Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Balancing sample design goals for the National Health and Nutrition Examination Survey

by Leyla Mohadjer and Lester R. Curtin

June 2008

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2008
Vol. 34, No. 1, pp. 119-126
Statistics Canada, Catalogue No. 12-001-X

Balancing sample design goals for the National Health and Nutrition Examination Survey

Leyla Mohadjer and Lester R. Curtin [1]

Abstract

The National Health and Nutrition Examination Survey (NHANES) is one of a series of health-related programs sponsored by the United States National Center for Health Statistics. A unique feature of NHANES is the administration of a complete medical examination for each respondent in the sample. To standardize administration, these examinations are carried out in mobile examination centers. The examination includes physical measurements, tests such as eye and dental examinations, and the collection of blood and urine specimens for laboratory testing. NHANES is an ongoing annual health survey of the noninstitutionalized civilian population of the United States. The major analytic goals of NHANES include estimating the number and percentage of persons in the U.S. population and in designated subgroups with selected diseases and risk factors. The sample design for NHANES must create a balance between the requirements for efficient annual and multiyear samples and the flexibility that allows changes in key design parameters to make the survey more responsive to the needs of the research and health policy communities. This paper discusses the challenges involved in designing and implementing a sample selection process that satisfies the goals of NHANES.

Key Words: Multistage Sampling, Domain Sampling, Weighted Measure of Size, Mobile Examination Centers.

1. Introduction

The National Health and Nutrition Examination Survey (NHANES) is one of a series of health-related programs sponsored by the United States Centers for Disease Control and Prevention through its National Center for Health Statistics (NCHS). The NHANES surveys have been used to assess the health and nutritional status of the non-institutionalized civilian population of the United States for over 45 years. The data collected in NHANES are used to estimate the prevalence of major diseases and risk factors for diseases. The nutritional data from NHANES provide temporal monitoring for the nation with respect to such factors as diet, cholesterol, hypertension, iron deficiency, anemia, and obesity. NHANES has also been designed to assess the relationship between diet, health, and the environment so that nutritional assessments can be linked to such diseases as cardiovascular disease, diabetes, hypertension, and osteoporosis.

Data collection for NHANES includes at least three stages: a household screener, an interview, and a medical examination. The primary objective of the screener is to determine whether any household members are eligible for the interview and examination. The screener collects basic information on household composition and demographic characteristics. The interview collects household-, family-, and person-level data on demographic and socioeconomic background, health, and nutritional characteristics. Upon completion of the interview, respondents are asked to participate in a medical examination. To standardize administration and protocols, these examinations are carried out in a specially designed and equipped mobile examination center (MEC). The examination includes physical measurements, tests such as eye and dental examinations, physiological measurements, and the collection of blood and urine specimens for laboratory testing. The NHANES website (http://www.cdc.gov/nchs/ nhanes.htm) provides detailed information about the NHANES medical components.

The development of an efficient design has involved consideration of several design issues unique to NHANES in addition to the ones normally involved in survey samples. This paper is focused on the unique and challenging aspects of the NHANES design. However, it is helpful to provide an overall summary of the NHANES design, as given below, before discussing the unique features.

The NHANES sample represents the total non-institutionalized civilian population of the United States. Active military and institutionalized persons are not part of the population of inference. NHANES is not an equal probability design; sampling fractions are set to "over-sample" Mexican Americans (and Hispanics in the 2007 and beyond samples), black Americans, low-income white/other Americans, persons below age 20, and persons above age 60. A four-stage sample design is being used. The primary sampling units (PSUs), often referred to as stands, are selected from a frame of all U.S. counties. The PSUs are mostly single counties; in a few cases, adjacent counties are combined to keep PSUs above a certain minimum size. There are close to 3,000 PSUs in the NHANES sampling frame. NHANES PSUs are selected with probabilities proportionate to a measure of size (PPS). There are 15 stands in each annual sample.

