SAMPLE DESIGN RESEARCH FOR THE NATIONAL NURSING HOME SURVEY

Karen E. Davis
National Center for Health Statistics, 6525 Belcrest Road, Room 915, Hyattsville, MD 20782

KEY WORDS: Sample survey, cost model

1. Introduction

The National Nursing Home Survey (NNHS) is a nationally representative sample survey of nursing home facilities, their residents, discharges, and staff conducted by the National Center for Health Statistics (NCHS). The NNHS estimates are for freestanding facilities and nursing care units in hospitals, retirement centers, or similar institutions where the unit maintained separate financial and resident records from the larger institutions. The NNHS has been conducted six times since 1973 (1973-74, 1977, 1985, 1995, 1997 and 1999). The NNHS was not fielded in 2001 to allow time to conduct survey and sample design developmental work that would facilitate future survey redesign efforts. This paper describes the research objectives, the current sample design, the design changes and ongoing research activities. A redesigned NNHS will be fielded in 2003.

2. Overview of Research Issues

A major goal of this redesign is to provide policymakers, policy analysts, and researchers with the data they need to address key "long-term care" policy issues. Changes in the health care industry have led to the development of new types of facilities (such as life care communities), and programs aimed at providing individuals with home care. There is also a greater variability in the types of people receiving long-term care services. Nursing homes traditionally served the elderly, however they now provide rehabilitation services for younger people for shorter periods of time, or for special populations. The sample redesign addresses these research issues as they relate to the overall survey objectives.

3. Survey Objectives

As in prior cycles, the redesigned NNHS is to produce estimates about the nursing and related care homes, their current residents, discharges, and their staff (Gabrel, et al., 2000). The survey design is to produce statistics for current residents with maximum precision for the available funds. In past cycles, the design was required to reduce respondent burden per sampled nursing home so interviewers could complete data collection in each facility in a single day. For the NNHS design research, specific sampling activities include:

- Investigate different scenarios for sampling admissions, current residents, discharges, or some combination of these encounters;
- Investigate sample sizes necessary to produce estimates of specific populations, or in specific types of settings;
- Investigate the possibility of sample designs that would have the ability to produce state-level estimates;
- Develop a cost model for detailed cost analysis.

Other research activities include creation of an ongoing comprehensive inventory of long-term care facilities/places that would serve as a sampling frame; investigation of linkages to the Center for Medicare and Medicaid Services'(CMS) Minimum Data Set database that would provide detailed resident-level clinical information; questionnaire development, programming, and testing for the switch from "paper and pencil" data collection to CAPI mode.

4. Sampling Design

The current sample design is a stratified two-stage probability design that is based on research undertaken for the 1985 NNHS (Shimizu, 1986). The first stage of selection is a probability sample of the nursing facilities in the sampling frame. The second stage of sample selection is done using a sample selection table to obtain systematic probability samples of current residents and discharges. The sampling frame for the most recent (1999) survey was derived from a frame that consisted of all nursing home facilities identified in the 1991 National Health Provider Inventory and updated with current (1999) files of nursing homes. The OSCAR file from the CMS and the 1998 SMG Marketing Group file were also used to select 1999 birth sample nursing homes. The universe was then stratified to draw the sample of nursing homes based on the certification status code, ownership code, census region, MSA code, bed size, and hospital-based status. Next, bed sizes are accumulated across facilities in order to compute sampling intervals. Facilities are then selected into the sample using the systematic sampling method. Facility sample weights are retained and are equal to the sampling interval divided by the facility's bed size.

The second stage of sample selection, involves sampling up to six current residents and up to six discharges within each facility. Sampling is done using a sample selection table with the sampled units determined by each possible count of residents and discharges in the facilities. The sample numbers in the tables are selected by systematic sampling. The current residents are usually selected with the same probability of selection within stratum, since resident counts and bed size tend to be highly correlated, and since facilities are selected with probability proportional to size. The sampled discharges tend to have unequal probabilities of selection, since discharge counts are less correlated with bed size. The sampling frames for within-facility samples consist of lists constructed by the interviewers at the time of the survey. The Current Resident Sampling List consists of all residents on the register of the facility on the evening prior to the day of the survey. The Discharged Resident Sampling List consists of all events in which persons were discharged (alive or dead) during the month ending on the day prior to the facility's survey date.

