DATA EVALUATION AND METHODS RESEARCH

Series 2
Number 39

Development of the Design of the NCHS Hospital Discharge Survey

DHEW Publication No. (HRA) 77-1199

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Public Health Service
Health Resources Administration
National Center for Health Statistics
Rockville, Maryland

NATIONAL CENTER FOR HEALTH STATISTICS

Library of Congress Catalog Card Number 70-605820

PREFACE

This report contains a detailed description of the survey design, estimating techniques, and quality control devices employed in the National Center for Health Statistics' continuing Hospital Discharge Survey. Thus it is an account of the technical structure of the undertaking. It is also the story of how that structure was put into place: the initial purposes and objectives, the available resources, the theoretical experimentation and exploration, the determination of unit costs, the balancing of contrasting requirements, efforts to optimize the sampling plan including use of some new patterns of controlled selection, and a scheme for gradual introduction of an increasing number of hospitals and items of data into the sample.

The authors of this report are responsible for design of the survey and the estimating methods which are utilized. As is true of any major statistical program, a number of people have played a part in planning or execution. Particularly substantial contributions were made by Garrie J. Losee and E. Earl Bryant. Robert J. Casady had a major role in setting up the quality control procedures. Special thanks are due the U.S. Bureau of the Census for carrying out several operational aspects of the sampling, and to the Survey Research Center of the University of Michigan for counsel and for use of planning data.

CONTENTS

SYMBOLS

Data not available----------	---
Category not applicable----------	...
Quantity zero----------	-
Quantity more than 0 but less than 0.05----	0.0
Figure does not meet standards of reliability or precision----------	*

DEVELOPMENT OF THE DESIGN OF THE NCHS HOSPITAL DISCHARGE SURVEY

Walt R. Simmons, *Assistant Director, National Center for Health Statistics*
and George A. Schnack, *Assistant to the Assistant Director*

INTRODUCTION

The central mission of the National Center for Health Statistics (NCHS) is the development and maintenance of a set of mechanisms which collectively provide a system of intelligence on matters of health, vital events, health resources, and related affairs. In discharging this responsibility, NCHS by 1965 had established continuing national household interview surveys,[1] standardized physical examination sample surveys of the population,[2] inventory programs of facilities which provide medical, personal, or domiciliary care,[3] sample surveys of selected classes of these facilities,[4] sample follow-back surveys anchored to registration of births and deaths,[5,6] and continued coordination and publication of statistics from the Nation's network of vital event registration agencies. The conduct of these programs was accompanied by research programs which emphasized development of improved vehicles and techniques for data collection and evaluation of already operating mechanisms.

Hospital experience is an increasingly significant component of the medical and health sector of American life. In its household Health Interview Survey, NCHS since 1958 had covered some aspects of this experience. Deliveries in hospitals are reported on birth certificates, and deaths in hospitals are reported on death certificates. But it had been evident from the passage of the Health Survey Act in 1956—and even before—that more information about persons admitted to or discharged from hospitals could contribute substantially to understanding and resolving health problems.

During the 5-year period ending with 1965, many discussions were held regarding the most appropriate character of a statistical reflection of hospital utilization—some of the discussions were internal in NCHS, some included other parts of the Public Health Service, others brought in agencies throughout the Federal Government, and still others took place in public advisory committees. Consideration was given to many matters: first of all, to both general and specific objectives and also to such topics as required tolerances; available relevant data; types and kinds of hospitals to be included in the universe to be studied; relationships among these hospitals, their associations, and the many consumers and potential collectors of data on hospital experience; possible general approaches to data collection; manpower and equipment resources; probable costs; financing; and timetables. When the NCHS Master Facility Inventory[3] was set up, attention was given to the likelihood that it might become a frame for a hospital sample. Concurrently, there was theoretical work on survey design, all leading ultimately to the formal design described in the following pages.

In several places and notably in NCHS reports [6-9] there are published descriptions of the Hospital Discharge Survey (HDS). This report gives additional detail, especially on the estimation processes. But it seeks to do something more—to reveal a part of the path that led to the design and relate why some features of the survey are what they are.

SPECIFICATIONS

The HDS, like most of the other major activities of NCHS, was to be a continuing "general pur-

pose" mechanism rather than a project to answer a single question. Accordingly, specifications tended to be somewhat general and required tolerances were flexible, but some matters were specified in relatively rigid terms. As is true in much survey designing, initial specifications almost necessarily are tentative, since there are likely to be inconsistencies among them in terms of tolerances, budgets, timetables, detail of data, and perhaps other factors. As the design develops, these inconsistencies are compromised.

Items, Scope, and Domains

Balancing desires of consumers for data, opinions regarding the willingness of hospitals to reply to queries of various length and difficulty, and evidence on the content of typical hospital records, the following specific list of objectives was identified as the immediate target of a hospital survey.

A. To estimate total hospital discharges per year for short-stay hospitals, by 28 age-sex groups (usually by 5-year age intervals), by type of discharge (living or dead).

B. To further classify discharges by:

Five length-of-stay classes

25 diagnostic classes

20 classes of operations

A variety of sociodemographic characteristics of persons discharged, including as much information of this type as pretests indicated is available in hospital records

Several geographic and population-density categories: perhaps standard metropolitan statistical areas and the other areas in each of nine census divisions.

C. To estimate and classify number of bed days in a manner similar to that for discharges.

D. To calculate average length of stay for these categories.

E. To collect several pieces of collateral data and tabulate as appropriate:

Numbers of hospitals by size and by type of service provided

Laboratory findings and the results of physician examinations

Financial data (at a later date; not initially), i.e. cost of stay and source of payments.

Time Patterns

The survey should be a continuing one and should cover all time within a reference period rather than a sample of that time. Reporting might be at monthly or quarterly intervals. It was expected that most published data would have a 1-year reference period, but figures might be published more often than annually—perhaps for a moving 12-month period at quarterly intervals.

Level, Trend, and Relationships

A continuing survey may be required to emphasize *levels* of its estimate or *changes* in those levels over time, or possibly *relationships* among statistics for different domains at a given point in time. It was decided that the hospital survey should be designed with first attention being given to "level" in its early years of operation. It was a new survey which was likely to go through a period of definitional and perhaps procedural changes before continuity over time could be established. Quite likely the first emphasis might shift to "changes" in trend in a few years.

Formal Tolerances

As observed earlier, tolerance specifications were to be flexible. In order to give some guide for general order of needed tolerance, two early indicator guides were chosen:

A. For an estimate of 500,000 discharges, the tolerance should be 10 percent with near certainty.

B. For an estimate of $1/100^{th}$ part of all discharges, the tolerance should be within 20 percent with 95 percent confidence.

It appeared shortly that the A-guide was the more demanding at the national level, but at a geographical division level the B-guide was the more severe.

Budget

No realistic budget ceiling was imposed at the beginning of planning. It was thought to be more reasonable to estimate the cost of a sensible program and then to determine whether that could be financed. But as study progressed, it began to appear that an annual budget of around $500,000 would be available, and much of the planning was done with that figure as the assumed variable cost resource.

THE SURVEY DESIGN

General Character

Further discussion of objectives, requirements, and specifications, supported by a pilot feasibility study carried out under contract with NCHS by the School of Public Health, University of Pittsburgh,[10] made clear the broad outlines which the survey plan should follow.

The Hospital Discharge Survey should be a continuing activity from which data on short-stay hospital experience of the civilian, noninstitutional population are obtained through a probability sample of all such discharges, both those alive and those not alive.

The plan should follow a basic two-stage highly stratified sample, the first stage being hospitals chosen from the NCHS Master Facility Inventory, with stratification by geography, size, population concentration, type of hospital, and type of ownership. Large hospitals should have a high probability of being in the sample and small hospitals, a relatively low probability.

Second-stage sampling should be a systematic selection of discharges from sample hospitals. If feasible the overall design should be nearly self-weighting.

The reporting medium should be periodic summary reports submitted by each sample hospital, accompanied by a discharge transcript for each person included in the sample. (See appendix II for illustrations of the forms which finally evolved.)

It would be necessary that an NCHS employee or agent make initial visits to the hospitals to secure cooperation and to assist in instituting reporting from that hospital. Subsequent occasional field visits also would be necessary for surveillance and control of the system.

