NYC HANES: DESIGN OF A COMMUNITY HEALTH AND NUTRITION EXAMINATION SURVEY

Jill M. Montaquila, Leyla Mohadjer, Westat; Lester R. Curtin, National Center for Health Statistics; R. Charon Gwynn, Lorna Thorpe, New York City Department of Health and Mental Hygiene
Jill Montaquila, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Sample design, local study, area sample

1. Introduction

The National Health and Nutrition Examination Surveys (NHANES) are part of several health related programs sponsored by the National Center for Health Statistics (NCHS). A unique feature of these surveys is the collection of health data by means of medical examinations on a nationally representative sample of the U.S. population. This information is essential for estimating the prevalence of various diseases and conditions, indicating which social and demographic groups are particularly vulnerable to various diseases, and planning for health policy. Because health and nutrition characteristics vary considerably by demographic and geographic characteristics, intensive study of particular subgroups through the "community" HANES model has been a longstanding interest.

The first survey that followed this model was the Hispanic HANES (HHANES, 1982-1984), which produced estimates for three major Hispanic subgroups. In 2003, NCHS and Westat agreed to provide technical expertise on a project spearheaded by the New York City Department of Health and Mental Hygiene to conduct a community HANES in New York City. Interviews and examinations for the New York City HANES (NYC HANES) were conducted during the summer and fall of 2004 on a representative sample of the NYC adult population. This paper presents an overview of the community HANES model, and gives a summary of the objectives, sample design, and analytic capabilities of NYC HANES. Section 2 gives a background of the NHANES survey program and the Community HANES model. The sample design for NYC HANES is described in Section 3. Section 4 contains a description of the expected analytic capabilities of NYC HANES.

2. Background of NHANES and Community HANES

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 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.) The current NHANES study, which is designed to be a continuous study with nationally representative annual samples, began in 1999. As mentioned earlier, unique to NHANES are the complete medical examinations for each respondent in the sample. In order to provide uniform methods of administration throughout the U.S., these examinations are carried out in mobile examination centers (MEC).

The "community" HANES model. The goal of a community HANES study is to provide a flexible and timely survey mechanism to examine health issues in a range of populations for which NHANES does not supply sufficient sample sizes. The set of analytic domains NHANES is designed to support consist of age-sex groups for blacks, Mexican-Americans, and the remainder of the U.S. population. Under the community HANES model, the study may be designed to allow estimates to be made for small areas, such as cities, counties, groups of counties or states, or for special populations, for example, an ethnic group. Operationally, a community HANES study might be more flexible than the NHANES, in that examinations may be conducted in one (or more) of several types of facilities—a trailer that requires transporting, a self-driven trailer, or a clinic or doctor's office.

Overview and objectives of NYC HANES. The New York City (NYC) Department of Health and Mental Hygiene (DOHMH) aims to address a number of health-related issues significant to NYC by establishing a community-level HANES, referred hereafter as NYC HANES. Specifically, DOHMH will determine the level of important indicators of health status, such as: hypertension; obesity; diabetes; other precursors to cardiovascular disease; environmental exposure to tobacco, metals and pesticides; mental health disorders; and certain infectious diseases.

Findings from the NYC HANES will complement, yet not duplicate, information gathered from other recent NYC survey initiatives. The NYC HANES will not have neighborhood-specificity, but will provide key prevalence information on conditions that require a physical exam (i.e. blood pressure, cholesterol level) and sensitive conditions that are not easily ascertained by telephone (i.e. mental health, sexual behavior). Results from the NYC HANES will provide critical baseline data on priority chronic and infectious diseases from which programmatic efforts can then be guided. Periodic repetition (approximately every 5 years) of the NYC HANES survey can be used to evaluate DOHMH interventions. Objectives of NYC HANES include:

1. To estimate the number and percent of persons in the NYC population with selected diseases and risk factors;
2. To estimate city-wide awareness, treatment, and control of selected diseases;
3. To estimate prevalence, awareness, treatment, and control of selected diseases among a limited set of demographic subgroups identified by race/ethnicity, gender, broad age bands, and borough (county);
4. To monitor prevalence and magnitude of environmental exposures in NYC;
5. To analyze risk factors for selected diseases in NYC; and
6. To establish a population-based serologic repository that can be used to explore emerging public health issues in NYC.