5. Sampling Unit

One specific redesign objective was to investigate whether admissions, current residents, discharges, or some combination of these encounters should be sampled in the 2003 NNHS. There are advantages and disadvantages to the various sample units. Although sampling admissions has several analytic advantages in studying the full spectrum of nursing home users, it is somewhat more complex and resource-intensive than sampling either current residents or discharges. Continued sampling of both current residents and discharged residents will permit future trend analyses since all past surveys have contained a current resident sample. The discharged sample provides an estimate of the short stay population, the volume of care, and allows for studying end of life care. Thus, it was agreed that taking samples of both current residents and discharges maximizes the likelihood of obtaining a sample representative of typical nursing home users.

6. Sample Sizes for Estimates of Specific Populations

A primary sampling goal of the NNHS is to improve NCHS' ability to address key quality issues and produce estimates with satisfactory precision (e.g., a relative standard error (RSE) of 30% or less for prevalence estimate "p", for $p \geq 1\%$). Thus, research was conducted to examine the precision levels that can be obtained from the current NNHS for selected health characteristics of demographic subdomains. The proposed demographic subdomains of interest for estimates are by race-ethnicity and age. They consist of the cross-classifications in Table 1 (see appendix).

The minimum sample size necessary for an analytical cell (e.g., white females 65-74 years old), assuming a design effect of 1.5 and a coefficient of variation less than or equal to 30% is 1,650 for $p \geq 1\%$. In order to produce national estimates, the current resident sample should total 39,600 persons for prevalence estimate $p \geq 1\%$.

The first priority was analyzing national estimates for total demographic (all race-ethnicity groups combined) domains before proceeding to examine individual subdomains within each race-ethnicity domain. The types of health variables that should be examined for measuring precision levels attainable from the current NNHS design should ensure the research includes a mix of variables with high and low prevalence estimates in order to examine the precision levels of a wide range of estimates for various sociodemographic groups. Analytic variables representing 6 health characteristics were selected from the 1999 NNHS current resident and facility files. These variables were:

1. Current Diagnosis: Decubitus Ulcer
2. Current Diagnosis: Fractures
3. Length of Stay - in days (5yrs)
4. Admission Diagnosis: Decubitus Ulcer
5. Admission Diagnosis: Fractures
6. Nursing (RN, LPN, Nurse's Aides, Orderlies FTEs)

The 1999 NNHS current resident data file was used to investigate the current precision level for the prevalence estimates of the analytic variables for the major race-ethnicity subdomains. SUDAAN was used in the precision analysis of prevalence estimates. The number of respondents to the health questions, prevalence estimates, standard error estimates, and design effects were computed using SUDAAN software.

Table 2. Results of Selected NNHS Health Characteristics

Category	Hispanics RSE (%)	Blacks & Others RSE (%)	Whites RSE (%)
All	**	23.6	12.9
Males	**	**	21.5
Females	**	**	7.6

Note:** indicates cell exceeded the precision requirement for an RSE of 30 percent or less.

The results given in Table 2 follow from investigations of the 6 analytic variables representing 6 health characteristics from the 1999 NNHS, applying RSE precision criteria of 30%, and prevalence level of $p \geq 1\%$. Prevalence estimates from each of the six analytic variables were examined, and a given race-ethnicity subdomain was considered to be capable of producing estimates that meet the RSE requirement if estimates for all six analytic variables that meet or exceed the specified prevalence level also met the stated precision requirements. The summary results indicated that the current resident sample was not sufficient to meet precision requirements. This sample would need to be increased by a factor of 4.8 in order to satisfy the precision levels for characteristics with $p \geq 1\%$. Since the cost of such an increase was prohibitive, additional research was conducted using a revised demographic subdomain format that combines Hispanic persons with Blacks and Others, and does not contain a cross-classification by gender. In order to produce national estimates for these revised subdomains, the current resident sample should total 13,200 persons for prevalence estimates $p \geq 1\%$.

As in the previous summary, the results for the revised demographic subdomains indicated that the current resident sample was still not sufficient to meet the precision requirements for several analytical cells. This sample would need to be increased by a factor of 1.7 in order to satisfy the precision levels for characteristics with $p \geq 1\%$. This factor includes inflation for nonresponse.

7. State-level Estimates

Ideally the NNHS should provide state-level estimates for every state. However, since the NNHS is designed as a national survey, the current design and the sample size of approximately 1,450 facilities (Jones, 2002) is currently insufficient to provide this level of detail. Further, the NNHS is presently not stratified by state. Therefore, research was conducted to examine what level of accuracy might be expected in each state using one year, or possibly two years, of data based on the current NNHS sample size.