Blocks and Panels

Preliminary review of likely sample designs indicated that ultimately several hundred hospitals and several hundred thousand transcripts per year would be included. It also appeared that it would be necessary, both for budgeting and for operational reasons, to stretch over several years the introduction of new hospitals into reporting status. Negotiation, training, and initial audit of reports are considerably more time consuming and expensive for a hospital just entering reporting status than for a hospital which is already an experienced respondent. But, useful estimates of hospital discharges were desired at as early a date as possible.

The solution to this problem was found in the concept of "blocks" and "panels" of hospitals. A new specification for design was added. This was that a master sample of hospitals should be formed in such a manner that it could be divided into panels of about 75 hospitals each and so that each panel would be a probability sample of the United States. Then a current operating sample would consist of one, two, three, or more panels of hospitals, with each current operating sample being itself a probability sample of the U.S. short-stay hospitals. One or more of the panels would be introduced into the sample each 6 months. For most purposes the terms "panel" and "block" can be used interchangeably in the HDS. But in order to give appropriate treatment to the very largest hospitals, the 18 hospitals with more than 1,000 beds each were designated "block 1" and scheduled for inclusion in the master sample with certainty. Block 1 plus block 2, the first noncertainty group, were integrated into panel 1 or the "lead panel" for the HDS. They consituted the first group of hospitals to become part of the Hospital Discharge Survey. Block 3 became panel 2, block 4 became panel 3, and so on.

Births, Deaths, Successors-in-Interest

As in any continuing sample survey of establishments, there was a host of problems concerning the identity, birth, merger, or dissolution of establishments, and these had to be resolved. They are among the most difficult problems in establishment sampling, and they cannot be treated adequately in a short space. The general principles followed in the HDS are stated in items C and E in the following section.

The Overall Plan for Sampling

A. The initial master sample would be drawn with the Master Facility Inventory as of a given date being the frame.

B. The master sample would be divided into blocks and panels, combinations of which would be the main sample at any stage of collection of data.

C. Questions of mergers, splits, reorganizations, successors-in-interest, and other changes in ownership or structure would be resolved under this principle:

   If the "acquiring" unit was in the original frame, then it and any subunit of which it became a parent continued, insofar as the sample and estimation were concerned, to be the original frame unit; any unit which was "acquired" by another lost its own identity and assumed the characteristics of a death.

D. Estimates for new hospitals not found in the original frame would come from a supplementary sample of such births treated as a birth stratum.

E. Deaths of hospitals would be reflected by zero measures for dissolution of an establishment in either the main sample or in the birth sample.

THE SAMPLE: EXPLORATION AND DESIGN

Principal Components

A stratified, two-stage master sample of hospitals and discharges had been decided upon, and along with it, there had been chosen a general scheme of data collection—at least the broad outline of required items of output and the domains for which estimates would be produced. Within that framework, the sample was to be designed. The principal components which would guide the designing were: total budget; unit costs—especially the cost per hospital and the additional cost per discharge; dimensions and boundaries of stratification; the key statistics on which optimization should be based; population variances of key statistics; the type of statistics, e.g., aggregates or ratios, and levels or trends; necessary tolerances; a basic model to permit formal and rational design determination; and finally, streamlined models which are geared to available resource data and computational capacity and which still are reasonably faithful to the overall model. Condensed summaries of the manner in which these components were treated are presented in the following paragraphs.

Budget

As observed earlier, while total available budget was somewhat uncertain, much of HDS survey planning assumed a budget of $500,000 to cover survey costs exclusive of overhead and fixed components. This figure had a significant impact on sample size, but the theory would have been little altered if actual budget had ultimately shown a different figure.

Unit Costs

One of the weakest links in the chain of reasoning which leads to optimum sample designs is the determination of unit costs for the various steps in data collection. Among the particularly difficult aspects of cost accounting for continuing surveys, and one of the more arbitrary features, is the period of amortization of initial, or "capital," outlays. In the HDS, the important initial outlay—beyond the fixed or nonvolume-related items—is the cost of introducing a hospital into the reporting panel. The introductory cost is of the order of four times the annual current cost of maintaining the hospital in the sample. Since both the level of total budget and the ratio of "per hospital" to "per discharge" costs have substantial impacts on sample design, it makes a real difference how the

introductory cost is handled in computations. For instance, if all introductory cost is charged to the first year, the unit cost per hospital is three times what it would be if that cost were amortized over a period of 10 years. The contrast would be more striking if, as has been considered, a decision had been made to install special transcribing equipment in each sample hospital. Neither of these two courses is the best choice: current charging is unrealistic, while 10-year amortization is scarcely feasible in the context of Federal government appropriation practices.

How to resolve this matter is a subject that deserves much more study than the literature of finite sampling suggests it may have received. In the case of the HDS, thoughtful consideration was given to the issue, but in the end the decision rested on arguments of uncertain merit. Many factors received attention, but three were given the most weight.

> Major survey programming is often expressed in cycles of 3-4 years: 1 year to plan and pretest, 1 year to collect data, and 1-2 years to process and publish.
>
> Governmental budgeting operates on planning intervals of 3-5 years, with most emphasis at any given point in time being keyed to a 3-year period.
>
> Sample and survey designs tend to be overhauled at 5-year or shorter intervals, at which time there are new "capital" costs. Thus, the life of a continuing survey may realistically be in the range of 3-5 years. Further, in the HDS it was expected that smaller hospitals would be rotated out of the sample after perhaps 3-5 years of reporting.

On these grounds, the introductory costs were amortized over a 3-year period in setting unit costs.

The absolute levels of unit costs were estimated on the basis of previous general survey experience, subjectively modified for contemplated HDS procedure; experience of NCHS and University of Pittsburgh in pilot surveys of 25 hospitals; and experience of the U.S. Bureau of the Census in contacting the first 60 hospitals chosen in a "lead panel" for the survey itself. These efforts resulted in the following figures:

Item	Lower bound	Central value	Upper bound
Unit cost per hospital (C_1)	$200.00	$250.00	$275.00
Additional unit cost per discharge (C_2)	$1.20	$1.60	$2.00

Actually, the lower and upper bound figures are subjective confidence bands around the expected or central value estimates and are not outer limits. Another set of assumptions upon which a good bit of preliminary planning was based used $165.00 as the value of C_1 and $1.00 as the value of C_2. It ought also to be noted that all these figures are *average* estimated costs. Observed costs varied for a number of reasons from place to place, with size of hospital being the most important differentiating factor.

Key Items for Designing Purposes

In designing multipurpose surveys, it is necessary to focus on a limited number of statistics among the objectives, since not all of the objectives will imply identical designs. In some manner, a few typical and important statistics will be chosen as key items for design computations. In the HDS, the choices were number of discharges, number of nights of care or "patient days," and the average length of patient stay per episode. These statistics were estimated for each of five important diagnostic conditions, which ranged in relative frequency from 0.6 of 1 percent to 12 percent of all discharges. Planners of the HDS were unusually fortunate in having relevant data on these items from not only NCHS efforts in the pilot study but also from unpublished materials in the files of the Survey Research Center (SRC) of the University of Michigan, to which NCHS had been given access. Those materials came originally from records of the American Hospital Association and from the Professional Activities Study directed by Dr. Vergil N. Slee.

Population Variances of Key Items

The pilot study and the data from Michigan SRC files were the basis, too, of estimates of needed population variances for the key statistics, both between hospitals and within hospitals. For optimization, the ratio of within- to between-relvariance is of equal or greater significance.

For each of the five diagnoses, the within-relvariances (W^2), the between-relvariance (B^2), and the variance ratio ($W^2/B^2 = \overline{VR}^2$) were assembled for each of a variety of hospital-size classes, using data from the Michigan SRC study. (NCHS pilot study data were too limited to permit separate estimates for hospital-size classes but did contribute evidence.) There are great differences among the different types of statistics and among the different diagnoses within the same types of statistics. Yet there is also a good bit of orderliness among the observations. For example, using median values of $\overline{VR}^2$ and W^2 among size classes for a given type of statistic and diagnosis, and with the different diagnoses as observations, $\overline{VR}^2$ is a fairly distinct linear function of W^2 for each of the three types of statistic. The fitted relations are:

For number of discharges:

$$\overline{VR}^2 = 30 + 2.2\,W^2$$

For patient days:

$$\overline{VR}^2 = 60 + 1.5\,W^2$$

For length of stay:

$$\overline{VR}^2 = 90 + 6.5\,W^2$$

The majority of values of $\overline{VR}^2$ were in the range 300-700 with some concentration near 500. A selection of observations is shown in the following table, in which the entries in the body are in the format $\frac{W^2}{B^2} = \overline{VR}^2$. These examples give a fair reflection of the pattern that emerged from the many computations.