The second sampling stage is area segments comprising Census blocks or combinations of blocks. Because PSUs

1. Leyla Mohadjer, Westat, 1650 Research Blvd., Rockville, Maryland, U.S.A. 20850. E-mail: LeylaMohadjer@Westat.com; Lester R. Curtin, National Center for Health Statistics, 3311 Toledo Road, Hyattsville, Maryland, U.S.A. 20782. E-mail: lrc2@cdc.gov.

vary in size, there is some variability in segment size and in the number of segments per PSU. Segments are formed with an average of about 150 households (or dwelling units) per segment. An average of about 5,000 segments are created within each PSU, and an average of 24 segments are sampled. The sample is designed to produce approximately equal sample sizes per PSU, and most PSUs have exactly 24 segments. The segments are also selected with PPS. The measures of size (MOS) of the segments, when combined with the subsampling rates used within the segments, provide approximately equal numbers of sample persons (SPs) per segment, although the relative variation in workload is greater among segments than among PSUs.

The third stage of sample selection consists of households and noninstitutional group quarters, such as dormitories. In a given PSU, following the selection of segments, all dwelling units (DUs) in the sampled segments are listed, and a subsample of households and group quarters within the DUs are designated for screening in order to identify potential SPs for interview and examination. SPs within the households or group quarters are the fourth stage of sample selection. All eligible members within a household are listed, and a subsample of individuals is selected. The subsampling rates for households within segments and for individuals within households are determined in advance. The combination of screening and differential sampling rates provides the increased sample size for those demographic subdomains of special interest (age, sex, race/ethnicity, and income). For example, in the 30 PSUs in which data were collected during the 2-year data cycle 2005-2006, 716 segments were selected and 26,529 households were selected for screening. After being screened for age, sex, and race/ethnicity composition and low-income status, 6,372 households had one or more individuals selected into the sample. A total of 12,862 individuals were selected, of whom 9,950 completed interviews and examinations.

The NHANES examination requires both highly specialized personnel and laboratory processing of collected specimens. As a result, examination components can be very costly to implement. To limit costs and reduce respondent burden, certain examination components are administered to only a subsample of MEC respondents. A single subsampling algorithm controls the amount of overlap among the various subsamples to allow analyses of correlations between various examinations and laboratory components. The SP's assignments to subsamples are fully determined before the SP arrives at the MEC.

The data collected in NHANES surveys have been extremely important in providing needed information about the health and nutritional status of the U.S. population. As a result, beginning with NHANES 1999, the survey has been implemented as a continuous, ongoing, annual survey (Montaquila, Mohadjer and Khare 1998). It is critical to devote a lot of attention to the development and maintenance of an efficient sample design for such an important and complex survey. This paper discusses the challenges involved in designing and implementing a sample selection process that satisfies the multiple goals of NHANES. The paper focuses on the sample design used through 2006 (in response to emerging analytical requirements, some aspects of the sample design changed starting in 2007).

Section 2 outlines the major purposes and goals of the survey, followed by an overview of the major factors affecting the design given in Section 3. The unique features of the NHANES sample design are described in Section 4. Finally, Section 5 provides a brief summary of the paper.

2. Major purposes and goals of NHANES

NHANES is an ongoing annual health survey of the noninstitutionalized civilian population of the United States. The main objectives of NHANES are to (1) estimate the national prevalence of selected diseases and risk factors; (2) estimate national population reference distributions of selected health parameters and environmental contaminants; (3) document and investigate reasons for secular trends in selected diseases and risk factors; (4) contribute to the understanding of disease etiology; (5) investigate the natural history of selected diseases; (6) study the relationship between diet, nutrition, environment, genetics, and health; and (7) explore emerging public health issues.

3. Major factors affecting sample design

As mentioned above, a unique feature of NHANES is the complete medical examination carried out in the MECs. In addition, the design needs to produce efficient sample sizes for a large number of subdomains of the general population. Many health and nutritional characteristics differ considerably by age, sex, and race/ethnicity and are also affected by income status. As a result, most analyses of NHANES data are conducted for defined age categories within various socioeconomic subgroups of the population. Therefore, the survey is designed to produce efficient sample sizes for a very large number of subdomains of the U.S. population.

