THE NEW STATISTICAL DESIGN OF THE NATIONAL HOSPITAL DISCHARGE SURVEY

Iris M. Shimizu, National Center for Health Statistics
6525 Belcrest Road, Room 915, Hyattsville, Maryland 20782

KEY WORDS: Sample design

1. Introduction

The National Center for Health Statistics (NCHS) has continuously conducted the National Hospital Discharge Survey (NHDS) since 1965 to collect data on inpatient episodes at short-stay, non-Federal hospitals in the U.S.A. It is the only source of nationally representative statistics on the characteristics of discharged patients, lengths of stays, diagnosis, surgical and non-surgical procedures.

In 1988, the NCHS implemented a redesigned NHDS sample in order to link the NHDS with other NCHS surveys and to improve survey efficiency through the use of technologies that were not available when the survey was first designed in the early 60's. This paper discusses the new design.

2. Background and objectives of redesign

As a result of the National Health Survey Act of 1956, a family of separate data systems was developed to meet the multiple needs for health statistics in the public and private sectors. These surveys evolved under the the National Center for Health Statistics (NCHS), which was formed in 1960, when responsibility for the National Health Survey and the National Vital Statistics Program was merged. Now, the NCHS is moving from a series of independently designed surveys to an integrated survey design.

The integrated design is directed towards linking the Center's surveys. The focus for this coordination is the National Health Interview Survey (NHIS), NCHS's main population based survey. The NHIS has been in operation continuously from 1957 to the present time as the principal source of information on the health of the civilian, non-institutionalized population of the U.S. A weekly survey, the NHIS collects data each calendar year from over 40,0000 households, including about 110,000 persons.

The redesigned NHDS is the first of the establishment surveys to be coordinated with the integrated design of the Center's population-based surveys. As part of the integration, the new sample design was to permit production of statistics for individual primary sampling units (PSU's) as well as for the nation and the four Census geographic regions.

In order to reduce both survey costs and response burden, the new design also shifts the NHDS from sole dependence on manual methods used since 1965 for abstracting discharge records to the purchase of some of these data in automated form from hospital abstracting service organizations. From the survey's beginning, data have been collected via a system in which the discharge sampling and data transcription are done manually within sample hospitals. Starting in 1985, some data have also been collected via a system in which NCHS purchases computer tapes containing discharge medical abstracts from commercial abstracting services and samples discharges from those tapes.

3. Target population and sampling frame

The NHDS universe consists of non-institutional, non-Federal hospitals in the 50 States and the District of Columbia. To be in the universe hospitals must be short stay (have an average length-of-stay for all patients of less than 30 days) and have six or more beds staffed for inpatient use. In the redesigned NHDS the universe was expanded to include general hospitals, regardless of their lengths of stay.

The NHDS sampling frame consists of 6,400 hospitals in the NHDS universe that were listed in the April 1987 SMG Hospital Market data tape (SMG Marketing Group Inc., 1989). For each hospital, the SMG file provides information on hospital location, ownership, type-of-service, beds, and data on admissions, births, and days of care for some 12 month reporting period. NCHS added other items to the file for each hospital. Hospitals were assigned to the primary sampling unit (PSU for their location in the universe for the National Health Interview Survey). Also for sampling purposes five classes of hospitals were defined by hospital specialty and size as shown in Table I.

Because the sampling design uses the volume of discharges typically expected in 12 months of business at each hospital, a 12-month total for discharges was imputed for hospitals which had not yet been admitting inpatients for a full 12 months or whose annual counts of admissions and births were otherwise unavailable. The imputed "annual" count of discharges was calculated by multiplying the hospital's bed size by 10 if the hospital was in the first specialty-size class and by 30, otherwise. The factors of 10 and 30 were observed in the SMG file to be rough minimums for discharges annually from hospitals in the respective specialty-size classes.