Disease	Statistic		
	Number of discharges	Patient days	Average length of stay
Malignant neoplasms----	$\frac{175}{0.47} = 375$	$\frac{320}{0.58} = 550$	
Diabetes------	$\frac{140}{0.37} = 375$	$\frac{370}{0.53} = 700$	
Heart disease-		$\frac{340}{0.76} = 450$	
Pneumonia-----			$\frac{60}{0.09} = 700$
Delivery------			$\frac{10}{0.07} = 150$

Basic Formal Model

The framework of the chosen sample structure is a two-stage stratified design in which size of hospital and geographic region are the primary strata and type of ownership and finer geographic classes are secondary strata. First-stage drawing of hospitals is random within strata by controlled selection, and second stage selection of discharges is systematic from lists. Overall probability of selection of a discharge is approximately constant. Allocation of number of sample hospitals to strata is proportionate to number of beds in the stratum. Additional detail on several features of the sample is found in subsequent sections of this report.

Most of the characteristics of the sample, but not all, are reflected in the mathematical model which was adopted. That model considers the estimate

$$x' = \sum_{h=1}^{L} \frac{M_h}{m_h} \sum_{i=1}^{m_h} \frac{\bar{N}_h}{\bar{n}_h} \sum_{j=1}^{\bar{n}_h} x_{hij} ,$$

where M_h is number of hospitals in the h^{th} stratum,

m_h is number of sample hospitals in the h^{th} stratum,

$\bar{N}_h$ is number of discharges in a hospital in the h^{th} stratum (since stratification is by size of hospital, all hospitals within one stratum are considered in the model to be of the same size),

$\bar{n}_h$ is number of sample discharges in a hospital in the h^{th} stratum,

L is number of strata, and

x_{hij} is the measure for j^{th} person in i^{th} hospital in h^{th} stratum.

This estimator has relvariance:

$$V_{x'}^2 = \frac{1}{\bar{x}^2}\left[\sum_h \frac{M_h^2}{m_h}\left(1-\frac{m_h}{M_h}\right)S_{1h}^2 + \sum_h \frac{N_h^2}{n_h}\left(1-\frac{\bar{n}_h}{\bar{N}_h}\right)S_{2h}^2\right]$$

in which $N_h = M_h \bar{N}_h$

$$S_{1h}^2 = \frac{\sum_{i=1}^{M_h}(x_{hi} - \bar{x}_h)^2}{M_h - 1}$$

$$S_{2h}^2 = \frac{1}{M_h}\sum_{i=1}^{M_h}\frac{\sum_{j=1}^{\bar{N}_h}(x_{hij} - \bar{\bar{x}}_{hi})^2}{\bar{N}_h - 1}$$

$$x_{hi} = \sum_{j=1}^{\bar{N}_h} x_{hij}$$

$$\bar{x}_h = \frac{1}{M_h}\sum_{i=1}^{M_h} x_{hi} \quad , \text{ and}$$

$$\bar{\bar{x}}_{hi} = \frac{1}{\bar{N}_h}\sum_{j=1}^{\bar{N}_h} x_{hij} \, .$$

For this model the optimum value of $\bar{n}_h$, the number of discharges per hospital in the h^{th} stratum, is

$$\bar{n}_{ho} = \left[\frac{W_h^2}{B_h^2 - \frac{W_h^2}{\bar{N}_h}} \cdot \frac{C_{1h}}{C_{2h}}\right]^{1/2}$$

where the relvariances and costs have meanings as assigned earlier with the added subscript showing that they apply to the h^{th} stratum. The optimum number of hospitals for stratum h is given by

$$m_{ho} = \frac{x_h W_h}{\bar{n}_{ho}[C_{2h}]^{1/2}} \cdot \frac{C}{\sum_h \left[(C_{1h} + C_{2h}\bar{n}_{ho})\frac{x_h W_h}{\bar{n}_{ho}[C_{2h}]^{1/2}}\right]}$$

in which C, total volume-related budget, is

$$C = \sum_h C_{1h} m_h + \sum_h C_{2h} m_h \bar{n}_h \, .$$

For some calculations this full model was used. But for most exploratory work, a simplified or streamlined version gives approximations which are adequate for comparing different versions of the central design—especially since knowledge of differences among parameters in the different strata is not very precise. In the simplified version, relvariance is approximated by

$$V_{x'}^2 \doteq \frac{B}{m} + \frac{W^2}{n}$$

in which m is total number of hospitals in the sample and n is total number of sample discharges. The cost equation is $C = C_1 m + C_2 n$. The approximate optimizing formulae are

$$\bar{n}_o = \left[\frac{C_1}{C_2} \cdot \frac{W^2}{B^2 - \frac{W^2}{\bar{N}}}\right]^{1/2}$$

and $m_o = \dfrac{C}{C_1 + \bar{n}_o C_2}$.

Stratification

The universe of hospitals, consisting of all short-stay civilian hospitals in the United States, as identified in the NCHS Master Facility Inventory,[3] was classified into 28 primary strata. Within each of the four broad geographic regions, hospitals were placed in one of seven categories classified by number of beds per hospital. The hospital-size classes were as follows:

6-49 beds
50-99 beds
100-199 beds
200-299 beds
300-499 beds
500-999 beds
1,000 beds and over.

Within primary strata there was a further classification by four types of ownership and by nine geographic divisions. In addition to these primary

and secondary strata, there was still further substratification through systematic sampling from the frame in which hospitals were listed by type of service and by State and county sequence within secondary strata. It should be noted that the primary size stratification accomplished in some degree a type-of-hospital classification.

No particular stratification is specified for discharges within a hospital. But usually discharges are arranged in some systematic order so that a pseudostratification is present.

Optimization

Several hundred combinations of cost and other input values were tested along with a central design which was based on the most likely cost figures, central tendency variance parameters, and optimum allocation of resources. The central design called for 720 hospitals and 200,000 discharges annually. (These figures came from the streamlined model and are comparable with figures for alternative designs to be discussed shortly. The full model produced an optimum of 686 hospitals and 209,000 discharges.) The design has a relative sampling error of 3.4 percent for a statistic $P \approx 1.0$ percent, that is, the 95-percent confidence interval for the estimate of P is 0.932-1.068 percent. At the universe level, this is well within specified tolerance requirements.

Table 1 summarizes the central design and 26 alternatives which would result from optimization under a $500,000 variable cost budget and reasonable parameters of unit cost and variance ratio. These results, supported by those from many other trial designs, made it rather clear that the proper number of hospitals almost surely was in the range 600-1,000 and probably was near

Table 1. Optimum sample size and relative standard error for central design and 26 alternate designs under reasonable parameters of unit cost and variance ratio; total variable budget=$500,000

[Standard errors calculated for "typical" statistics[1]]

Hospital unit cost and sample size[2]	Variance ratio = 300			Variance ratio = 500			Variance ratio = 750		
	Discharge unit cost			Discharge unit cost			Discharge unit cost		
	$1.20	$1.60	$2.00	$1.20	$1.60	$2.00	$1.20	$1.60	$2.00
$200									
m_0	1,065	980	915	915	835	770	800	725	670
n_0	240,000	190,000	160,000	260,000	210,000	180,000	280,000	220,000	190,000
V x 100	3.3	3.6	3.9	3.0	3.3	3.6	2.8	3.1	3.4
$250									
m_0	910	835	785	785	720	665	690	625	580
n_0	230,000	180,000	150,000	250,000	200,000	170,000	270,000	210,000	180,000
V x 100	3.5	3.8	4.0	3.1	3.4	3.7	2.9	3.2	3.4
$275									
m_0	850	785	735	735	670	625	645	590	545
n_0	230,000	180,000	150,000	250,000	200,000	170,000	270,000	210,000	180,000
V x 100	3.6	3.8	4.1	3.2	3.5	3.8	2.9	3.2	3.5

[1] See section "Population Variances of Key Items," p. 6.

[2] m_0 = sample size of hospitals

n_0 = sample size of discharges

V x 100 = relstandard error

700-800. About 200,000 discharges should be transcribed annually.

In the course of experimentation a variety of significant pieces of information emerged. Some of this variety is suggested by the following points.