In general, the sample design for NHANES must create a balance between the requirements for efficient subdomain samples and the need for an efficient workload for examination staff at the MEC, while keeping response rates as high as possible. More specifically, the NHANES design attempts to (1) obtain prespecified self-weighting sample sizes for a set of about 75 predesignated subdomains; (2)

produce sample sizes per PSU that will result in an efficient workload for the interview and examination staff at the MEC; (3) design samples that are likely to achieve high response rates; (4) be as cost effective as possible; (5) produce efficient annual samples; (6) allow for accumulation of samples, especially for rare subdomains or rare diseases over time; and (7) be flexible to allow changes in key parameters, including sampling domains, and sampling rates to respond to emerging health issues.

In the remainder of this section, we provide brief summaries of how each of these seven goals affects the design and implementation of NHANES.

NHANES subdomains - The sample design for NHANES meets a prespecified level of precision for cross-sectional data and comparisons over time for a set of predesignated subdomains. Specifically, 77 sampling domains (in the 2006 sample) are defined by race/ethnicity, sex, age, income, and pregnancy status. The sample includes oversamples of blacks, Mexicans, the very young, adolescents, the elderly, pregnant women, and the low-income population.

When estimates of universe totals for the entire population are considered to be of the greatest importance, then the best available estimate of the total population is used as an MOS in the sample selection process. For NHANES, where the interest is in subdomains of the total population, an alternative MOS is needed to improve the accuracy of the estimates and provide better control of the sample size. Section 4 describes the MOS used for sampling PSUs and segments in NHANES.

The objective of oversampling (using differential probabilities of selection) is to achieve a sample containing proportionately more members of certain population subdomains than there are in the population. The goal is to obtain adequate sample sizes to make inferences for subdomains representing relatively small proportions of the total universe of interest and to do it in such a way as to minimize variances for the budget available for the survey. Different oversampling strategies are used depending on the domains of interest. For example, oversampling of the minority subpopulations is accomplished through stratifying geographic areas by concentration of these minority groups and selecting segments in high-density areas at a higher rate. On the other hand, a large screening sample may be required to oversample persons within specific age categories. The subsection on Cost Ratios below describes why oversampling procedures used in NHANES are different from those commonly used in many area frame sample surveys.

Workload for mobile examination centers (*MECs*) - The MEC consists of four specially designed and equipped trailers and contains all of the medical equipment. Each trailer is approximately 45 feet long and 10 feet wide. Detachable truck tractors drive the trailers from one location to another. MECs travel to survey locations throughout the country. The trailers are set up side by side and connected by enclosed passageways. The area in the MEC is divided into rooms to allow privacy during the examinations and interviews. The examination includes a variety of physical and dental assessments and measurements, laboratory tests, and health interviews.

Because of the logistical issues related to the traveling MECs, the sample size in each sampled location must be derived ahead of time and considered fixed so that field operations can be scheduled in an efficient and manageable way. Also, it is necessary to establish a firm time schedule for each stand so that appointments can be made for examinations. It is not possible to change the time schedule since it must be coordinated with the MEC's visits to other stands, which are also planned in advance.

Response rates - Achieving high response rates is a concern for practically every sample survey. With NHANES, this is a particular challenge because of the extensive nature of the interviews and examinations. Remunerations have been used in NHANES as a means of improving response rates. In addition, NHANES has an extensive outreach program that includes contacts with local organizations and individuals to gain cooperation, as well as local media coverage to reach as many SPs as possible. As a sample design issue, one approach that has been proven to favorably affect response rates is selecting larger sample sizes within sampled households. One of the factors thought to be responsible for the increased response rates in multiple-SP households is that each person is given remuneration for his or her time and participation, and it is generally more convenient for household members to come to the MEC at the same time. Table 1 shows the examination response rates for SPs coming from households where only one person was selected compared to the response rates for SPs coming from multiple-SP households. As the table indicates, response rates increase by about 4 to 7 percent depending on the type of household.