The hospitals in the sampling frame were also labeled as subscribers to abstracting services if they were found on one of the then current clientele lists obtained from commercial abstracting service organizations. Research prior to the redesign of the NHDS had shown that, while a hospital may stop subscribing to a particular service, hospitals tended to switch to an alternate abstracting service organization. Only a few switched to use of in-house (owned by the hospital) computers and none reverted to manual methods for processing their discharge medical records.

4. Sampling Design

The NHDS sample includes, with certainty, all hospitals with 1000 or more beds or 40,000 or more discharges annually. The remaining sample of hospitals is based on a stratified three-stage design.

The first stage consists of 112 PSU's that comprise a probability subsample of PSU's used in the 1985-94 National Health Interview Survey (NHIS). The PSU's are counties or groups of

counties or county equivalents (such as parishes or independent cities) or towns and townships (for some PSU's in New England). The NHDS sample includes, with certainty, the 26 PSU's with the largest populations according to the 1980 Census. In addition, the sample includes half of the next 26 largest PSU's, and one PSU from each of 73 PSU strata formed from the remaining PSU's for the NHIS sample design. Those 73 PSU strata were defined within four geographical regions and MSA or non-MSA status by using 1980 Census of Population data and a computer program that minimized the between PSU variances for NHIS stratification variables. (MSA is a metropolitan statistical area defined by the U.S. Office of Management and Budget on the basis of the 1980 Census.) From the 73 strata thus formed, the PSU's were selected with probability proportional to the projected 1985 population. For details of the NHIS PSU sample design, see NCHS, et al. (1989).

The second stage consists of non-certainty hospitals selected from the sample PSU's. At this stage, methods for maximizing the overlap between the old and new NHDS hospital samples were considered because of the potential for maximizing response rates and minimizing start-up costs in the new sample. Those methods were not used, however, because they require the probabilities of selection to the old sample for all hospitals in the universe. Except for hospitals selected to the old sample, it is not possible to determine every hospital's sampling stratum (and, hence, its probability of selection) when it was subjected to sampling for the old NHDS sample.

The sampling design that was used at the second stage differed depending on whether the hospitals were located in certainty PSU's or in non-self-representing sample PSU's. The certainty PSU's were collapsed and the hospitals selected across PSU's when practical, instead of within PSU only. Primary strata of hospitals in the certainty PSU's were defined by region (Northeast, Midwest, South, and West) and size of the PSU (the 12 largest PSU's versus the remaining certainty PSU's) containing the hospital. In order to permit over-sampling of hospitals where costs and burden would be reduced, hospitals in the dominant region of the 12 largest PSU's were further stratified by abstracting service status (whether or not the hospital subscribed to a commercial abstracting service). In each of the region-straddling certainty PSU's, hospitals which were located outside the dominant region for their PSU were made a separate primary stratum to assure representation in the sample from these areas for both regional and PSU estimates. Some individual PSU's were also made separate primary strata when necessary to assure a minimum number of sample hospitals from each to permit PSU estimates. A total of 17 strata were thus formed from the non-certainty hospitals in the certainty PSU's.

Secondary stratification was accomplished within the primary strata by arraying stratum hospitals in a specific order prior to systematic sampling of those hospitals. Within the constraints of assuring geographic distribution of the sample, the ordering was designed to maximize the potential for at least some overlap between the new and old samples and to sample speciality class hospitals in proportion to the discharges they generate. Hospitals within each primary stratum were first arrayed by PSU, where applicable. Within PSU, hospitals were arrayed by old-sample-response status (inscope and responding in the 1987 NHDS versus all others) crossed with hospital class group (first specialty-size class versus all the other classes). Within these groups, hospitals were arrayed by abstracting service status (unless this status was used to define primary strata) and within abstract service status, by specialty-size class. Within specialty-size class, hospitals were arrayed by type-of-service and within type-of-service by annual numbers of discharges occurring at the hospitals. From the ordered array in each stratum, hospitals were selected using systematic random sampling with probability proportional to size (pps) where size was the annual number of discharges.