A. The optimum is broad, that is, the sampling error changes slowly as the number of sample hospitals departs some distance from the ideal. For example, for a given set of unit costs and population variances, a 10-percent decrease or increase from the optimum in number of sample hospitals increases the sampling error by less than 1 percent.

B. However, more radical deviations from the optimum could have more drastic impacts. For instance, cutting the number of hospitals to one-half the optimum would increase overall variance by a quarter and would have even sharper effect on many domains. To hold tolerances fixed would nearly double the cost under the design. Increasing the number of hospitals over the optimum by the same margin would have about two-thirds the impact caused by a corresponding decrease in number of hospitals from optimum. Any reduction in number of hospitals makes the estimation of sampling variances more unstable and increases the hazard from always possible boners and mistakes in survey execution. It also makes carrying out a controlled selection more difficult—as will be described subsequently.

C. The optimum number of discharges per hospital is not affected by total overall budget.

D. Numbers of discharges and number of patient days as single statistics would lead to the same central design. If the ratio of length of stay were the determining statistic, the design could consist of about 80 percent as many hospitals and 10 percent more discharges.

E. There is a tendency for common and simple cases, such as pneumonia and deliveries, to require fewer discharges and more hospitals for effective treatment than do some more complex or rarer diseases, e.g., malignant neoplasms and diabetes. This tendency is, however, confounded with the "type of statistic." Perhaps typically, the common diseases may be best handled with 25 percent more hospitals and 10-15 percent fewer discharges than the other reasons for admission.

THE LEAD PANEL

Strategy for Sample Hospitals

Preliminary investigation indicated the HDS sample size would be several hundred hospitals. Prior to making a final determination as to the exact sample size and its composition, however, the decision was made to survey a small number of hospitals. There were two good reasons for doing this. First, as indicated, the optimization was based on fragmentary and assumed information concerning costs and relvariance components and, hence, the optimization might not be accurate enough. Second, for operational reasons a very large survey could not be undertaken at any one time. In order to gain the field experience needed and collect more precise data to evaluate the contemplated design, it was decided that a small number of hospitals, called a lead panel, would be surveyed for a period of time before deciding on the number of hospitals and patient abstracts in the final HDS sample. It was recognized, however, that the lead panel was to become an integral part of the final design after its use as a pilot survey.

This strategy proved to be highly successful. In 1964 a survey of 90 hospitals gave the knowledge and experience necessary to determine a HDS sample of 690 hospitals located in 10 blocks, where:

Block 1.—This contains all 18 hospitals in strata with hospitals having 1,000 or more beds for inpatient use;

these strata are termed "certainty strata."

Block 2.—This contains 72 hospitals chosen from the remaining strata, termed "noncertainty strata," and together with block 1 constitutes the lead panel.

Blocks 3-10.— These contain 75 hospitals each chosen from the frame of hospitals remaining after the selection of hospitals for the lead panel.

A discussion of blocks 1 and 2 is given below, followed by a discussion of blocks 3-10.

Controlled Selection

Determination and allocation of sample size.—The determination of sample size came from the decision that all hospitals having 1,000 or more beds would be included with certainty in the sample in addition to enough hospitals from the remaining frame to yield a few national statistics with a tolerable sampling error and the field experience that was being sought. There were 18 certainty hospitals, while a convenient sample rate in the other strata of one hospital per 10,000 beds would yield a sample of 72 hospitals, giving a lead panel of 90 hospitals. It was thought likely, in the early design stages, that the final sample rate would be about one hospital per 1,000 beds, making the 1-in-10,000 overall sampling fraction easily convertible.

Using the nearest integer value from an allocation scheme of 1 in 10,000 gave, in most instances, two to six sample hospitals in each of the noncertainty primary strata. In those few instances where the number of beds in a primary stratum was less than 20,000, two hospitals were allocated. This step was taken to simplify the estimation of variance components used in the later design.

Selection of sample hospitals.—The sample size allocation to the primary stratum was distributed to the substrata, formed by the type of ownership and geographic division (later called State clusters), within the primary stratum. The latter distribution was proportionate to the number of hospitals in the substrata. Initially because the sample sizes in the substrata were small numbers, these values were not rounded to integers but were retained as calculated to two decimal places. This secondary distribution became the "expected values" in the selection of the substrata, from which a systematic selection from among all hospitals gave the sample of hospitals.

In addition to using the fairly extensive stratification procedures described above, the selection of hospitals also incorporated a modified form of the Goodman-Kish controlled selection technique.[11] This technique allows some element of judgment in obtaining a better sample while retaining the essential characteristics of probability sampling.

The technique permitted the HDS to maintain a constant probability of hospital selection in the primary strata, while, within definite limits, setting the probability of selecting combinations of hospitals representing the different substrata. This is accomplished by assigning high probabilities of selection to favorable sets of sampling units and low probabilities to unfavorable sets. The objective of the technique is to obtain a sample which is more closely representative of the universe than a randomly drawn sample would likely be.

The process used in selecting the hospitals is presented in detail in appendix III and only a brief description is given here.

The lead panel sample of noncertainty hospitals was designed to be 72 hospitals drawn from 24 primary strata having 216 substrata. With a known overall sample size for each primary stratum, it is possible to calculate an "expected number" of hospitals which should fall into each of the substrata if the allocation of sample hospitals is made proportional to the number of hospitals in the frame in that stratum. This expected number is likely to be fractional. Obviously since there are 216 strata and only 72 sample hospitals to be allocated, many of the expected values in the substrata must be less than unity.

It is necessary to work through the entire process, as set forth in appendix III, to get a full understanding of how controlled selection accomplished its objective, but a condensed illustration can clarify some key features. Suppose that for a given primary stratum it had been determined

that two sample hospitals are to be drawn into the sample. The primary stratum is made up of four substrata or cells labeled A-1, A-2, B-1, and B-2; the "ideal" number of sample hospitals from each of these is shown in the table below.

Stratification I	Stratification II		
	Total	Class 1	Class 2
Total----------	2.00	1.07	0.93
Class A----------	0.55	0.15	0.40
Class B----------	1.45	0.92	0.53

Controlled selection is a technique for drawing the hospitals into the sample in such a fashion that the numbers in this table set upper and lower limits and indeed the complete probability distribution for the number of sample hospitals in each cell, row, and column. For example, in the illustration, the probability is 0.40 that there will be one hospital from cell A-2 and 0.60 that there will be none. Similarly the probability is 0.45 that there will be two hospitals from class B and 0.55 that there will be one. It is certain that there will be exactly two from the entire primary stratum.

In subsequent steps, described in detail in appendix III, a consolidation over all 24 primary strata of the control procedure just outlined enabled the final sample to be drawn with a high degree of representativeness assured for geographic region, size of hospital, type of ownership, and State clusters. This was true despite the fact that there were only 72 hospitals drawn from a classification network of 216 cells. Further, the allocation of the number of sample hospitals to the 24 primary strata was proportional to the number of beds in the stratum.

THE MASTER SAMPLE

Sample Size and Allocation

The determination of sample size was based on the best estimates of available budget, unit costs per hospital and discharge, and a ratio of within- to between-hospital relvariance available after one-quarter of a year's survey experience with the block 2 hospitals. These considerations led to a master sample design of 690 hospitals and a quarter of a million patient abstracts annually. As noted in the section "Blocks and Panels," page 3, it was found desirable to allocate sample hospitals to a series of subsamples.

Allocation to representative subsamples was made to that multiple of eight hospitals nearest the optimum, after adjustment for the previous allocation in the lead panel. The reason for this step in the design was to allow the partition of the sample into eight blocks of 75 hospitals each with controlled selection to insure that each block would be representative of the frame.

Selection of Sample Substrata

First and second stages.—The method of controlled selection was used here in the same way as in the block 2 selection in the lead panel. Patterns were formed in each primary stratum and then in each region. For the selection of 600 hospitals, however, a third stage was incorporated.

Third stage.—Each region produced several patterns or combinations of substrata. These patterns in each region then became the group values, as in tables of appendix III, and patterns were joined together over the regions. The control in this phase was the expected number of hospitals in each ownership class of the entire United States.

This final step generated 28 different combinations of 600 substrata each. Using a random number and the probabilities associated with each combination, one was selected. From this combination a random selection of hospitals was made in each substratum. If a substratum appeared more than once in the combination, indicating more than one sample hospital from that substratum, a systematic selection was made.