3. Sample Design

Target population. The target population for NYC HANES was persons 20 years or older residing in New York City. Persons residing in group quarters were excluded from the target population; that is, NYC HANES covered the civilian, household population, since it is often very difficult for listers and interviewers to gain entry into group quarters. Furthermore, the group quarters population comprises only 2.3 percent of the total population, and only 1.3 percent of the total civilian noninstitutional population in New York City. Thus, the group quarters population was excluded from the target population for NYC HANES.

General design parameters and sample design research. The following assumptions were made for purposes of designing the sample for an NYC HANES study:

- Data collection was to occur in 4 different sites. The sites were for operational purposes only; that is, they did not correspond to primary sampling units;
- The target sample size was 2,000 examined sample persons (SPs) 20 years of age and older; and
- An overall examination response rate of about 60 percent was expected. In this research, a screener response rate of about 80 percent and a 75 percent examination response rate conditional on having completed the screener were assumed. Thus, it was expected that a total of about 2,667 identified SPs would be required in order to examine 2,000 SPs.

A three-stage sample design was used for NYC HANES. The first stage of sample selection was the segment. A segment consists of a block or group of proximal blocks having a required minimum size. A measure of size (based on the number of households in the segment) was associated with each segment in the frame and the sample of segments was selected with probability proportional to the measure of size (PPS). The second stage of selection in NYC HANES was the household (within sampled segments), and the third stage was the selection of persons (*sampled persons*, or SPs) within households. The sampled segments were field listed and housing units subsampled as necessary to yield an expected target number of about 2,667 identified SPs.

During the design phase, several sample design alternatives employing this three-stage design were considered for NYC HANES. One sample design option involved sampling dwelling units within segments and then sampling exactly one eligible person from each household. Although all eligible persons in a household would have the same chance of selection, choosing one person per household implies that the chance of selection of a person depends on the household size. Persons in one-eligible-person households will have twice the chance of selection of persons in two-eligible-person households, three times the chance of selection of those living in three-eligible-person households, etc. Although these differential probabilities will be compensated for in the weighting process to allow the creation of unbiased estimates, the resulting variation in sampling rates will considerably reduce the precision of the survey estimates. Furthermore, in our experience with NHANES III and NHANES '99, we have observed that response rates are higher when more than one SP is selected in a household. (This is likely attributable

to the incentives paid to each respondent, resulting in higher payments to larger families.)

A second alternative design was to sample more than one person per household, in households with more than one eligible person. This alone would still result in variable selection probabilities, but these variable selection probabilities could be negated by screening a larger number of households and then subsampling households depending on the number of eligible persons in the household. We considered several alternatives involving sampling more than one person per household, with additional screening to yield an equal-probability sample of SPs.

The alternative requiring the smallest screening sample is the one in which all eligible SPs in the household are selected into the sample. Also, the design costs will be lower under this alternative, since there is no need to develop a screener to conduct any within household sampling. However, as noted in the following discussion, the price paid for this is a reduction in the statistical efficiency of the sample.

One potential feature of samples with more than one SP per household is the effect of intra-household correlation on the precision of survey estimates. For many health and nutrition characteristics, there is some degree of correlation—typically positive—among household members. (Intra-household correlation is not a concern if sampled persons in the same household are in different analytic domains.) Estimates of intraclass correlation due to within-household clustering were calculated using data from NHANES 1999-2000 survey results. Variables of interest to NYC were used in the analysis including incidence of diabetes, blood glucose levels, incidence of hypertension, blood pressure levels, body mass index, and cholesterol variables. These variables were all thought to be clustered within-households. The estimated intraclass correlation for these variables averaged 0.12, with a minimum value of 0.05 and a maximum value of 0.20. As noted above, these intraclass correlation estimates reflect only the within-household correlation. This type of correlation is relevant and useful for comparing the alternatives presented here, since the alternatives differ only in the numbers of persons sampled within households. However, it should be noted that intraclass correlation also arises due to within-segment clustering of households.