Using 1999 NNHS data, we examined for each state the total facilities, the current sample, the expected annual sample size for a self weighting sample, and the effective sample size when we combine two years' worth of data. It was determined that in the current design, only 6 states (CA, IL, NY, OH, PA, TX) are able to produce state-level estimates of total facilities (without breakdowns by bed size, certification, etc.). If we were interested in data on facility characteristics, doubling the sample would not be sufficient to produce these estimates for all 50 states.

The possibility of producing state-level statistics for a combination of 2 years of NNHS data was also examined. Note however, that the NNHS is not conducted in 2 consecutive years. With a self-weighting sample design, 17 of the states would have adequate sample to produce estimates of total facilities using 2 years of data. However, the coefficient of variation for only 6 of these 17 states meet the precision requirement, having a coefficient of variation of 30 percent or less.

State-level NNHS data would necessitate more than doubling the current facility sample size. In some states, a census of facilities would be required. However if the current sample size is maintained, reliable estimates of total facilities can be produced for six states.

8. Variable Cost Model

In order to develop an efficient sample design for the NNHS, variable survey costs were estimated (i.e., costs that increase when sample sizes increase). A simple overall cost model was assumed, then alternative sample allocations using the model were developed. Estimation costs were considered for both current residents and discharged residents and reflect the "paper and pencil" data collection method. Note that a fixed number of sample cases per home can increase precision for current resident statistics when the homes are selected with probability proportional to size and size is bed size. However, the same is not true of discharge or admission statistics. Therefore, both designs were considered.

The initial cost function has components for sampling facilities and patients. In principle, the costs include training, travel, and other expenses of interviewers and supervisors, and other expenses associated with data collection and processing. For a given sample, the cost could be expressed as:

$$C = c_0 + \sum_{i=1} m_i c_i + \sum_{i=1} \sum_{j=1} m_{ij} c_{ij}$$

where C = overall cost

c_0 = fixed costs

c_i = variable costs of including an additional facility in the sample at the i th stage.

m_i = number of sample facilities selected at the i th stage.

c_{ij} = variable costs of including an additional resident in the sample at the j th stage.

m_{ij} = number of sample residents selected at the j th stage.

In this model, fixed costs would contain the budget for development activities such as frame development, survey design, and questionnaire development; weighting and analysis would also be considered as fixed costs, since weights would be computed regardless of the number of facilities or patients sampled.

The cost per facility would include the time for enrolling the facility, listing and sampling at the facility, time for the facility questionnaire, and local travel (e.g. miles driven) for carrying out these activities. It would also include staff training (since the more facilities, the greater the number of interviewers to be trained); also, while travel expenses would be included in this component, these costs would be for long distance travel and thus more expensive.

Estimates for the nursing home level were estimated by facility bed size (i.e., under 100 beds, and at least 100 beds). We expect clustering of nursing homes in the population with at least 100 beds to be more pronounced than the geographical clustering of those with less than 100 beds. Travel costs for data collection at the nursing home level were estimated by bed size. The larger nursing homes would be more likely to be located in urban areas than in rural areas. The travel expense for data collection for smaller nursing homes (i.e., <100 beds) is expected to be somewhat greater than the travel expense for data collection in larger nursing homes.

In recent years, data collection for the NNHS has been conducted by the U.S. Bureau of the Census. The sampled nursing home facilities are geographically dispersed throughout the United States, and are not necessarily located in Census Primary Sampling Units. The travel cost for a one-day data collection visit by a single interviewer were estimated according to the distance between the interviewer's residence and the sample nursing home. Table 3 shows the estimated distribution of distances by nursing home bed size.

Table 3. Estimated Distribution of Distances Between Interviewer's Residence and Sample Nursing Home by Bed Size

Distance	Nursing Home Bed Size	
	< 100	>= 100
	Proportion	
25-99 miles	0.25	0.30
100-149 miles	0.30	0.55
150 miles or more	0.45	0.15

The cost for interview training generally falls into several areas. One area is the cost for the interviewer and another area is the cost for staging the training, but the cost for staging the training is considered an overhead (or fixed) expense. The remaining area is the time required for interviewer home study prior to training. The interviewer travel costs for training include wages, travel to airport, airfare, travel to training site, per diem, and home study.

The costs at the nursing home level include several expenses. These facility expenses include several activities. Table 4 provides an estimate of the time needed to complete the listed activities.