Controlled selection to form blocks.—One of the requirements in the allocation of sample to the primary strata was that the number of hospitals selected be a multiple of 8. Because of this it was possible to run through the three stages of controlled selection to assign sample hospitals to one of eight blocks in a manner which made each block representative of the HDS frame. This process is presented in appendix III.

WITHIN-HOSPITAL SAMPLING

The sample of discharges is selected in a systematic manner, usually on the basis of the patient's medical record number. The primary numbering systems are termed *unit* and *serial*. In the unit system, the patient receives a number on his first admission and he retains this through subsequent admissions to the same hospital. In the serial system, the patient receives a number on each admission and each medical chart is filed under its own number. Thus, if a patient were admitted to a hospital three times, there would be three serial numbers and three separate records for him. There is also a combination of the two systems, i.e., the *serial-unit* system. There are other special systems, but the unit and serial are the most common.

If the hospital uses medical record numbers and maintains a record of discharges, and if the discharges are listed in such an order that the discharges for a calendar month are readily accessible as a unit distinct from discharges of other months, then the discharge list is used as the sampling frame. A daily list of discharges or a card file of discharges filed by day, week, or month are examples of sampling frames.

The number and identity of the sample discharge cases to be selected from the discharge frame are determined from a table of randomly selected terminal digits. Each abstractor is provided with a table based on the "within hospital" sampling rate, e.g., in a hospital having a "within hospital" rate of 2 out of 10, the abstractor takes all discharges having the terminal digit 2 or 7. Any discharge case for which a patient was discharged in the survey month and which has the terminal digits of its medical record number as specified in the instructions is defined to be an eligible sample case. It becomes, in estimation, a sample case if it is not an out-of-scope discharge.

It is not possible to select discharges by this method from a hospital which does not have a numbering system for its medical records. In such hospitals, it is necessary to use a random start followed by a systematic selection.

ESTIMATION AND SAMPLING VARIANCE

Estimation

Statistics produced by the Hospital Discharge Survey are derived by a complex estimating procedure in which the basic unit of estimation is the patient abstract. The procedure used to produce essentially unbiased estimates has three principal components: inflation by the reciprocals of the probabilities of selection; two levels of nonresponse adjustment; and two levels of ratio adjustment to known totals. These are described in general terms below, and appendix IV contains exact mathematical expressions of them.

Inflation of sample data.—Simple inflation of the abstracted statistical data by the reciprocals of the sampling probabilities takes into account all stages of survey design. Since the survey described in this report utilizes a two-stage design, there are two probabilities—the probability of selecting the sample hospitals and the probability of selecting the patient abstract.

Imputation.—Often, at each stage of the design, there is an attrition of sample units from the survey. Two types of inflation, other than expansion which accounts for the sampling fraction, are used in the Hospital Discharge Survey estimator. One results from abstracting fewer than all patient discharges. Another comes about when sample units do not respond. An inflation of the recorded data by the reciprocals of the response ratios at each stage in the design provides a partial correction for the missing data. There are two response ratios in this survey—a hospital ratio and a patient abstract ratio.

Imputation for nonresponding hospitals is carried out within each of the size-by-region strata for each calendar month. The adjustment is made by a multiplier ratio, the numerator of which is the number of beds in the sample hospitals as recorded in the Master Facility Inventory (MFI) and the denominator of which is the number of beds in those sample hospitals responding for that month. This adjustment has

the effect of imputing to the nonresponding hospitals the information obtained from the responding hospitals.

Each hospital transmits a report that shows the number of discharges which the sampling procedure produces and that should be accompanied by a transcript of each sample discharge record. Sometimes one or more of these transcripts is missing. To adjust for this contingency, each sample record is multiplied by a factor, the numerator of which is the number of transcripts which should have been received and the denominator, the number which were actually received.

Ratio adjustment.—It is well known that a ratio estimate for a statistic is superior to an ordinary inflation estimate if there is sufficient positive correlation between the numerator and denominator of the ratio. This principal is used at both stages of the HDS. Each is discussed briefly.

A first-stage ratio adjustment is included only in the estimation of patients discharged from sample hospitals in the 24 noncertainty strata. The adjusting multiplier ratio is obtained by dividing the total number of MFI beds in a stratum by the number of beds estimated from the sample hospitals in that stratum. The second-stage ratio adjustment is made for each of the responding in-scope sample hospitals for each calendar month for all statistics which were derived from two stages of estimation. The adjustment is made using a multiplier factor obtained by dividing the total number of discharges in a month (as reported by the hospital) by the product of the number of sample discharges in that month and the reciprocal of the within-hospital sampling fraction. The purpose of this adjustment is to correct for deviations from the expected within-hospital sample size.

Composite estimator.—One of the distinctive features of the HDS is that each noncertainty block of sample hospitals yields essentially unbiased national estimates for the universe of hospitals having fewer than 1,000 beds. As can be seen in the equation (3), appendix IV, the final estimator averages the estimates from the several blocks. A slightly better method might have been to make a block's weight inversely proportional to its sampling variance, but the different blocks have sufficiently similar variances so that any gain from such an attempt would be doubtful.

Sampling Variance

A well designed survey includes a plan for discovering what the sample data can say about the reliability of estimates based upon the sample data. Since the data obtained in a survey are subject to variation, estimates using these data are also subject to variation and to some extent are uncertain. The standard error is a measure of this uncertainty and can be used to judge the variation that might occur by chance because only a part of the universe is surveyed. Appendix IV contains detailed equations used in measuring the sampling variance.

QUALITY CONTROL OF DATA

A significant number of HDS resources are used to monitor the quality of the abstracts received from the reporting hospitals and the subsequent coding and processing of the data in them. A quality control procedure which is designed to insure low error rates in the medical coding is typical of these activities and is presented below. Other control activities common to many NCHS surveys have been mentioned in previous NCHS publications and are not treated here.

Medical Coding and Error Rate

Because the diagnostic description of a discharge is particularly important, a rather elaborate scheme was developed to guarantee high quality in the numerical classification and coding of the medical information on the patients appearing in the abstracts. The medical abstracts—items 11 and 12 of appendix II—are received at the data processing center in small groups, usually between 25 and 100 abstracts. Each group is a sample of discharges from a hospital for a calendar month. A "coder assignment batch" of about 2,000 abstracts is formed by arbitrarily combining several of these groups from different hospitals.

The quality control process is founded on three independent manual edits of the abstracts and a machine evaluation of these edits. It proceeds in this manner:

From each batch a systematic sample of abstracts is selected. This sample is normally 10 percent of the batch.

The first two available coders are given the job of independently coding, on separate forms, the diagnoses and operations appearing in the sample.

The next available coder, assuming he is not one of the above, becomes the "production" coder who does an independent coding of the diagnoses and operations appearing in the whole batch.

These data—coder numbers, listed codes, abstract numbers, and batch numbers—are then put on punched cards and machine processed.

Two separate error rates on diagnostic coding are computed for each coder:

(1) based on the first listed diagnostic code.
(2) based on all listed operative codes.

A similar pair of error rates is calculated for operative coding:

(1) based on the first listed operative code.
(2) based on all listed operative codes.

Coding errors, using the first listed code, are assigned to the coders by the following rules:

Rule 1. if all three coders have the same first listed code, then none of the coders receives an error;
Rule 2. if two of the three coders have the same first listed code, while the other coder has a different code, this latter coder receives an error;
Rule 3: if all three coders have different first listed codes then each coder receives an error.

The number of errors, *using all listed codes*, is counted differently. For each abstract coded by the three coders, a set of codes is formed which contains those codes that are given by at least two of the coders. Added to this set are enough dummy codes such that the set has as many codes in it as the abstract with the most listed codes. Suppose, for example, coder 1 listed codes for three diagnoses, coder 2 listed four, and the production coder gave six. The set for this abstract would then contain six numbers, those codes which two or three of the coders had in common plus enough dummy codes to yield a set of six values. This set having been formed, errors are assigned to the coders by one of the following rules:

Rule 4: a coder is given an error for every code he listed which is not in the set; or
Rule 5: a coder is given an error for every code in the set which is not listed by him.

The rule which yields the larger error applies.

Batch and Coder Controls

For individual batch control, a batch is rejected when the coding error rate exceeds 15 errors per 100 codes. If this happens, the batch is 100 percent dependently coded by a coder known to have a low error rate. For individual coder control, a coder is retrained when his coding error rate exceeds by 2σ the average error rate obtained from a set of about 50 batches, i.e., when about 50 batches have been completely processed, all error rate data are sent to a statistician in the HDS. He obtains from these data an average error rate and the variance of the error rates. The 2σ limit is the criterion determining when retraining is required. The statistician also compiles the error rates into a series of graphs like those shown in figures 1 and 2. Graphs like figure 1 are compiled from error rates for production coders who processed the total batch. Graphs like figure 2 are prepared for each coder, regardless of the source of the error rates.