NHANES is, therefore, designed to maximize the number of SPs per household. Such an approach is feasible for studies like NHANES, where the sample is composed of a large number of subdomains. That is, the effect of within-household clustering is not a large concern for NHANES because most analyses are done within age-sex-specific subdomains (or some limited groups of subdomains) and there is generally little within-household clustering at the subdomain level. The average number of SPs selected per household (in households where at least one SP was selected) within the defined sampling domains ranges from 1 to 1.24 in the 1999-2006 sample. Combining the domains down to 12 to 15 domains by collapsing over age and/or race/ethnicity will result in average numbers ranging from 1.01 to 1.37 SPs per household. Therefore, some level of clustering is present to the extent that collapsed domains are used for analysis. Note that the SP sample is basically used for SP-level analysis (*e.g.*, health and nutrition statistics). The clustering of SPs is, of course, higher at the family and household levels. However, household- or family-level variables are used for such analysis (*e.g.*, household dust levels, family income, or insurance). Refer to Curtin and Mohadjer (2008) for a discussion of the impact of clustering, and unequal probabilities of selection of subdomains, on the precision levels of various estimates.

Table 1
Examination response rates by number of SPs in household, by household type, in 1999-2006 NHANES sample

Household type	Number of SPs selected per household		Response rate (%)	
	One SP	Two or more SPs	One SP	Two or more SPs
Black/Mexican	4,892	20,222	76.5	82.3
Other low-income[1]	1,362	3,349	77.6	84.5
Other non-low-income	5,597	15,508	68.8	72.6

[1] The Other group includes all SPs who are not Black or Mexican. The low-income threshold is set at 130 percent of poverty.

Cost ratios - The field data collection cost in area survey samples includes the cost of listing DUs, screening households to locate eligible respondents, and conducting the interview to collect data. In NHANES, the data collection phase includes both the household interview and the MEC examination. NHANES requires highly specialized medical equipment, personnel, and laboratory processing. As a result, the cost of an examination is very high compared to other costs in the survey. In fact, the cost of listing and screening is only about 3 to 4 percent of the cost of interviewing and examination. This cost ratio (the cost of interviewing and examination relative to the cost of listing and screening) greatly affects the design of NHANES.

As mentioned above, many of the predesignated subdomains of NHANES require some method of oversampling to achieve the required sample sizes. For the minority populations, substantial reductions in screening are possible with oversampling of highly concentrated minority areas. In general, an optimum design is developed by ascertaining the effect on cost and variance of alternative sampling procedures and choosing the one that minimizes the variance for a fixed cost. In the evaluation of trade-offs between cost and variance, suppose that a particular oversampling strategy reduces the number of households to be listed and screened while increasing the variance for most statistics. The savings brought about by the reduction in cost of listing and screening could be used to increase the size of the sample and thereby lower the variance. However, in NHANES, listing and screening a household is only a very small fraction of the cost, and thus, it takes very large savings in listing and screening costs to justify a moderate increase in variance. As a result, the oversampling procedures established for the survey reflect the NHANES cost ratio and are different from those of typical area surveys.

Annual and multiyear samples - To facilitate potential linkage with other large-scale surveys, to retain flexibility in the sample design, and to allow for the production of annual estimates for broad subdomains, NHANES became a continuous, annual survey starting in 1999. The travel requirements for nationally representative annual samples in the United States are challenging. Three MECs – two of which are stationed at PSUs and one of which is traveling at any given time – work on a very carefully designed schedule to meet the design requirements of the study.

The ability to make meaningful inferences from any survey is affected by both the precision of the estimates themselves and the precision of the variances of the estimates used in the analysis. One of the main limitations of an NHANES annual sample is the small number of PSUs (15 per year), which results in a small number of degrees of freedom for both estimation and analysis and thus design-based variance estimates that are relatively imprecise. Additionally, the effective sample sizes for most subdomains are too small in annual samples. Most subdomain analyses will need to accumulate a number of annual samples to provide both precision and statistical power for comparisons. The procedures for combining years of the survey must be relatively simple, and appropriate for commercial software packages, to maximize the usefulness to the wide variety of users of the NHANES data. Thus, it is critical to employ a sample design that allows efficient accumulation of the annual samples across years.