For the hospital sample from non-self-representing PSU's, hospitals in each PSU were arrayed in the same order as used for hospitals within PSU's in the certainty PSU strata. To allow for expected non-response of 15-30% in the new sample and still have two respondent hospitals needed to compute variances of PSU estimates, three hospitals were then selected from the ordered array in each PSU. Again systematic random sampling with pps was used with size being the annual number of discharges. In PSU's containing fewer than three hospitals, all the hospitals in the PSU were included in the sample.

For 1988, the sample consisted of 542 hospitals. Of the 542 hospitals, 11 were found to be out of scope because prior to 1988 they went out of business or otherwise failed to meet the criteria for the NHDS universe. Of the 531 inscope hospitals, 422 hospitals responded (NCHS collected data for at least half of the number of sample discharges expected in half or more of the months these hospitals were inscope). The number of hospitals in the universe, the sample, and the responding sample are shown by region in Table II.

At the third stage a sample of discharges from each hospital was selected by a systematic random sampling technique. For hospitals using the manual system of data collection, the discharges were selected at the hospital or its abstract service agent from daily listing sheets, computer files, or other lists in which discharges were listed in some chronological order. For most of these hospitals, the sample discharges were selected on the basis of the terminal digits(s) of the patient's medical record number. In some cases, an admission number, billing number, or other number was used. If none of the available patient numbers were useful for sampling purposes, the sample was selected by starting with a randomly selected discharge and taking every kth discharge thereafter.

For hospitals whose data were collected via the automated system, the discharges were selected by NCHS from discharge medical abstract files after sorting the records in those files. The records were first sorted on the first two digits of the ICD-9-CM code of the first-listed diagnosis. Within the diagnostic codes, the records were sorted by patient age group at time of

admission (under 1 year, 1-14 years, 15-44 years, 45-64 years, 65-74 years, 75-84 years, 85 years and over, and age unknown). Within age group, the records were sorted by sex and within sex by date of discharge. These samples were selected by starting with a randomly selected discharge and taking every kth discharge thereafter.

The third stage sampling rate was determined by the hospital's sampling stratum and the system (manual or automated) used to collect data from the hospital. One percent and five percent of discharges in the certainty hospitals were selected under the manual and automated systems, respectively. Except for certainty hospitals, the target sample size was 250 discharges each from all manual system hospitals and from the automated system hospitals which had fewer than 4,000 discharges annually according to the 1987 sampling frame data. Samples of 2,000 were targeted for each of the remaining non-certainty automated system hospitals. The final sample for 1988 included about 250,000 discharge medical record abstracts.

5. Data collection and processing

Two data collection procedures were used for the survey. One was a manual system of sample selection and data abstraction. The other was an automated method, used with approximately 37 percent of the respondent hospitals in 1988, that involved the purchase of data tapes from abstracting service organizations.

In the manual system, the sample selection and the transcription of information from the hospital records to abstract forms were performed at the hospitals and the completed forms, along with sample selection control sheets, were forwarded to NCHS for coding, editing, and weighting. A few of these hospitals submitted their data via computer printout or tape. Of the hospitals using the manual system in 1988, about two-thirds had the work performed by their own medical records staff. In the remaining hospitals using the manual system, personnel of the U.S. Bureau of the Census did this work on behalf of NCHS. For the automated system, NCHS purchased tapes containing machine-readable medical record data from abstracting service organizations. Upon receipt of these tapes, NCHS selected the sample discharges.

For each sampled discharge, data were collected via either the manual or the automated system on items relating to personal characteristics of the patient, including birth date, sex, race, ethnicity, marital status, ZIP Code, (but not name and address), and expected sources of payment; administrative information, including admission and discharge dates, discharge status, and medical record number; and medical information, including diagnoses, surgical and nonsurgical operations or procedures, and dates of surgery. These data items conform with the Uniform Hospital Discharge Data Set, or UHDDS (National Archives and Records Administration, 1985). The PSU, hospital name, medical record number, and patient birth date and ZIP Code are confidential information and are not available to the public.