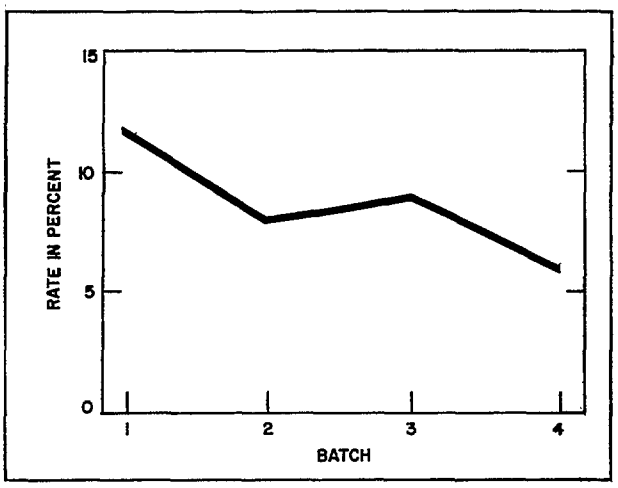

Figure 1. Typical graph of the error rate by batch for diagnosis or operation.

Periodically these graphs are updated and sent with a brief memorandum to the medical coding supervisor summarizing the overall status of the program. In this memorandum the supervisor receives the analysis of the set of batches and is advised when the coders are not in control.

The three independently-assigned codes, along with associated descriptive information on the discharge, when punched or typed into a computer, permit four distinct types of analysis from printed computer output:

1. Summary of the experience of any single interviewer.
2. Time-trend analysis of error rates over all interviewers combined—the usual quality control procedure.
3. Identification of particular diagnostic conditions which have high error rates, thus opening the door to possible retraining of coders.
4. Identification of particular hospitals which have high error rates and accordingly may need special attention.

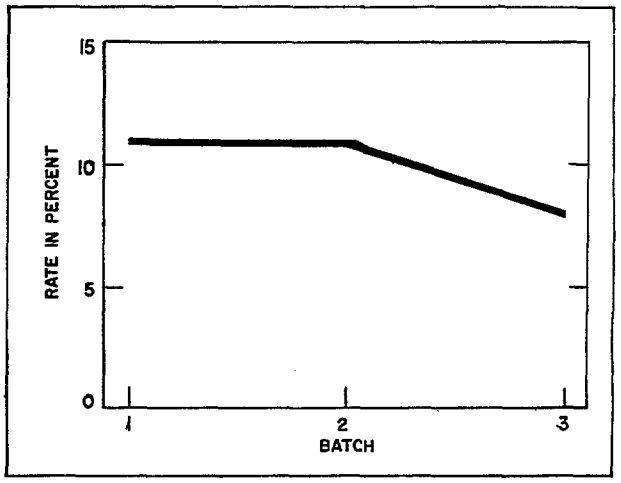

Figure 2. Typical graph of the error rate for coder A.

REFERENCES

[1]U.S. National Health Survey: The statistical design of the Health Household-Interview Survey. *Health Statistics*. PHS Pub. No. 584-Series A-No. 2. Public Health Service. Washington, U.S. Government Printing Office, July 1958.

[2]National Center for Health Statistics: Plan and initial program of the Health Examination Survey. *Vital and Health Statistics*. PHS Pub. No. 1000-Series 1-No. 4. Public Health Service. Washington. U.S. Government Printing Office, July 1965.

[3]National Center for Health Statistics: Development and maintenance of a national inventory of hospitals and institutions. *Vital and Health Statistics*. PHS Pub. No. 1000-Series 1-No. 3. Public Health Service. Washington. U.S. Government Printing Office, Feb. 1965.

[4]National Center for Health Statistics: Design and methodology for a national survey of nursing homes. *Vital and Health Statistics*. PHS Pub. No. 1000-Series 1-No. 7. Public Health Service. Washington. U.S. Government Printing Office, Sept. 1968.

[5]National Center for Health Statistics: Hospitalization in the last year of life, United States, 1961. *Vital and Health Statistics*. PHS Pub. No. 1000-Series 22-No. 1. Public Health Service. Washington. U.S. Government Printing Office, Sept. 1965.

[6]National Center for Health Statistics: Methods and response characteristics, National Natality Survey, United States, 1963. *Vital and Health Statistics*. PHS Pub. No. 1000-Series 22-No. 3. Public Health Service. Washington. U.S. Government Printing Office, Sept. 1966.

[7]National Center for Health Statistics: Utilization of short-stay hospitals, summary of nonmedical statistics, United States, 1965. *Vital and Health Statistics*. PHS Pub. No. 1000-Series 13-No. 2. Public Health Service. Washington. U.S. Government Printing Office, Aug. 1967.

[8]Sirken, Monroe G.: Federal survey collecting data on discharged patients. *Hospitals*. 40. Aug. 1, 1966.

[9]Rossoff, Milton G.: Obtaining patient statistics through a national survey. *Inquiry*. III (3). Sept. 1966.

[10]National Center for Health Statistics: Participation of hospitals in the pilot study of the hospital discharge survey. *Vital and Health Statistics*. PHS Pub. No. 1000-Series 2-No. 19. Public Health Service. Washington. U.S. Government Printing Office, Oct. 1966.

[11]Goodman, J.R., and Kish, L.: Controlled selection — a technique in probability sampling. *J. Am. Statist. A.* 45: 350-372, Sept. 1950.

———o o o———

APPENDIX I

STATISTICS OF THE HOSPITAL DISCHARGE SURVEY MASTER SAMPLE

Table I. Number of hospitals in the frame and in the master sample, by size of hospital and geographic region

Size of hospital	Geographic region									
	Total		Northeast		North Central		South		West	
	Frame	Sample	Frame	Sample	Frame	Sample	Frame	Sample	Frame	Sample
	Number of hospitals									
All sizes	6,965	690	1,107	180	1,979	208	2,620	201	1,259	101
Under 50 beds	3,113	89	199	10	830	26	1,438	35	646	18
50-99 beds	1,623	99	288	18	442	27	587	36	306	18
100-199 beds	1,144	143	277	36	378	45	332	44	157	18
200-299 beds	552	125	182	44	151	36	134	27	85	18
300-499 beds	386	134	110	36	129	44	96	36	51	18
500-999 beds	129	82	42	27	46	27	28	18	13	10
1,000 beds or more	18	18	9	9	3	3	5	5	1	1

Table II. Approximate expected number of annual discharges[1] in thousands in the master sample, by size of hospital and geographic region

Size of hospital	Geographic region				
	Total	Northeast	North Central	South	West
	Number in thousands				
All sizes	281	74	84	81	42
Under 50 beds	28	2	7	12	7
50-99 beds	44	8	12	16	8
100-199 beds	65	17	21	19	8
200-299 beds	49	17	14	11	7
300-499 beds	49	13	16	13	7
500-999 beds	36	12	12	8	4
1,000 beds or more	10	5	2	2	1

[1]Estimated on the basis of number of beds in the hospitals in 1963.

Table III. Approximate sampling rates by size of hospitals in the master sample

Size of hospital	1st stage (hospital)	2d stage (discharges within hospitals)	Overall fraction of discharges
Under 50 beds	1/40	4/10	1/100
50-99 beds	1/20	2/10	1/100
100-199 beds	1/10	1/10	1/100
200-299 beds	1/5	5/100	1/100
300-499 beds	1/3	3/100	1/100
500-999 beds	1/2	2/100	1/100
1,000 beds or more	1	1/100	1/100

APPENDIX II. PRINCIPAL FORMS USED IN THE SURVEY

Discharge Abstract—Front

CONFIDENTIAL- All information which would permit identification of an individual or an establishment will be held confidential, will be used only by persons engaged in and for the purposes of the survey and will not be disclosed or released to other persons or used for any other purpose (22 FR 1687).