Flexible design - A critical objective of NHANES is to explore emerging public health issues. The survey needs to be flexible and able to adapt to changing requirements and new challenges. Thus, the sample design must balance the need for efficient subdomain samples with the flexibility needed to make changes in key parameters. To date, the current NHANES design has been able to incorporate some changes in subdomain definitions and sampling rates when these changes have been made after the selection of PSUs. However, in extreme circumstances, substantial changes in subdomain definitions or sample size requirements would necessitate the selection of a new PSU sample.

4. Unique features of the NHANES design

The factors described in Section 3 have played major roles in the development of the sample design and have resulted in some design features that are unique to NHANES. The unique features of the sample design include (1) weighted PSU and segment MOS; (2) efficient annual and multiyear samples; (3) maximized number of SPs per household; (4) controlled sample sizes for PSUs; (5) sequential release of the PSU sample; (6) special methods to deal with deterioration of the efficiency of the optimum design over time; and (7) special methods to reduce the risk of data disclosure through geographic identification.

The following paragraphs briefly describe these unique features of the NHANES design.

Clustering and measures of size (*MOS*) - In NHANES, the sample size must be large enough to produce an efficient workload for each PSU, considering the time and the cost involved in moving a MEC between survey locations and the time required to set up and break down the MECs for travel. Experience gained in earlier NHANES surveys has indicated that an average of 340 examined SPs is an approximately optimum number that provides the maximum number of PSUs while keeping the sample size in each area large enough to justify the costs associated with moving the MECs. In addition, the PSUs for NHANES are typically defined as individual counties to reduce the amount of travel necessary for respondents to visit a MEC, and thereby increase the likelihood of achieving high response rates.

The NHANES sample is designed to yield a self-weighting sample for each sampling subdomain while producing an efficient workload in each PSU. PSUs and segments are selected with probabilities proportionate to a weighted MOS, reflecting the PSU population in subdomains of interest. The selection probability of a PSU determines the maximum rate at which persons residing in that particular PSU can be selected. Refer to *Vital and Health Statistics, Series 2*, *No. 113*, *September 1992*, *CDC/NCHS*, available at http://www.cdc.gov/nchs/products/pubs/pubd/series/sr02/120-101/120-101.htm, for a description of the MOS used in NHANES.

Annual and multiyear samples and stratification - One way to achieve nationally representative annual samples is to select an independent sample of PSUs each year. Because of the limited number of NHANES PSUs and the fact that PSUs are selected proportionate to size, this approach would be likely to lead to substantial overlap in PSUs from year to year. Sample overlap, even at the PSU level, could lead to loss of precision in survey estimates when survey years are combined (due to increased clustering of the sample). Thus, rather than sampling PSUs independently each year, the approach in NHANES has been to select a 6-year sample, from a nested structure of major and minor strata (as described below), and then allocate one PSU from each major stratum to each year. This nested structure for the 6-year sample avoids overlap of non-self-representing PSUs during the 6 years.

The design for the NHANES 6-year sample is a stratified two-PSU-per-stratum design and has been developed with the primary goal of efficiency for the 6-year sample, as well as efficient multiyear samples. The stratification scheme is designed to ensure that the PSUs comprising the annual and multiyear samples are distributed evenly in terms of geography and certain population characteristics.

The NHANES design (through 2006) included 18 self-representing PSUs. These PSUs ranged from those that were self-representing for the annual samples to those that were self-representing for 3-year or 6-year samples. These PSUs were assigned such that each year had an equal number of self-representing PSUs, with 3-year self-representing PSUs being 3 years apart. The non-self-representing PSUs were stratified into 12 major strata, defined based on geography and the metropolitan statistical area status of the PSUs. Seventy-two minor strata were defined based on the demographics of the PSUs. The minor strata were constructed to be of equal size to the extent possible (in terms of total MOS). The variables used to form the boundaries of the minor strata were minority status and the percentages of the population below poverty level. Each major stratum included six minor strata, and one PSU was selected from each of these final strata. Within each major stratum, minor strata were paired to create pseudo-strata. Each pair was randomly assigned to the study 3 years apart. The assignment of the pairs to the particular sets of study years and the assignment of the study years within the pair were random within the first major stratum, and all other major strata followed the same pattern.

