsecondary Education Data System (*IPEDS*)—A system of interrelated surveys conducted annually by the U.S. Department of Education's National Center for Education Statistics. IPEDS gathers information from every college, university, and technical and vocational institution that participates in the federal student financial aid programs. The Higher Education Act of 1965, as amended, requires that institutions that participate in federal student aid programs report data on enrollments, program completions, graduation rates, faculty and staff, finances, institutional prices, and student financial aid.

Listing—The field process where interviewers are sent to selected sampled areas to identify and make a list of all HUs for the purpose of drawing a sample of HUs.

Living quarters—All places where people live or stay or could live or stay. There are two types of living quarters: HUs and group quarters.

Master Address File (*MAF*)—The Census Bureau's official inventory of known living quarters and selected nonresidential units in the United States. The file contains mailing and location address information, geocodes, and other attribute information about each living quarters unit. The Census Bureau continues to update the MAF using the USPS DSF and various automated, computer assisted, clerical, and field operations.

Metropolitan statistical area (*MSA*)—A large population center together with adjacent communities that have a high degree of economic and social integration with that center. MSAs are made up of one or more adjacent counties or county equivalents. Some MSAs are defined around two or more population centers; for more information, visit: https://www.census.gov/programs-surveys/metro-micro.html.

Noncity style address—A mailing address that does not use a house number and street or road name. This includes rural routes and box numbers and highway contract routes and box numbers; post office boxes and drawers; and general delivery.

No-Stat address—A business or dwelling under construction, demolished, blighted, or otherwise identified as unlikely to become active for some time, or a rural route address that has not been receiving mail for 90 days or longer.

Post office (PO) box throwback—An address that receives free PO box service at a post office that has no carrier delivery service or at a post office in which the address is within one-quarter mile of the post office.

Rural route—A delivery route served by a USPS rural carrier. A rural carrier is a USPS employee assigned to case, deliver, and collect mail using a vehicle along a rural route and to provide most services available at a small post office. Some mailing addresses in rural areas consist of the combination of a rural route number and a box number.

Screening—An interviewing procedure in which all household members are enumerated before sample selection for the purpose of identifying eligible households or people within households to manipulate the distribution of particular attributes in the sample (for example, to increase the numbers of households including Asian, Black, or Hispanic people).

Street record type—A range of addresses on a street block, block face (one side of a street), cove, cul-de-sac, or other address grouping. Generally, all named or numbered streets with mail delivery or potential for mail delivery have this record type.

Topologically Integrated Geographic Encoding and Referencing (*TIGER*)*/Line*—Publicly available geographic shapefiles that are generated from the Census Bureau's TIGER database. This is a geographic database that automates the mapping and related geographic activities required to support the Census Bureau's census and survey programs.

Tract—A small, relatively permanent statistical subdivision of a county defined by a local committee of census data users for the purpose of presenting data. Census tracts nest within counties, and their boundaries usually follow visible features, but they may follow legal geography boundaries and other nonvisible features in some cases. Census tracts ideally contain about 4,000 people and 1,600 HUs.

Unique ZIP code—A ZIP code that is assigned to a company, government agency, or an entity that receives a high volume of mail from the USPS at one location that is then distributed internally.

Unit frame—A portion of the 2016–2025 NHIS sample frame consisting of addresses subset from a list of addresses provided by a vendor such as the Marketing Systems Group (MSG).

ZIP + N centroids—The latitude and longitude coordinates that approximate the geographic center of a ZIP + N area (where ZIP represents the standard five digit USPS ZIP code). ZIP + N values do not define a geographic entity. However, for the purposes of providing latitude and longitude coordinates for an address, MSG creates geographic entities to approximate a ZIP + N value's area.

Vital and Health Statistics Series Descriptions

Active Series

Series 1. **Programs and Collection Procedures**
Reports describe the programs and data systems of the National Center for Health Statistics, and the data collection and survey methods used. Series 1 reports also include definitions, survey design, estimation, and other material necessary for understanding and analyzing the data.

Series 2. **Data Evaluation and Methods Research**
Reports present new statistical methodology including experimental tests of new survey methods, studies of vital and health statistics collection methods, new analytical techniques, objective evaluations of reliability of collected data, and contributions to statistical theory. Reports also include comparison of U.S. methodology with those of other countries.