PHS-4734-2
REV. 11-66

DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE
PUBLIC HEALTH SERVICE
NATIONAL CENTER FOR HEALTH STATISTICS

Form Approved
Budget Bureau No 68-R620 R2-2

1. HOSPITAL NUMBER

ABSTRACT OF PATIENT RECORD- Hospital Discharge Survey

2.	HDS NUMBER	0 1 2 3 4 5 6 7 8 9
3.	MEDICAL RECORD NUMBER	0 1 2 3 4 5 6 7 8 9
4.a.	DATE OF BIRTH	MONTH: JAN. FEB. MAR APR MAY JUNE JULY AUG. SEPT. OCT NOV. DEC.
		DAY: TENS 0 1 2 3; UNITS 0 1 2 3 4 5 6 7 8 9
	Complete 4b and 4c if date of birth is not given.	YEAR: 1800 1900; TENS 0 1 2 3 4 5 6 7 8 9; UNITS 0 1 2 3 4 5 6 7 8 9
4.b.	AGE	TENS 0 1 2 3 4 5 6 7 8 9; UNITS 0 1 2 3 4 5 6 7 8 9
4.c.	AGE IS STATED IN	YEARS MONTHS DAYS
5.	SEX	MALE FEMALE
6.	RACE OR COLOR	WHITE "NONWHITE" NEGRO NOT STATED OTHER NONWHITE
7.	MARITAL STATUS	MARRIED DIVORCED SINGLE SEPARATED WIDOWED NOT STATED
8.	DATE OF ADMISSION	MONTH: JAN FEB MAR APR MAY JUNE JULY AUG SEPT OCT NOV DEC; DAY: TENS 0 1 2 3; UNITS 0 1 2 3 4 5 6 7 8 9; YEAR 0 1 2 3 4 5 6 7 8 9
9.	DATE OF DISCHARGE	MONTH: JAN. FEB. MAR APR. MAY JUNE JULY AUG SEPT OCT NOV DEC; DAY: TENS 0 1 2 3; UNITS 0 1 2 3 4 5 6 7 8 9; YEAR 0 1 2 3 4 5 6 7 8 9
10.	DISCHARGE STATUS	ALIVE DEAD

IBM H95191

Discharge Abstract—Back

HDS NUMBER

11. FINAL DIAGNOSES:

12a. WAS AN OPERATION PERFORMED ? YES NO

12b. OPERATIONS:

COMPLETED BY ABSTRACTOR _____ DATE _____

FOR NCHS USE ONLY

DIAGNOSIS CODES	OPERATION CODES
(1)	(1)
(2)	(2)
(3)	(3)
(4)	
(5)	

PHS-4734-2 (BACK) REV. 11-66

IBM H88193

Discharge Listing Sheet

PHS-4734-5
REV. 2/68

DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE
PUBLIC HEALTH SERVICE
NATIONAL CENTER FOR HEALTH STATISTICS

FORM APPROVED:
BUDGET BUREAU NO. 68-R0620

SAMPLE LISTING SHEET

Hospital Discharge Survey

A. HOSPITAL	B. STATISTICAL DATA	C. SAMPLING
NAME	TOTAL BEDS (excluding bassinets)	MONTH
NUMBER	TOTAL ADMISSIONS (excluding newborn)	KEY
LIST USED	LIVE BIRTHS	NUMBER IN SAMPLE
	TOTAL DISCHARGES (including newborn)	SAMPLE SELECTED BY_____ DATE

HDS NUMBER	DATE OF DISCHARGE OR (ADMISSION)	MEDICAL RECORD NUMBER	OTHER IDENTIFICATION *(if needed)*	DATE ABSTRACTED OR (OUT-OF-SCOPE REASON)
1	2	3	4	5

Page ____ of _____

Control Sheet and Transmittal Notice

PHS-4734-3
REV. 2/68

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
PUBLIC HEALTH SERVICE
NATIONAL CENTER FOR HEALTH STATISTICS

FORM APPROVED:
BUDGET BUREAU NO. 68-R0620

TRANSMITTAL NOTICE

Hospital Discharge Survey

HOSPITAL NAME AND ADDRESS	DATE OF TRANSMITTAL
	HOSPITAL NO.

Type of abstracts transmitted: Medical ☐ or Ledger ☐

This shipment includes:

Abstracts for _____ (Month) 19____ . . . . . (Number) _____

Back records . . . . . (Number) _____

Total . . . . . (Number) _____

BACK RECORDS SUBMITTED				RECORDS NOT AVAILABLE			
HDS NO.	MEDICAL RECORD NUMBER	HDS NO.	MEDICAL RECORD NUMBER	HDS NO.	MEDICAL RECORD NUMBER	HDS NO.	MEDICAL RECORD NUMBER

CHANGES: *Check all items below and indicate any changes from previous reporting period.*

YES NO

☐ ☐ NEW ADMINISTRATOR
FULL NAME: _____
TITLE: _____

☐ ☐ NEW MEDICAL RECORD LIBRARIAN:
FULL NAME: _____
TITLE: _____

☐ ☐ NEW CONTROLLER
FULL NAME: _____
TITLE: _____

☐ ☐ CHANGE IN OWNERSHIP OR CONTROL: FROM _____ TO _____

REMARKS *(including a list of supplies and/or forms, if needed):*

_____
Signature and title of transmitter

— ooo —

APPENDIX III

CONTROLLED SELECTION OF THE MASTER SAMPLE AND BLOCKS 3-10

III-1. The method of controlled selection developed to select the sample of 600 hospitals forming the eight HDS blocks 3-10 uses three similar but distinct steps. Each step is given in detail and the data appearing are designed to facilitate understanding.

III-2. The first step in the selection process was to allocate 75 units to the 24 noncertainty primary strata. This allocation was proportional to the number of beds in the strata. Next the number of sample hospitals, or sample take, was set at eight times this figure. This was done to insure that eight samples of 75 hospitals each would be formed and that each sample would have the identical distribution by primary strata.

For example, consider the population stratum of hospitals in the North Central Region having 50-99 beds for inpatient use after removing the block 2 selections.

The proportional allocation called for 25 sample hospitals in this primary stratum. This was rounded to 24, the nearest multiple of eight, and then distributed to the substrata proportionally to the universe data in table IV. The result is shown in table V. (See also the section "Stratification," p. 7.) Thus the figures in table V show what will be called the "expected" number of sample hospitals in each substratum, including the cross-classification cells of the two secondary modes of classification. The figure 15.42 is, for example, the expected number of sample hospitals representing the States in group 1. Since in ordinary usage an expected value, $E(x)$, can be written as

$$E(x) = \sum_i x_i \; Prob\,(x_i),$$

Table IV. Number of hospitals in primary stratum of the universe having 50-99 beds each and located in the North Central Region

Ownership	State cluster		
	Total	1	2
	Number of hospitals		
Total	439	282	157
A	47	29	18
B	90	45	45
C	139	82	57
D	163	126	37

Table V. Expected number of hospitals in blocks 3-10 from primary stratum having 50-99 beds each and located in the North Central Region

Ownership	State cluster		
	Total	1	2
	Number of hospitals		
Total	24.00	15.42	8.58
A	2.57	1.59	0.98
B	4.92	2.46	2.46
C	7.60	4.48	3.12
D	8.91	6.89	2.02

it is instructive to observe that the value of 15.42 can be written as

$$15.42 = 15 \cdot (1 - 0.42) + (15 + 1) \cdot (0.42)$$
$$= 15 \cdot (0.58) + 16 \cdot (0.42) .$$

Thus if the chances are 58/100 of including 15 State group 1 hospitals and 42/100 of including 16 State group 1 hospitals, the expected number of State group 1 hospitals is 15.42. Controlled selection is a way of forming patterns of 15 and 16 State group 1 hospitals and assigning the probabilities 0.58 and 0.42, respectively.

Table V is rewritten into tables VI and VII, where the values in VI are the integer values from table V and those in table VII are the remaining values. Table VI defines a certainty pattern of sample hospitals. Of the 24 sample hospitals in this example, 20 will come from the distribution shown in this table. Thus the sample will have at least six hospitals representing the D-1 substratum.

