# THE ENHANCED SAMPLE DESIGN OF THE FUTURE NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY (NHANES)

**Jill M. Montaquila, Leyla Mohadjer, Westat; Meena Khare, NCHS**
**Jill M. Montaquila, Westat Inc., 1650 Research Boulevard, Rockville, Maryland 20850**

**Key Words:** Survey Integration, Sample Design, Annual Samples.

## 1. Introduction

The National Health and Nutrition Examination Survey, NHANES, is taking a new direction, beginning in 1999. The major differences from previous Health and Nutrition Examination Surveys are that the future NHANES will be implemented as a continuous, annual survey and that it will be linked to related Federal government data collections, in particular, the National Health Interview Survey (NHIS) and, potentially the U.S. Department of Agriculture's food consumption surveys.

The possibility of using the NHIS as a sampling frame for the NHANES survey has been under consideration for many years. Plans for the integration of periodic surveys became a higher priority issue for NCHS after a National Academy of Sciences (NAS) panel strongly urged such consideration.

### 1.1 The NHANES Survey Program: History and Highlights

The National Health and Nutrition Examination Surveys are one of the series of health related programs conducted by NCHS. A unique feature of these surveys is the collection of health data by means of medical examinations carried out for a nationally representative sample of the U.S. population. NHANES I, the first cycle of NHANES, was conducted from 1971 to 1975. The second National Health and Nutrition Examination Survey (NHANES II) was conducted from 1976 to 1980.

Although NHANES I and NHANES II each examined in excess of 20,000 individuals, the representation of the minority population in the sample was not large enough to adequately estimate the health status of Mexican-Americans, Cuban-Americans, or Puerto Ricans, or even the three groups combined. The objective of the Hispanic Health and Nutrition Examination Survey (HHANES), conducted from 1982 to 1984, was to produce estimates of health and nutritional status for the three major Hispanic subgroups that were comparable to the estimates available for the general population.

NHANES III was fielded between 1988 and 1994. (For details, see DHHS, 1996.) In the tradition of the past national surveys, it continued to be a keystone in providing critical information on the health and nutritional status of the U.S. population. This information is essential for estimating the prevalence of various diseases and conditions, explaining the mechanisms of disease development, and planning for health policy. NHANES III provides information for more than a dozen individual agencies and, thus, already reflects coordination within the Department of Health and Human Services (DHHS) in the collection of direct physical measurement data.

Data collection for NHANES comprises three levels: 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 interview collects household and person level data on health and nutrition characteristics. The examination includes physical measurements, tests such as eye exams and dental exams, and blood and urine specimens used to obtain laboratory measurements.

### 1.2 Survey Objectives and Constraints

The analytic goals of the NHANES are as follows:

- To estimate the number and percent of persons in the U.S. population and designated subgroups with selected diseases and risk factors;
- To monitor trends in the prevalence, awareness, treatment, and control of selected diseases;
- To monitor trends in risk behaviors and environmental exposures;
- To analyze risk factors for selected diseases;
- To study the relationship between diet, nutrition, and health;
- To explore emerging public health issues and new technologies; and
- To establish a national probability sample of genetic material for future genetic testing.

As mentioned earlier, unique to NHANES are the complete medical examinations for each respondent in the sample. In order to standardize their administration, these examinations are carried out in mobile examination centers (MEC). Considering the time and the cost involved in moving a MEC between survey locations, the sample size per primary sampling unit (PSU) must be large enough to produce an efficient workload at each PSU. In addition, to reduce the

amount of travel necessary for respondents to visit a MEC and thereby have the survey achieve as high a response rate as possible, the PSUs for NHANES are defined as individual counties.

The set of domains for which specified reliability is desired in NHANES consists of age-sex groups for blacks, Mexican-Americans, and the remainder of the U.S. population. In order to increase precision of estimates for certain subdomains, oversampling will be carried out for adolescents, the elderly, Mexican-Americans, and the black population. In addition, the future NHANES will include a national sample of pregnant women.