6. Medical coding and editing

The medical information recorded on the sample patient abstracts that were collected by the manual system was coded centrally by NCHS staff. A maximum of seven diagnostic codes was assigned for each sample abstract. In addition, if the medical information included surgical or nonsurgical procedures, a maximum of four codes for these procedures was assigned. The coding system currently used for coding the diagnoses and procedures on the medical abstract forms, as well as the data that appear on the commercial abstracting services data tapes, is the International Classification of Diseases, 9th Revision, Clinical Modification, or ICD-9-CM (Public Health Service and Health Care Financing Administration, 1980).

Following conversion of the data on the medical abstract to computer tape and combining it with the automated data tapes, a final medical edit was accomplished by computer inspection and by a manual review of rejected abstracts. If the sex or age of the patient was incompatible with the recorded medical information, priority was given to the medical information in the editing decision.

When the patient age or sex was not stated (about 2 percent of the sample discharges in 1988) the patient was assigned an age or sex consistent with the age or sex of other sample patients with the same diagnostic code. Race was not imputed if it was missing (9 percent of the sample discharges in 1988). If a date of admission or discharge was missing (0.08 percent of sample discharges in 1988) a length-of-stay was imputed by assigning the patient a length-of-stay characteristic of the stays for other patients in one of the 4 major age groups or for newborns or women with delivery.

In addition to the edits performed by NCHS, data obtained through the automated system may have been edited and imputed by an abstract service. The extent of this imputation, if any, is unknown.

7. Estimation procedures

7.1. Basic estimator

The basic estimator consists of the basic inflation weight and adjustments for nonresponse. The basic inflation weight is the inverse of the product of the probabilities for the sampling stages, that is, the probability of selecting the PSU, the probability of selecting the hospital, and the probability of selecting the discharge within the hospital.

NHDS data were adjusted to account for two types of nonresponse. The first type of nonresponse occurred when an inscope (NHDS eligible) sample hospital was nonrespondent for more than half of the months during which the hospital was inscope, thus becoming a non-respondent hospital. In this case, the weights of discharges from respondent similar hospitals were inflated to account for discharges represented by the nonrespondent hospitals. For this purpose, hospitals were judged similar if they were in the same region, hospital specialty-size group and, if possible, the same sampling stratum (for example, the same abstracting status group if the

nonrespondent hospital was in the 12 largest PSU's or the same PSU if the hospital was in a non-self-representing PSU). The adjustments for this nonresponse were made separately for admission types--that is, for discharges to newborns (whose hospital stay began with their own birth to a hospital inpatient) and for discharges to other than newborns. The adjustment consisted of a ratio for which the numerator was the weighted number of discharges of the admission type in all similar sample hospitals (regardless of response status) and the denominator was the weighted total of discharges of that admission type from the respondent similar hospitals. Data on the annual number of discharges for each admission type for each hospital came from either the hospitals or the current SMG Hospital Market Data Tape. The current SMG tape for a data year is taken to be that which is released in April of the subsequent year. For example, the 1989 tape was used for 1988 estimates.

The second type of nonresponse occurred when NCHS failed to collect all the discharge abstracts expected (the number expected is the product of the hospital's total discharges each month and the discharge sampling rate assigned to the hospital). In each month when the hospital was respondent (at least half the expected abstracts were collected), the weights of abstracts collected for the month were inflated to account for the missing abstracts. For a hospital's month(s) of nonresponse, the weights of discharges in the hospital's respondent months were inflated by ratios that varied with discharge groups defined by the ICD-9-CM diagnostic classes of those discharges' first listed diagnoses. The adjustment ratio for each partially respondent hospital and each discharge group was calculated using only data from sample hospitals which were both NHDS eligible and respondent for all 12 months of the data year. The ratio's numerator was the weighted sum of discharges in that discharge group for all months in which the partially respondent hospital was inscope, and the ratio's denominator was the weighted sum of discharges which occurred in that discharge group during the months when the partially respondent hospital did respond to the NHDS.