This stratification scheme resulted in a sample of 72 non-self-representing PSUs that produces efficient annual and multiyear estimates without compromising the efficiency of the 6-year estimates. The 6-year sample has a one-PSU-per-minor-stratum design (or a two-PSU-per-pseudo-stratum design), and each annual sample has a one-PSU-per-major-stratum design. In addition, this design allows for the flexibility needed to address changes in the sample requirements (if a new sample needs to be selected), since the first 3 years of the sample follow a one-PSU-per pseudo-stratum design.

Maximized number of SPs per household - After the sample of screened households is identified, a sample of persons to be interviewed and examined from individual households is selected. All eligible members within a household are listed, and a subsample of individuals is selected based on sex, age, race/ethnicity, and income (all pregnant women are selected with certainty). SPs are selected at rates established to ensure that the target sample sizes by subdomains will be achieved.

The sample of SPs is selected in a way that maximizes the average number of SPs per household in order to increase the overall response rate in the survey. If independent random selections are made for the subdomains, in most cases only one person in a household would be selected and the average sample size per household would be quite low, not much above 1. Therefore, instead of unrestricted randomization, a pseudo-random procedure is used that maximizes the number of SPs per households. Refer to Waksberg and Mohadjer (1991) for a description of the approach.

Controlled sample sizes per PSU - The sample size in each PSU (stand) that is actually generated from a self-weighting sample in each domain is based on a number of assumptions such as the age and race/ethnicity distribution in the PSU. These assumptions hold only approximately. Once the sample sizes have been calculated, they are treated as quotas, and the number of SPs in each stand is forced to adhere closely to the quota. The reason for this procedure is to have a manageable and efficient field operation. It is necessary to establish a firm, and fixed, time schedule for each stand so that appointments can be made for SP examinations. The time schedule obviously takes into account the expected number of SPs in each stand. As mentioned above, it is difficult to change the time schedule for a stand since it must be coordinated with the MEC's visits to other stands, which are also planned in advance.

There is no way of knowing in advance whether the assigned quota for a particular stand is lower or higher than what would arise from self-weighting samples within the various domains. Part of the reason for the uncertainty is that the MOS used for sample selection is based on the latest decennial Census and may not be quite up to date. The issue is further complicated by variations in response rates from stand to stand, as well as sampling variation in the number of identified SPs. Consequently, it is necessary to use a sample selection procedure that can produce samples that are either somewhat larger or somewhat smaller than those arising from the application of the self-weighting sampling rates.

Sequential release of the sample in each stand - To accomplish the above objective, an initial sample is selected in each stand that uses sampling rates 50 percent larger than those required to attain the target sample sizes in each domain. Each stand's initial sample is then divided into a group of subsamples. Each subsample is a systematic subsample of the initial sample, with the households sequenced by segment number and a temporary, geographically based sequence number prior to subsampling. Thus, each subsample cuts across all segments, except when limited by sample size.

As a general rule, the 50 percent subsample (*i.e.*, subsample A) is released to the interviewers first. The yield from this subsample is monitored and used to project estimates of the total number of SPs expected when screening of this subsample has been completed. Based on these figures, additional subsamples are released as needed. The sample is monitored on a daily basis to determine whether additional subsample releases are required.

The one operational problem with the procedure for monitoring the sample yield is that it cannot completely control the subdomain sample sizes. The distribution of subdomains differs, to some extent, from the expected numbers based on the most recent Census data (used to derive the sampling rates). Experience with NHANES indicates that some population changes that will affect the sample sizes can be expected. Other factors that affect subdomain sample yield are patterns of nonresponse and undercoverage in stands. One option to correct the shortfall (or overage) in subdomain sample sizes is to change the sampling rates in future stands. However, such changes will increase heterogeneity in sample weights, thus adversely affecting the precision of the subdomain estimates, and are not advisable except under extreme circumstances.