The sample design chosen for NYC HANES was one that reduces the number of SPs per household to a reasonable level, but keeps the design effect due to intra-household correlation at a relatively low level. Table 1 presents the details of this design, based on population totals for New York City from the Census 2000 Summary File 1 (SF1).

As a basis for comparison, it should be noted that in a "take all" sample design, screeners would need to be attempted in 1,930 units and completed in 1,420 households. Of course, the design effect due to intra-household correlation would be expected to be larger than in the designs considered above. Using the same estimates of intra-household correlation (.05, .12, and .20), we would expect the design effect to be in the range of 1.05-1.22.

Segment selection. Previous research conducted by Westat for NHANES indicated that for a PSU in which 333 SPs are examined, the optimal size—accounting for the counter effects of intraclass correlation (within segments) and operational efficiency—is about 24 segments, or an average of about 14 examined SPs per segment. Thus, for NYC HANES with a target of 2,000 examined SPs, the sample was designed to include a total of 144 segments. In order to facilitate examinations at 4 different sites within the city while controlling sample yield and still attain a the same weighting scheme throughout the city, all segments (across all 5 boroughs) were selected at once, and these 144 segments were then geographically grouped into sites. In forming the sites, segments in Manhattan and Staten Island were combined into a single site, and segments in each other borough formed separate sites (i.e., Bronx, Brooklyn, and Queens each had its own site). The division of the sample into sites was used for operational (field management) purposes only; the sites were not used for sampling purposes.

The measure of size used to select segments for NYC HANES was the number of households in the segment. This information was obtained for each block from the 2000 census SF1 file, and blocks were then aggregated to form segments that met a minimum measure of size criterion. The minimum segment measure of size was determined based on the need to attain the household sampling rates discussed above. A sample of 144 segments was selected systematically with probabilities proportional to the measure of size.

Household and SP selection. The household was the second-stage unit in the NYC HANES sample. A sample of households was selected from the field listing of the sample segments. The screened households were subsampled for SP selection at rates ranging from 0.5 to 1, depending on the number of eligible enumerated persons in the household. (See below.) The SP was the third-stage unit.

Table 1. Sample design for NYC HANES, with variable numbers of SPs per household

Number of eligible persons in household	Number of households	Population	Number of eligible persons to be selected in household	Number of households with completed screener	Overall sampling rate for designating households for SP selection*	Overall SP sampling rate	Total number of	
							Households with SPs	SPs
1	1,170,514	1,170,514	1	1,101	0.000588	0.0005	550	550
2	1,677,047	3,354,095	1	1,577	0.001175	0.0005	1,577	1,577
3	260,158	780,473	2	245	0.000881	0.0005	183	367
4	66,613	266,451	2	63	0.001175	0.0005	63	125
5 or more	20,173	100,865	2	19	0.001175	0.0004	19	38
Group quarters population		182,430						
Total	3,194,505	5,854,828		3,004			2,392	2,657

*Rates have been inflated to account for expected screener nonresponse of 20 percent. During the sample design process, it was observed that a considerable reduction in screening could be realized by reducing that rate for households with 5 or more eligible persons to 0.001175 (the maximum among other household sizes), with a negligible loss of statistical efficiency.