7.2. Final estimator

The final estimator (used to produce the published estimates) consists of the basic estimator with a population weighting ratio adjustment that is applied separately for newborns and for other than newborns. These adjustments were made within each of 16 noncertainty hospital groups defined by region and hospital specialty-size classes to adjust for over- or under-sampling of discharges reported in the sampling frame for the data year. For discharges other than newborn infants, the adjustment is a multiplicative factor that had as its numerator the number of admissions reported for the year at sampling frame hospitals within each region-specialty-size group and as its denominator the estimated number of those admissions for that same hospital group. The adjust-ment for discharges to newborns was similar but numbers of births were used in place of admissions. The ratio numerators were based on the figures obtained from the current SMG Hospital Market Data Tape and the ratio denominators were obtained through a simple inflation of the SMG figures for the NHDS sample hospitals.

8. Variance procedures

For the redesigned NHDS, the multistage, unequal probability sampling design and the ratio adjustment complicate variance estimation. Rust (1985) and Wolter (1985) review methodologies to estimate the variances for complex surveys. Of the methods available, linearized Taylor series procedures were chosen to estimate the variances for the redesigned NHDS.

There are several reasons for selecting the linearized Taylor series methods for the NHDS variances. The primary reason is that the NCHS balanced repeated replication (BRR) computer procedure (NCHS HES-BRR) assumes that units are selected independently at the first stage where units are sampled (without certainty), whereas the NHDS sample is selected without replacement. Theoretically, this assumption of independence can lead to considerable overestimates of the sampling variance for the NHDS. Linearization does not require an independence assumption. Second, linearization methods are computationally simple. They also allow greater flexibility in analyzing operational and statistical aspects of survey design and estimation methodologies.

To compute variances, each stratum must contribute at least two PSU's to the sample. Because the NHDS sample contains only one PSU from each non-certainty PSU stratum, it was necessary to collapse strata to form pseudo-strata of multiple sample PSU's. Each pseudo-stratum was defined to consist of one of the following:

A. A self-representing hospital.

B. A stratum used for selecting hospitals from self-representing PSU's.

C. Two or three non-self-representing PSU strata that would be collapsed for variance purposes in a half-sample design for the NHIS. These were collapsed within region.

Three PSU strata were collapsed into a single pseudo-stratum only when there was an odd number of non-self-representing PSU strata within a region. The strata were collapsed by combining strata that were similar with respect to the original NHIS stratification variables. (Massey, et al 1989).

The final NHDS estimator is not a linear estimator. Hence, a mathematical expression for its variance is not tractable. Woodruff (1971) suggests procedures for linearizing estimators of aggregates so that variances of the linearized estimators could be used to estimate variances for the original estimator. However, such linearization of the estimator is impractical.

Variances were instead approximated using the basic (non-ratio adjusted) NHDS estimator for aggregates, which expressed generically is:

$$X' = \sum W x ,$$

where W is the non-ratio adjusted (basic) weight

computed for each sample discharge according to the estimator. The basic estimator X′ is a linear estimate if the basic weight can be treated as the inverse of the probability of selection. That is, the responding hospitals can be considered as being selected at an extra sampling stage with probability techniques from their respective sampling strata. Likewise, abstracted discharges can be considered as being selected by probability methods at an extra sampling stage within each hospital.

The linearity of the basic estimator facilitates mathematical formulation of variances. The variances for the basic estimates overstate the theoretic values of the variances for the final estimates to the extent that the ratio adjustment in the final estimator reduces variances.