Dealing with deterioration of the efficiency of the optimum design over time in a tightly controlled sample - The usual practice in area samples is to list all households in the sample segments and apply a prespecified sampling rate to the listed households. This approach gives all households the desired probabilities of selection. For example, if the sampling rate is 50 percent, then one-half of the housing units listed in the segments will be included in the sample. If the number of housing units has tripled due to new construction (*i.e.*, housing units built since the most recent decennial Census), the same sampling rate will produce three times as many interviews and examinations as the number originally expected. Such dramatic changes in the segment size are expected when the data collection period is several years after the most recent decennial Census for which data files are available.

For NHANES, highly variable sample sizes are not feasible because of the scheduling requirements of the MECs. Subsampling within PSUs, in an effort to obtain equal sample sizes across PSUs, is not recommended either, because it will introduce unequal weighting factors that would reduce the efficiency of the sample.

NHANES has used two procedures to update the segment MOS: (1) creation of new construction segments and (2) two-phase sampling to update the MOS. A third approach under consideration involves using purchased commercial address listings to update the MOS in a two-phase sample design.

Under the new construction approach (Bell, Mohadjer, Montaquila and Rizzo 1999), newly constructed units are excluded from area segments and new segments are created based on U.S. Census Bureau information on permits issued for new construction since the most recent decennial Census. New construction segments comprise clusters of building permits issued during one or several adjoining months by a building permit office. Census Bureau files from the Building Permits Survey are used as sources of the data on the number of residential building permits issued by the building permit offices.

Two-phase sampling is used in a number of statistical applications. One of the applications of two-phase sampling is to update a sampling frame when the sample is to be selected with respect to an MOS but a reliable estimate of the MOS is not available. With this approach, a larger sample of units (segments, in the case of NHANES) is selected. An updated value of MOS is then collected for this larger sample (also referred to as the first-phase sample). The final sample of units (segments) is selected from the first-phase sample using the updated MOS.

Starting in 2000, the NHANES segment MOS has been updated (for stands for which such updating seemed necessary) using a two-phase sampling procedure (Montaquila, Bell, Mohadjer and Rizzo 1999). In these cases, listers travel to the stand to obtain a count of the number of DUs in each segment in the first-phase sample. Using the listers' counts, an updated MOS that reflects the ratio of the actual number of DUs to the expected number of DUs is calculated for each first-phase segment. The final sample of segments is then selected by subsampling from the first-phase segments using the updated MOS.

Risk of data disclosure through geographic identification - In today's world, confidentiality concerns and the risk of data disclosure present real challenges to survey sponsors. The ability to identify survey respondents, either through unique combinations available on a single data file or by linking different databases, is of great concern. This is particularly true for NHANES, because of the extensive amount of sensitive data collected on each SP and the small number of PSUs in the sample. Therefore, NHANES evaluates the risk of disclosure on two fronts: geographic disclosure and disclosure from individual characteristics. Various methods (limited or suppressed data release) are used by NCHS to mask the individual characteristics that have a high risk of identifying individuals in the NHANES sample. Sensitive, limited, or non-released data items are available through a Research Data Center. At this time, only national estimates can be produced from publicly available data files; detailed geographic analyses must be done in the Research Data Center.

Although only national estimates can be produced, the direct estimation of sampling errors for those national estimates requires the release of design variables such as stratum and PSU identifiers. Typically, these variables indicate that a group of SPs are all in the same county but do not identify that county. Geographic disclosure is of a particular concern because (1) NHANES has a small number of PSUs, (2) PSUs are limited in geography to one county, and (3) an extensive amount of outreach activity is conducted within each PSU to improve response rates. The outreach program includes contacting various organizations and individuals at each stand to seek their support and using media (newspapers, television, and radio) to reach as many SPs as possible. It is therefore relatively easy to determine the counties in the NHANES sample. The racial/ethnic composition of a county, along with metropolitan/non-metropolitan status, is enough information to correctly match a list of known counties with groups identified as a county cluster on the public data file. To limit geographic disclosure, probabilistic record swapping methods are used at the second stage of sampling (segment swapping) to create masked variance units. The goal is to reduce the risk of identifying individuals by masking their location. Refer to Park, Dohrmann, Montaquila, Mohadjer and Curtin (2006) for a description of the swapping procedures applied to the NHANES sample.