The future NHANES has two new design features aimed at enhancing the analytic utility of the survey. The first is the requirement for nationally representative annual samples. The second is linkage to other national health and nutrition surveys, beginning with the NHIS.

## 2. Methodology for Sample Selection

The future NHANES sample will represent the total civilian noninstitutionalized population in the 50 states of the United States and the District of Columbia. A four-stage sample design will be used: (1) Primary Sampling Units (PSUs) comprising mostly single counties, (2) segments within PSUs, (3) households within segments, and (4) persons within households. In order to reduce the amount of screening required, segments are stratified according to the proportion of the population that is Mexican-American. (These strata are referred to as "density strata.") Higher sampling rates are used to sample within the density strata with higher proportions of Mexican-Americans.

The NHANES sample is designed to yield a self-weighting sample for each sampling domain while producing an efficient workload for each PSU. PSUs are selected with probabilities proportionate to a measure of size (MOS). The selection probability of a PSU determines the maximum rate at which persons residing in that particular PSU can be selected. For NHANES, the PSU MOS is defined to be

$$M_h' = \frac{M_h}{\pi_h} = \frac{1}{\pi_h} \sum_k A_k C_{hk}' ,$$

where

$$A_k = \sum_{i,\ell} P_{ik} r_{ik\ell} \frac{C_{\cdot\cdot k\ell}}{C_{\cdot\cdot k\cdot}} ;$$

$h$ = NHANES PSU;
$i$ = Mexican-American density stratum;
$k$ = race/ethnicity subdomain;
$\ell$ = age/sex subdomain;
$\pi_h$ = selection probability for the NHIS PSU containing NHANES PSU $h$;
$r_{ik\ell}$ = overall sampling rate in density stratum $i$ for the $(k,\ell)$-th race-ethnicity/sex-age subdomain;
$C_{hk}'$ = most recent population estimate for race/ethnicity subdomain $k$ in PSU $h$;
$P_{ik}$ = U.S. proportion in the $i$-th density stratum for the $k$-th race/ethnicity subdomain;
$C_{\cdot\cdot k\ell}$ = most recent projection of the year 2000 total population count for race-ethnicity/sex/age subdomain $(k,\ell)$; and
$C_{\cdot\cdot k\cdot}$ = most recent projection of the year 2000 total population count for race-ethnicity subdomain $k$.

By defining the measure of size in terms of the sampling rates as well as population counts, PSUs with larger populations for subdomains that are oversampled in NHANES have a greater probability of being selected. This results in reductions in the amount of screening required, compared to a MOS that is a function of population counts alone.

The probability of selection of an NHANES PSU, conditional on the NHIS PSU having been selected, is

$$p_{1h} = k_1 \frac{M_h'}{\sum_h M_h'} = k_1 \frac{M_h/\pi_h}{\sum_h (M_h/\pi_h)} ;$$

where

$k_1$ = number of PSUs selected.

Within each PSU, a sample of segments is selected. In NHANES, there are two types of segments: (1) Area segments, which are typically blocks or groups of blocks, and (2) New construction segments, which are sets of building permits for new residential construction. Some area sample surveys sample new construction units in area segments. This approach creates two problems for NHANES: (1) The segment measure of size may be inaccurate, if the segment contains much new construction; and (2) if a segment with a large proportion of new construction is selected, either a larger than expected sample will have to be drawn from that segment or a weighting factor must be applied. For NHANES, highly variable sample sizes are not economical due to the scheduling of the MECs. Applying a weighting factor would reduce the efficiency of the sample. Therefore, the exclusion of new construction from area segments and creation of separate

new construction segments is the operationally and statistically preferred approach for NHANES. When the data from the 2000 decennial census comes available, the process of creating separate new construction segments will be stopped until later in the decade, since a small amount of variation in sample sizes is tolerable.