Due to the large number of statistics from the survey, it is impractical to present variances for every statistic. Hence, a generalized variance function is produced for each class of aggregate statistics by fitting curves to points whose coordinates are survey estimates X′ and their corresponding estimated relvariances $[= S^2(X')/(X')^2]$. For each class of statistics, a sample of 100 points are selected from tables planned for publication. After eliminating from the tables duplicate statistics (statistics with the same value and the same standard error) and statistics based on fewer than 30 observations, the remaining statistics are arrayed in order of magnitude. The points corresponding to the ten largest values in the array are then selected, with certainty, and 90 additional points are selected using systematic random sampling from the remaining array.

A curve of the form

$$\text{relvariance } (X') = A + B/X'$$

is then fit to those points by using a weighted least squares approach. The fitted curve and the 100 points are plotted to test for adequacy of fit and to see whether one curve may be used for two or more classes of statistics. These curves and their derivatives are then used to present the sampling errors for statistics from the NHDS.

REFERENCES

National Archives and Records Administration. 1985. Federal Register; Vol. 50, No. 147. Washington: National Archives and Records Administration. 31038-40.

National Center for Health Statistics; Massey, J.T.; Moore, T.F.; Parsons, V.L.; and Tadros, W. 1989. Design and Estimation for the National Health Interview Survey, 1985-94. Vital and Health Statistics. Series 2, No. 110. DHHS Pub. No. (PHS) 89-1384. Public Health Service, Washington: U.S. Government Printing Office.

Public Health Service and Health Care Financing Administration. 1980. International Classification of Diseases, 9th Revision, Clinical Modification. DHHS Pub. No. (PHS) 80-1260. Public Health Service. Washington: U.S. Government Printing Office.

Rust, K. 1985. Variance estimation for complex estimators in sample surveys. J. Official Statistics 1. pp. 381-397.

Shah, B.V.; KaVange, L. M.; Barnwell, B. G.; Killinger, J. E.; and Wheeless, S. C. 1989. SUDAAN: Procedures for Descriptive Statistics User's Guide. Research Triangle Institute, P.O. Box 12194, Research Triangle Park, NC 27709.

SMG Marketing Group, Inc. 1989. Hospital Market Database. Healthcare Information Specialists, 1342 North LaSalle Drive, Chicago, Illinois 60610.

Wolter, K. M. 1985. Introduction to Variance Estimation. New York: Springer-Verlag.

Table I. Hospital specialty size classes for 1988 redesign of the National Hospital Discharge Survey

Hospital Group	Bed size		Discharges annually	Type of service
1	6-999 beds	and	< 40,000 discharges	Selected specialties [1]
2	6-174 beds	and	< 40,000 discharges	General and other specialties [2]
3	175-349 beds	and	< 40,000 discharges	General and other specialities [2]
4	350-999 beds	and	< 40,000 discharges	General and other specialties [2]
5	1,000 or more beds	or	≥ 40,000 discharges	All specialties

[1] Includes psychiatric; tuberculosis and other respiratory disease; rehabilitation; chronic disease; mentally retarded institution; alcoholism and other chemical dependency; and children's psychiatric.

[2] "Other specialties" include: obstetrics and gynecology; eye, ear, nose, and throat; orthopedic; other specialty; children's general; children's TB and other respiratory disease; children's eye, ear, nose and throat, children's rehabilitation; children's orthopedic; children's chronic disease; and children's other specialty.

Table II. Hospitals in the National Hospital Discharge Survey universe and sample and the number of sample hospitals that were inscope[1] and respondent[2] by geographic region: United States, 1988

Hospital Region and size	Universe	Total sample	Sample inscope[1]	Respon-dents[2]	Response Rate
All hospitals	6400	542	531	422	.79
Region					
Northeast	931	117	116	101	.87
Midwest	1797	120	118	87	.74
South	2458	219	215	174	.80
West	1214	86	82	60	.73

[1] Excludes hospitals which, for the whole year, either were out of business or failed to meet the definition of a general, a children's general or short-stay hospital.

[2] Hospitals for which NCHS collected data for at least half of the number of sample discharges expected in half or more of the months the hospitals were in scope.