Table 4. Facility Expense Category (Less than 100 Beds)

Purpose	*Time (min)*
Telephone Prescreening	5
Induction for Data Collection	60
Round-trip Travel	**
Introduce Survey to Administrator	15
Introduce Survey to Respondent	12
Prepare CR Sample List	20
Prepare DR Sample List	20
Data Collection for FQ	20
Printing Forms	**
Editing Forms	10
Down Time for Staff Lunch	60
Training	**
Post Site Visit Activities	20
Keying Forms	20
	262
Number of hours	4.37

**Note: The time needed for round-trip travel, printing forms, and training extends over several days.

Many of the larger facilities (i.e., at least 100 beds) can provide computer listings of residents and discharges. It is therefore expected that less time will be needed to prepare the sample lists. This results in about 30 minutes less time to complete the tasks at the larger facilities (4.03 hours), assuming one interviewer for each facility.

The cost for an ultimate sampling unit (USU) is the survey cost for a sample current resident or a sample discharge. There are three components for the cost for a resident USU: the cost for data collection, the cost for field edit, and the cost for data keying. According to current information, the direct data collection burden for a sample current resident (CR) is 10 minutes. An estimated 10 minutes will be required for keying each survey CR questionnaire, and about 5 minutes is needed for the interviewer to examine each completed questionnaire and sampling list for completeness. The total survey cost for a current resident ultimate sampling unit
(=1 completed current resident questionnaire) was estimated by multiplying the interviewer's hourly wage by the time required for each component. The data collection cost for a discharged resident sampling unit was estimated in the same fashion.

The cost data for the model came from reports provided by the U.S. Bureau of the Census' Cost and Response Management Network (CARMN). CARMN reports costs separately for individual surveys conducted by the U.S. Bureau of the Census, (Shimizu, et al., 2001). These reports provide cost data by survey task and include salaries, training, mileage, per diem, travel, and telecom services.

To allocate the variable costs, we used 1999 NNHS data and total expenses for CARMN tasks at the nursing home level. Using these data, if the cost per resident case was $1.00, the simple model for variable costs would be:

Variable costs = $1.00 (number of current residents)
+ $0.84 (number of discharged residents)
+ $3.30 (number of facilities with fewer than 100 beds)
+ $5.95 (number of facilities with at least 100 beds).

Using this model, doubling the number of residents sampled per facility is financially cheaper than doubling the number of facilities. However, the facility response rate may be adversely affected by the extra survey burden for the individual facility.

Based on these variable costs, several resident sample allocations were assessed using varying nursing home sample sizes. Although estimated costs for each sample are not displayed in Table 5, the effective sample sizes for the current level of resident sampling (up to six residents per home), when the sampled number of facilities vary are shown.

Table 5. Effective Sample Sizes Expected in NNHS

Units	Full Sample	2/3 Sample	1/2 Sample
Nursing Homes	3000	2000	1500
Inscope	2940	1960	1470
Responding			
Facility Q	2852	1901	1426
CR List	2766	1844	1383
DR List	2652	1768	1326
Current Residents			
Responding	15436	10290	7718
Discharged Residents			
Responding	14481	9654	7240
Total Residents	29917	19944	14958

Using the varying sample allocations from the above table, the expected relative standard errors were estimated using the health characteristics that were described in Section 6.

Depending on the variable of interest, the greatest decrease in variances will come by increasing the number of facilities due to clustering of resident characteristics by facility (i.e., the residents in a facility tend to be more like each other than residents in other facilities).

9. Summary

Sample design research has been conducted in preparation for implementation of a redesigned NNHS in 2003. Data from the NNHS were used to study the utilization of nursing facilities. This information will be used to redesign the survey, and to support research directed at finding effective means for treatment of long-term health problems.

References

Gabrel C., Jones A. *The National Nursing Home Survey: 1997 Summary*. National Center for Health Statistics. Vital Health Stat 13(147). 2000.

Jones A. *The National Nursing Home Survey: 1999 Summary*. National Center for Health Statistics. Vital Health Stat 13(152). 2002.

Shimizu, I. *The 1985 National Nursing Home Survey Design*. Proceedings of the American Statistical Association Section on Survey Research Methods, 516-520. 1986.

Shimizu, I., Lan F. *Approximation of Variable Costs for the National Health Interview Survey* . Proceedings of the American Statistical Association Section on Survey Research Methods, 2001.

Appendix

Table 1. NNHS Proposed Demographic Subdomains

Age	**Hispanics**		**Blacks & Others**		**Whites**	
	Male	Female	Male	Female	Male	Female
<65	X	X	X	X	X	X
65-74	X	X	X	X	X	X
75-84	X	X	X	X	X	X
85+	X	X	X	X	X	X