Under the design given in Table 1, in households with 1 or 2 eligible persons, exactly 1 SP was selected; in households with 3 or more eligible persons, 2 SPs were selected. Thus, for example, in 1-person households, that SP was selected with certainty, but in 2-person households, one person was selected with probability 0.5. In order to negate these differential selection probabilities arising at the final stage of selection, the screening sample was increased and households were subsampled based on the number of persons enumerated in the household. In order to perform this subsampling, households were sampled for screening at the maximum rate (0.001175, or about 1 in 851 households). That is, of the approximately 3 million households in New York City, 3,754 (3,194,505*0.001175) were sampled for screening. In the screening interview, household members were enumerated and eligibility was assessed. Based on the enumeration, households were subsampled for SP selection. For example, all households with 5 or more eligible persons (a group with the highest household sampling rate) were designated for SP selection, but among 1-person households, only 50 percent (or 0.000588/0.001175) were designated for SP selection. In this manner, a constant overall sampling rate of 0.0005 was attained[1] (i.e., this approach yields an approximately self-weighting sample of SPs) while controlling for the number of SPs selected per household.

According to the Census 2000 SF1, the overall vacancy rate in New York City is 5.6 percent. (This is lower than the national vacancy rate of 9 percent.) Additionally, a very small proportion of listed units could be expected to be found by the interviewers to not be DUs. Thus, we inflated the DU sample by about 8 percent to account for vacant units and "not a DU" units. Therefore, to attempt screening in 3,754 households, we selected 4,081 units.

In summary, the NYC HANES sample design was aimed to achieve a target sample size of about 2,667 sample persons to yield about 2,000 examined SPs, assuming an 80 percent screener response rate and a 75 percent examination response rate conditional on having completed the screener. The design was a three-stage area probability sample. The first stage of selection was the selection of area segments. About 144 segments were selected with probability proportional to the total number of households in the blocks comprising the segment. Segments were defined so they were large enough to allow for a 20 percent loss due to screener nonresponse, an 8 percent loss due to vacant units and not-a-dwelling-unit, plus an unreleased reserve sample (which would have been released only if it had become necessary due to extreme, unexpected circumstances resulting in a substantial loss of sample). Table 2 summarizes the characteristics of the sample design for NYC HANES.

[1] There is a slight variation in the sampling rates in households with 5 or more persons because of the reduction in the sampling rate to the maximum needed for 1- to 4-person households (in order to reduce screening). However, negligible bias will be introduced by assuming a constant rate.

Table 2. Characteristics of the sample design for NYC HANES

Total eligible (age 20 or older) population of New York City	5,854,828
Total number of households in New York City	3,021,588
Number of sampled segments	144
Expected minimum segment size (total number of DUs)	43
Expected number of:	
Units in which screening is attempted	4,081
Households with completed screeners	3,004
Households designated for SP selection	2,392
Identified SPs	2,657
Examined SPs	2,000

4. Analytic Capabilities

NYC HANES was developed with the aim of assessing a number of health-related issues significant to New York City, including hypertension, obesity, diabetes, other precursors to cardiovascular disease, environmental exposure to tobacco, metals and pesticides, mental health disorders, and certain infectious diseases. Key analytic subgroups of interest in NYC HANES include those defined by race/ethnicity (non-Hispanic White, non-Hispanic Black, Hispanic, other), gender, broad age bands (20-44, 45-64, 65 and older), and borough. Table 3 lists the research questions NYC HANES was designed to address and the data collected to address those research questions.

Expected precision of key estimates for key subgroups. Due to the broad nature of the NYC HANES survey content and the historical precedent of specific sample size considerations for the CDC's NHANES survey, sample size targets were developed using a combination of information. First, precision requirements for NHANES were examined and determined to be a gold standard for which NYC HANES should strive to attain. The precision requirements established for NHANES III (1988-1994) were:

1. An estimated prevalence statistic of the order of 10 percent in an age-sex domain should have a relative standard error of 30 percent or less.
2. Estimated absolute differences between domains of at least 10 percent should be detectable with a Type I error rate of 0.05 and a Type II error rate of 0.10.

Assuming a design effect of 1.5, the sample size necessary to satisfy the first condition is about 150 examined persons; the sample size necessary to satisfy the second condition is about 420 examined persons.