Series 3. **Analytical and Epidemiological Studies**
Reports present data analyses, epidemiological studies, and descriptive statistics based on national surveys and data systems. As of 2015, Series 3 includes reports that would have previously been published in Series 5, 10–15, and 20–23.

Discontinued Series

Series 4. **Documents and Committee Reports**
Reports contain findings of major committees concerned with vital and health statistics and documents. The last Series 4 report was published in 2002; these are now included in Series 2 or another appropriate series.

Series 5. **International Vital and Health Statistics Reports**
Reports present analytical and descriptive comparisons of U.S. vital and health statistics with those of other countries. The last Series 5 report was published in 2003; these are now included in Series 3 or another appropriate series.

Series 6. **Cognition and Survey Measurement**
Reports use methods of cognitive science to design, evaluate, and test survey instruments. The last Series 6 report was published in 1999; these are now included in Series 2.

Series 10. **Data From the National Health Interview Survey**
Reports present statistics on illness; accidental injuries; disability; use of hospital, medical, dental, and other services; and other health-related topics. As of 2015, these are included in Series 3.

Series 11. **Data From the National Health Examination Survey, the National Health and Nutrition Examination Surveys, and the Hispanic Health and Nutrition Examination Survey**
Reports present 1) estimates of the medically defined prevalence of specific diseases in the United States and the distribution of the population with respect to physical, physiological, and psychological characteristics and 2) analysis of relationships among the various measurements. As of 2015, these are included in Series 3.

Series 12. **Data From the Institutionalized Population Surveys**
The last Series 12 report was published in 1974; these reports were included in Series 13, and as of 2015 are in Series 3.

Series 13. **Data From the National Health Care Survey**
Reports present statistics on health resources and use of health care resources based on data collected from health care providers and provider records. As of 2015, these reports are included in Series 3.

Series 14. **Data on Health Resources: Manpower and Facilities**
The last Series 14 report was published in 1989; these reports were included in Series 13, and are now included in Series 3.

Series 15. **Data From Special Surveys**
Reports contain statistics on health and health-related topics from surveys that are not a part of the continuing data systems of the National Center for Health Statistics. The last Series 15 report was published in 2002; these reports are now included in Series 3.

Series 16. **Compilations of Advance Data From Vital and Health Statistics**
The last Series 16 report was published in 1996. All reports are available online; compilations are no longer needed.

Series 20. **Data on Mortality**
Reports include analyses by cause of death and demographic variables, and geographic and trend analyses. The last Series 20 report was published in 2007; these reports are now included in Series 3.

Series 21. **Data on Natality, Marriage, and Divorce**
Reports include analyses by health and demographic variables, and geographic and trend analyses. The last Series 21 report was published in 2006; these reports are now included in Series 3.

Series 22. **Data From the National Mortality and Natality Surveys**
The last Series 22 report was published in 1973. Reports from sample surveys of vital records were included in Series 20 or 21, and are now included in Series 3.

Series 23. **Data From the National Survey of Family Growth**
Reports contain statistics on factors that affect birth rates, factors affecting the formation and dissolution of families, and behavior related to the risk of HIV and other sexually transmitted diseases. The last Series 23 report was published in 2011; these reports are now included in Series 3.

Series 24. **Compilations of Data on Natality, Mortality, Marriage, and Divorce**
The last Series 24 report was published in 1996. All reports are available online; compilations are no longer needed.

For answers to questions about this report or for a list of reports published in these series, contact:

Information Dissemination Staff
National Center for Health Statistics
Centers for Disease Control and Prevention
3311 Toledo Road, Room 4551, MS P08
Hyattsville, MD 20782

Tel: 1–800–CDC–INFO (1–800–232–4636)
TTY: 1–888–232–6348
Internet: https://www.cdc.gov/nchs
Online request form: https://www.cdc.gov/info
For e-mail updates on NCHS publication releases, subscribe online at: https://www.cdc.gov/nchs/email-updates.htm.

U.S. DEPARTMENT OF HEALTH & HUMAN SERVICES

Centers for Disease Control and Prevention
National Center for Health Statistics
3311 Toledo Road, Room 4551, MS P08
Hyattsville, MD 20782–2064

For more NCHS Series Reports, visit:
https://www.cdc.gov/nchs/products/series.htm

Series 2, No. 191
CS330163