Segments are also selected with probability proportionate to a measure of size. The segment MOS has the same form as the PSU MOS. The MOS is based on the sampling rates as well as the population counts, in order to give segments with larger populations for subdomains that are oversampled in NHANES a greater probability of being selected. The segment MOS is

$$M_{hj} = \sum_k A_{ik} C_{hjk} ,$$

where

$j$ denotes the segment;

$C_{hjk}$ = most recent population estimate for race/ethnicity subdomain $k$ in segment $j$ in PSU $h$;

and

$$A_{ik} = \sum_\ell P_{ik} r_{ik\ell} \frac{C_{\cdot\cdot k\ell}}{C_{\cdot\cdot k\cdot}} .$$

Thus, the conditional probability of selection for a segment is

$$p_{2hj} = k_{2h} \frac{M_{hj}}{\sum_j M_{hj}} ;$$

where

$k_{2h}$ = number of segments selected in PSU $h$.

Within segments, dwelling units (DUs) are selected with equal probability, at a rate equal to the maximum within-segment sampling rate required in order to attain the subdomain sampling rates.

Persons are selected within DUs using the ratio of the subdomain sampling rate to the maximum subdomain sampling rate. In NHANES, sampled persons are referred to as SPs.

Thus, the overall selection probability for a person in age-sex-race/ethnicity subdomain $(k,\ell)$ in density stratum $i$ is

$$\pi_h \left[ k_1 \frac{M_h / \pi_h}{\sum_h \left( M_h / \pi_h \right)} \right] \left[ k_{2\ell} \frac{M_{hi}}{\sum_{i \in h} M_{hi}} \right] * \left[ \frac{\max_{k\ell} \{ r_{ik\ell} \}}{p_{1h} p_{2hj}} \right] \left[ \frac{r_{ik\ell}}{\max_{k\ell} \{ r_{ik\ell} \}} \right]$$

$$= r_{ik\ell};$$

and it can easily be shown that these probabilities yield approximately equal sample sizes for each PSU.

## 3. Enhanced Features of the Future NHANES Survey

### 3.1 Annual Samples

To facilitate potential linkage with other surveys, retain flexibility in the sample design, and allow for the production of annual estimates for broad subdomains, the future NHANES will have a continuous, annual sample design. Each annual sample will be nationally representative. In NHANES III—a 6-year study—the first 3 years and the second 3 years of the study were each nationally representative. While attempts were made to cover each region of the country each year, NHANES III did not have nationally representative annual samples. The travel requirements for nationally representative annual samples will be challenging. Three MECs—two of which will be stationed at PSUs and one of which will be traveling at any given point in time—will work with a carefully designed schedule to meet the requirements of the study.

Although the annual samples will be nationally representative, annual estimates should only be produced for the nation as a whole, for race/ethnicity subdomains, or for very broad age-sex subdomains within race/ethnicity. Since NHANES can visit only a small number of PSUs on an annual basis, the user should be aware of the relative instability of variance estimates for annual estimates.

Table 1 gives the expected annual sample sizes by age-sex-race/ethnicity subdomain. Due to expected variations in the sample distribution on an annual basis, these numbers should be taken to be rough approximations. As the sample is aggregated across years, the age-sex-race/ethnicity subdomain sample sizes will increase and are expected to stabilize. Table 1 also contains the expected 2-, 3-, 4-, 5-, and 6-year sample sizes by age-sex-race/ethnicity subdomain. It is expected that accumulations of at least three years of data will be required to obtain reliable estimates for the age-sex-race/ethnicity sampling domains given in Table 1. It should also be noted that the NHANES

sample is designed to produce reliable cross-sectional estimates, but is not designed to detect year-to-year differences.