5. Summary and conclusions

A unique feature of NHANES is the complete medical examination carried out in the MECs. In addition, the survey is designed to produce efficient sample sizes for a large number of subdomains of the U.S. population, since most analyses of NHANES data are conducted for defined age categories within various socioeconomic subgroups of the population. Thus, the sample design for NHANES must create a balance between the requirements for efficient subdomain samples and the need for an efficient workload for examination staff at the MEC, while keeping response rates as high as possible. In addition, the design must be as cost effective as possible, produce efficient annual samples, and allow for accumulation of samples for rare subdomains or rare diseases over time. Furthermore, the design must be flexible to allow for changes in key parameters, including sampling domains, and sampling rates to respond to emerging health issues.

The above requirements result in a very complex design with some design features that are unique to NHANES. In particular, the current sample is designed to produce efficient annual and multiyear samples. NHANES uses weighted PSU and segment MOS to yield self-weighting samples for each subdomain, while producing an efficient workload in each PSU. Once the sample sizes are calculated, they are treated as quotas. The sample sizes are strictly controlled in each PSU to create a manageable and efficient field operation. A very large screening sample is used to oversample most of the age and income subdomains, and oversampling of highly concentrated areas is used for some of the very rare minority subdomains. The sample of SPs is selected using a pseudo-random procedure to maximize the average number of SPs per household because it has appeared to increase the overall response rate in previous surveys.

The challenges described in this paper are focused on the main aspects of the NHANES. There remain many other features unique to NHANES that analysts must take into account when analyzing data from the survey. For example, not only are there very few PSUs in each annual sample, but data collected within these PSUs are not randomly collected across the seasons. In particular, if there is a seasonal by geographic region interaction for a variable of interest, the current NHANES design will not be able to estimate it. Because of the small number of PSUs in each data release cycle, any contextual data linkage at the geographic level must be done in the NCHS Research Data Center. Because of the many subsamples within NHANES, special care must be taken to use the appropriate subsample weight; for example, estimates for undiagnosed diabetes must use the special fasting weight.

To facilitate the efficient use of MECs for data collection, there has been no attempt to randomly allocate the sample of PSUs across time in annual samples. However, the time dimension plays a major role in some health indicators, such as nutrition. Furthermore, analysis of nutrition data may also be affected by the complex nature of the design and data collection. Special sample weights constructed for the 2 days of the 24-hour recall data account for variation in the number of examinations by day of the week. A web-based tutorial is now being developed to provide assistance in the analysis of NHANES nutrition data. A general tutorial for design-based analysis of NHANES data can be found at http://www.cdc.gov/nchs/tutorials/.

Acknowledgements

The authors are grateful to the associate editor and the referees for their helpful comments and suggestions, which have greatly improved the paper.

References

Bell, B., Mohadjer, L., Montaquila, J. and Rizzo, L. (1999). Creating a frame of newly constructed units for household surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Curtin, L.R., and Mohadjer, L. (2008). Design trade-offs for the National Health and Nutrition Examination Survey. *Proceedings of the Ninth Conference on Health Surveys Research Methods*, to appear.

Montaquila, J., Bell, B., Mohadjer, L. and Rizzo, L. (1999). A methodology for sampling households late in a decade. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Montaquila, J., Mohadjer, L. and Khare, M. (1998). The enhanced sample design of the future National Health and Nutrition Examination Survey (NHANES). *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Park, I., Dohrmann, S., Montaquila, J., Mohadjer, L. and Curtin, L.R. (2006). Reducing the risk of data disclosure through area masking: Limiting biases in variance estimation. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Waksberg, J., and Mohadjer, L. (1991). Automation of within-household sampling. *Proceedings of the Survey Research Methods Section*, American Statistical Association.