Although the NHANES precision requirements were used as a guide, power calculations were used to determine the target sample sizes across a wide range of expected frequencies. Balancing sample size requirements and study cost considerations, a target sample size of 2,000 examined adults was established. With such a sample size, NYC HANES will have slightly higher power than the current annual NHANES for subgroups of the same size. Briefly stated, NHANES averages a little over 2 SPs per household with SPs; in NYC HANES, the average will be closer to 1. The clustering of SPs within census segments will be comparable to NHANES. And, unlike NHANES, there is no variation in sampling rates for different sampling domains (e.g. demographic groups). Additionally, in NYC HANES, unlike NHANES, there is no between-PSU variance component. All of these design factors are expected to result in improved power relative to NHANES for subgroups of the same size. The anticipated lower response rates for NYC HANES will necessitate weighting adjustments to adjust for nonresponse that will introduce variation into the weights and reduce the power somewhat; however, the increases in power relative to NHANES due to the factors described above are expected to more than offset the reductions in power resulting from lower response rates.

5. Reference

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 3. Data items and corresponding research questions for NYC HANES

Data collected	Research questions to be answered
Blood pressure	1. What proportion of NYC adults has hypertension? Of those with hypertension, what proportion is on medication? Is controlled? 2. What proportion of NYC adults has hypertension by major demographic and risk groups (gender, race/ethnicity, age (above/below 60), smoking, obesity, hypercholesterolemia, etc.)?
Height and weight	1. What proportion of NYC adults is overweight? Obese? 2. What proportion of NYC adults is overweight (obese) by major demographic groups?
Standing waist girth	1. What proportion of NYC female adults has a standing waist girth of more than 35 inches? Male adults with a standing waist girth of more than 40 inches?
Lipid profile	1. What proportion of NYC adults has hypercholesterolemia? Is on medication? Is controlled? 2. What proportion of NYC adults has hypercholesterolemia by major demographic and risk groups (gender, race/ethnicity, age +/- 60, smoking, obesity, hypertension, etc.)? 3. What proportion of NYC adults has high triglyceride levels? 4. Low HDL levels? 5. Metabolic syndrome?
Fasting plasma glucose and glycohemoglobin	1. What proportion of NYC adults has diabetes? Undiagnosed diabetes? 2. What proportion of NYC adults has diabetes by major demographic and risk groups (gender, race/ethnicity, age +/- 60, smoking, obesity, hypertension, etc.)? 3. What proportion of NYC adults with diabetes have relative glycemic control? 4. What proportion of NYC adults have impaired fasting glucose (IFG)?
Hepatitis C	1. What proportion of NYC adults test positive for HCV antibody?
Herpes Simplex Virus-2	1. What proportion of NYC adults test positive for HSV-2 antibody? By demographic and risk groups?
Serum cotinine	1. What is the prevalence of current smoking, according to self-report and cotinine levels >15ng/ml? 2. What is the prevalence of substantial environmental exposure to smoking, according to self-report and cotinine levels 1-15ng/ml? 3. What proportion of NYC adults do not report smoking but have cotinine levels indicative of current smoking?
Mental health ACASI	1. What is the prevalence of major depression? Generalized anxiety disorder? 2. What proportion of NYC adults with these mental disorders is currently receiving treatment/medication, or are in care? 3. What is the prevalence of these mental disorders by demographic groups? 4. What proportion of NYC adults report attempting suicide in the past 4 weeks? Report suicidal thoughts in the past 4 weeks?
Sexual behavior ACASI	1. What is the prevalence of the following self-reported STDs among NYC adults – gonorrhea, chlamydia? 2. What is the mean/median age at first sexual intercourse among NYC adults and how does this result using ACASI compare to telephone survey results? 3. What is the median number of sexual partners in the past 12 months among NYC adults? By demographic risk factors? How does this compare to telephone survey results? 4. What proportion of NYC adults report sex with persons of the same gender in the past 12 months? How does this compare to telephone survey results?
Additional questionnaire components	Physical activity, hospital utilization and access to care, smoking and tobacco use, secondhand smoke exposure, medical conditions and medical history (including cancers, asthma, arthritis, etc.)