### 3.2 Sample Linkage With the NHIS

In 1995, in response to the NAS recommendation, the Department of Health and Human Services (DHHS) submitted a plan to the Office of Management and Budget proposing to integrate and coordinate data collection systems within the Department. There were a number of goals sought for this survey integration plan, including:

- Maximize the comparability of questions, sampling techniques, and other aspects of surveys to enable analytic linkage of data from several surveys, enhance analytic utility and ease of use for research analysts;
- Reduce costs by sharing knowledge and resources, instrument testing and design techniques;
- Allow longitudinal studies of diseases and disorders by linking to mortality registers and cancer registers, for example;
- Allow linkage between survey samples at the county or other geographic level or even at the level of individuals to enhance the study of the etiology of diseases; and
- Reduce data redundancy and waste of time and resources for Departmental staff and for public researchers, allowing staff to concentrate on new areas.

Based on preliminary research on linkage alternatives, the current design is for the NHANES to be linked to the NHIS at the PSU level and at the content level. Further research and evaluation of linkage alternatives is in progress.

The NHIS is an ongoing survey of the noninstitutionalized civilian population that collects data on general health-related issues such as health status, health care utilization and access, chronic ailments, and acute conditions. It also collects demographic data, as well as data on health insurance coverage, income, and program participation.

The sample design for the NHIS has undergone periodic revision since its inception in 1957; however, the basic design is a multistage, stratified area sample with metropolitan statistical areas or groups of counties as the PSUs, area segments consisting of Census blocks or block groups as the second-stage sampling units, and households as the third-stage sampling units. For the 1995 NHIS, the requirements of the survey were expanded to include improved precision for statistics for various domains defined by race, ethnicity, and geography. Extensive research was conducted to investigate possible design options that would meet the expanded requirements to the extent feasible. Partly as a result of this research, the 1995 NHIS sample was redesigned and now includes about 350 PSUs (compared to about 200 in the 1985-94 surveys) and a total of about 6,000 segments. For a review of the current status of NHIS integration plans and references to the NHIS sample design, refer to Ezzati-Rice *et al.* (forthcoming).

### 3.3 The Integrated Survey Information System (ISIS)

The future NHANES will contain an integrated survey information system (ISIS) that links all survey operations through a wide area network. Telecommunications will be set up to connect all field offices with the home office. The ISIS will allow for nearly instantaneous transmittal of data; automated processing, scheduling, and tracking; and enhanced sample monitoring. The system will also enhance and expedite the editing and data preparation process.

## 4. Current Plans for the Future NHANES Design

To achieve the goals of the Survey Integration Plan, the sample design of the future NHANES will be based on a continuous ongoing annual survey of the civilian, non-institutionalized population of the U.S. To meet additional Survey Integration Plan objectives, NHANES will be linked to the NHIS at the PSU level as well as the content level.

The counties in PSUs from two panels of the 1997 NHIS will be used as the sampling frame for NHANES. NHANES will visit about 14 to 15 PSUs per year. Each single year and any combination of consecutive years will comprise a nationally representative sample of the U.S. population. This design will facilitate potential linkage to other health and nutrition surveys that provide yearly estimates, in particular, the Continuing Survey of Food Intake by Individuals (CSFII), conducted by the U.S. Department of Agriculture, and will allow aggregate-level national estimates from NHANES each year. In addition to the 14 to 15 PSUs selected for the main study, the NHANES design will allow for the possibility of the selection of additional PSUs either to be used for a special study or to be held in reserve.

To achieve content-level linkage, questions from the NHIS are being used in the NHANES interviews. To further evaluate linkage between the NHIS and the NHANES, some PSUs in the future NHANES will include NHIS sample selected in the same county. This will allow for an evaluation of a more fully linked NHIS and NHANES design, to evaluate the potential effect on response rates, to develop more accurate cost estimates, to test the ability of local area estimation, and to quantify the analytic potential of a linked design.

## 5. References


Arnett, R.A., Hunter, E., Cohen, S., Madans, J., and Feldman, J. (1996). "The Department of Health and Human Services' Survey Integration Plan." *Proceedings of the Section on Government Statistics.* Alexandria, VA: American Statistical Association.

Ezzati-Rice, T., Cohen, S., Khare, M., and Moriarty, C. (forthcoming). "Using the National Health Interview Survey as a Sampling Frame for Other Health-Related Surveys." 1998 *Proceedings of the Section on Survey Research Methods.* Alexandria, VA: American Statistical Association.

Judkins, D., Marker, D., and Waksberg, J. (1994). *National Health Interview Survey: Research for the 1995 Redesign.* Report prepared for the National Center for Health Statistics, Public Health Service, Centers for Disease Control and Prevention. Rockville, MD: Westat.

U.S. Department of Health and Human Services (DHHS), National Center for Health Statistics, Centers for Disease Control and Prevention (1996). NHANES III Reference Manuals and Reports: Plan and Operation of the Third National Health and Nutrition Examination Survey, 1988-94 (CD-ROM). Hyattsville, Maryland.


Table 1. Expected NHANES annual, 2-year, 3-year, 4-year, 5-year, and 6-year sample sizes

| | | | Expected yield (Number of examined SPs) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Annual | 2-Year | 3-Year | 4-Year | 5-Year | 6-Year |
| Black | M&F | <1 year | 50 | 100 | 151 | 201 | 251 | 302 |
| | | 1-2 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 3-5 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | M | 6-11 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 12-15 years | 95 | 190 | 285 | 380 | 475 | 570 |
| | | 16-19 years | 95 | 190 | 285 | 380 | 475 | 570 |
| | | 20-39 years. | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 40-59 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 60+ years | 88 | 176 | 265 | 353 | 441 | 530 |
| | F | 6-11 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 12-15 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 16-19 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 20-39 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 40-59 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 60+ years | 88 | 176 | 265 | 353 | 441 | 530 |
| Mexican-American | M&F | <1 year | 92 | 185 | 277 | 369 | 462 | 554 |
| | | 1-2 years | 93 | 186 | 280 | 373 | 466 | 560 |
| | | 3-5 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | M | 6-11 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 12-15 years | 95 | 190 | 285 | 380 | 475 | 570 |
| | | 16-19 years | 95 | 190 | 285 | 380 | 475 | 570 |
| | | 20-39 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 40-59 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 60+ years | 90 | 181 | 272 | 362 | 453 | 544 |
| | F | 6-11 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 12-15 years | 95 | 190 | 285 | 380 | 475 | 570 |
| | | 16-19 years | 95 | 190 | 285 | 380 | 475 | 570 |
| | | 20-39 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 40-59 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 60+ years | 95 | 190 | 285 | 380 | 475 | 570 |

Table 1. Expected NHANES annual, 2-year, 3-year, 4-year, 5-year, and 6-year sample sizes (continued)

| | | | Expected yield (Number of examined SPs) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Annual | 2-Year | 3-Year | 4-Year | 5-Year | 6-Year |
| White/Other | M&F | <1 year | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 1-2 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 3-5 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | M | 6-11 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 12-15 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 16-19 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 20-29 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 30-39 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 40-49 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 50-59 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 60-69 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 70-79 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 80+ years | 88 | 176 | 265 | 353 | 441 | 530 |
| | F | 6-11 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 12-15 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 16-19 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 20-29 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 30-39 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 40-49 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 50-59 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 60-69 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 70-79 years | 88 | 176 | 265 | 353 | 441 | 530 |
| | | 80+ years | 88 | 176 | 265 | 353 | 441 | 530 |
| Total for age-sex-race/ethnicity domains | | | 4,702 | 9,404 | 14,105 | 18,807 | 23,509 | 28,210 |
| Additional pregnant women* | | | 124 | 249 | 374 | 499 | 624 | 749 |
| Overall Total in National Sample | | | 4,826 | 9,653 | 14,479 | 19,306 | 24,133 | 28,959 |

*It is expected that about 300 pregnant women will have been sampled (in a 6-year sample) based on their age-sex-race/ethnicity. The additional pregnant women are those who will have been sampled only due to their pregnancy status.

**There are no explicit targets for infants (age <1 year). The numbers given here are the expected yield, given the estimated amount of screening.