The remaining four sample hospitals are distributed in table VII. It is the task of controlled selection to form patterns of four hospitals each, not more than one hospital being from the same cross-classification cell, and to assign to these pattern probabilities of selection. The values in the cells of this table are the probabilities of selecting a hospital from that cell. The marginal values, on the other hand, are the constraints on forming the patterns. For example the value in cell D-1 means that there is a probability of 0.89 for a hospital from this substratum being in the sample of four while the value of 0.91 in the D-total cell means that there is a probability of 0.91 for a hospital from the D ownership group

Table VI. Certainty selections

Ownership	State cluster		
	Total	1	2
	Number of hospitals		
Total----------	20	13	7
A----------	1	1	0
B----------	4	2	2
C----------	7	4	3
D----------	8	6	2

Table VII. Noncertainty selection—the control matrix

Ownership	State cluster		
	Total	1	2
	Number of hospitals		
Total----------	4.00	2.42	1.58
A----------	1.57	0.59	0.98
B----------	0.92	0.46	0.46
C----------	0.60	0.48	0.12
D----------	0.91	0.89	0.02

being in the sample of four and there is a probability of 0.09 for no hospital from this ownership being in the sample of four.

A mechanism for forming patterns and assigning the probabilities is as follows:

(1) form n groups of cells, where n is the total number of hospitals shown in the noncertainty selection matrix, in such a way that the sum of the "weights" in each group is one. The weight associated in the group with any cell or "selection unit" can be any part or all of the "expected value" for that cell in table VII. In our illustration $n = 4$, and the third group (see table VIII), composed of parts of cells A-2, C-1, and C-2, has been assigned probabilities, the sum of which is: 0.40 + 0.48 + 0.12 = 1.00;

(2) form a pattern by selecting from each group one cell; in the example, A-2, B-2, C-1, and A-1 form the first pattern;

(3) assign the pattern a probability; for example the first pattern has a probability of 0.09; this can be any value up to the smallest remaining weight for any cell in the pattern;

(4) reduce the weights of all selected cells by the value of the chosen probability in step 3; and

(5) repeat steps 2, 3, and 4 until all weights are reduced to zero.

Now if $a_{ijk} = 1$, when cell ij is in the k^{th} pattern,
$= 0$, otherwise,

and if P_k = the probability assigned to the k^{th} pattern, where $k = 1, \ldots m$, then at the conclusion of the above five steps:

$$\sum_{k=1}^{m} P_k = 1 ,$$

$$\sum_{k=1}^{m} a_{ijk} P_k = E(ij) , \text{ where } E(\Theta)$$

is the expected value in the Θ^{th} cell in table VII.

$$\sum_{k=1}^{m} \sum_{j} a_{i\,jk} P_k = E(i.)$$

$$\sum_{k=1}^{m} \sum_{i} a_{i\,jk} P_k = E(.j)$$

$$\sum_{k=1}^{m} \sum_{i} \sum_{j} a_{i\,jk} P_k = E(..) = n$$ — i.e. the grand total number of sample hospitals for the primary stratum.

The formation of groups and selection of patterns are displayed for the example in table VIII. The figures in the column for pattern 1 are the "weights" referred to above. It can be noted that the sum of weights in this column for any selection unit, e.g., for A-1 the sum is 0.59, is the expected value for the A-1 cell in table VII. The underscored numbers in the columns indicate the membership of selection units in the pattern of the column. After the pattern 1 column, the weights are what remains of the initial weight after the subtraction of the cumulative "probability" of including the particular selection unit in previous patterns.

The strategy for forming both groups and patterns will vary from one practitioner to another. However, if one proceeds according to the guidelines, all or most of the patterns will assure "desirable" samples, and indeed in all patterns the universe cells will be represented proportionately to within one hospital. Some designers try, subjectively, to go even further, and arrange as nearly as feasible to assign to the "most desirable" groups the larger probabilities of selection so that the odds in favor of drawing a "good" sample are even greater, e.g., study of table VIII reveals that patterns 3,4, and 7 all have especially good representation and collectively have a probability of 0.78 that one will be selected.

The general guide in forming groups is to make them internally homogeneous and externally as unlike one another as feasible while still retaining the written guidelines.

Table VIII. Group and pattern formation

Group	Substratum (selection unit)	Pattern						
		1	2	3	4	5	6	7
1	A-1	0.50	0.50	0.42	0.42	0.42	0.40	0.37
	A-2	0.50	0.41	0.41	0.10	-	-	-
2	A-2	0.08	0.08	-	-	-	-	-
	B-1	0.46	0.46	0.46	0.15	0.05	0.03	-
	B-2	0.46	0.37	0.37	0.37	0.37	0.37	0.37
3	A-2	0.40	0.40	0.40	0.40	0.40	0.40	0.37
	C-1	0.48	0.39	0.31	-	-	-	-
	C-2	0.12	0.12	0.12	0.12	0.02	-	-
4	A-1	0.09	-	-	-	-	-	-
	D-1	0.89	0.89	0.81	0.50	0.40	0.40	0.37
	D-2	0.02	0.02	0.02	0.02	0.02	-	-
Probability		0.09	0.08	0.31	0.10	0.02	0.03	0.37

NOTE: The underscored numbers identify membership in the patterns. See text for explanation of the numbers themselves.

It should be emphasized again that there is not a unique way of forming the groups and the patterns at this stage. Indeed a slightly different formation of groups led in an alternative trial (not shown here) to two patterns similar but a little different from patterns numbered 3 and 7 and having a total probability of 0.81. The key observation is that there usually are at least several "very desirable" patterns which may be recognized.

III-3. The selection could be completed in the following manner. The probabilities in the last row of table VIII (multiplied by 100) are cumulated into a table IX. A random number between one and 100 is drawn. Say it happens to be 74. This selects pattern 7 and requires a random selection of hospitals from each of cells A-1, A-2, B-2, and D-1. The total allocation to the primary stratum is through the 20 certainty designations plus these four. For example, the allocation to cell D-1 would be a total of seven hospitals, which would be selected in a systematic random manner from the 126 in the population in that cell. The overall probability of selection of one of these hospitals is 24/439=0.05467 (not $\frac{7}{126} = 0.05556$). The entire procedure would be repeated for the other primary strata, and the master sample selected accordingly. Under this procedure, it would be possible, through rounding approximations accumulating in the 24 strata, to have in the overall sample some slight over- or under-allocation of hospitals to an ownership class. This contingency is not significantly troublesome, but in the HDS, additional precautions were taken to avoid it, by introducing a refinement which employed further application of controlled selection methods.

Table IX. Cumulated pattern probabilities (decimal point removed)

Pattern	Cumulated probability
1	9
2	17
3	48
4	58
5	60
6	63
7	100

III-4. To avoid unnecessary reading of detailed tables, the refinement is described in outline form only. (The rounding problem relates only to noncertainty cases, so in the following text only the noncertainty cases are referred to, with the understanding that the certainty selections are simply added together over all 24 primary strata.)

A second application of controlled selection procedure is carried through in the same manner as the one described previously. The objective of this new round is to control ownership over *all* size-classes within each region. New tables V, VI, and VII are formed with the "expected numbers" applying to the region rather than the primary stratum.

A new table VIII is formed in which the "groups" are the six size classes; the "patterns" of the previous step become the "selection units" of the new step; the probabilities of the patterns in the previous steps become the "weights" which are associated with their respective patterns and which are distributable among new patterns, as in the old table VIII; the new patterns are combinations of the new selection units, to each of which is assigned a new pattern probability in the same manner as in III-3.

III-5. The final phase of the refinement for the master sample is to repeat the procedure of III-4, for the Nation as a whole, with the new groups being the four regions, the new selection units being the patterns of III-4; and the probabilities of III-4 patterns the new weights. The complete master sample of cells is selected then with the random drawing of a single III-5 pattern.

III-6. The consequence of this set of steps is that the sample has been allocated first to each of 24 size-class-region primary strata and then controlled to ownership over the Nation, within region, and within size class; and in addition, there is control to State cluster within each region-size class and to ownership-State-cluster within region-size class.

III-7. Block formation.—Controlled selection was also used to put the selected hospitals into blocks, the process being quite similar to that described above. Assume, for illustration, a primary stratum having the following 24 selected hospitals:

Ownership class	Region total	State group	
		1	2
All classes------	24	15	9
A----------	3	2	1
B----------	5	2	3
C----------	8	5	3
D----------	8	6	2

This is the certainty pattern (table VI) and pattern 1 in table VIII. Each block was to have one-eighth of this distribution:

Ownership class	Region total	State group	
		1	2
All classes------	3	1-7/8	1-1/8
A----------	3/8	2/8	1/8
B----------	5/8	2/8	3/8
C----------	1	5/8	3/8
D----------	1	6/8	2/8

The interpretation and use of this table is exactly the same as that of table VI. Each block could have, for example, only one C-type hospital—not none and not two or more. Similarly one block must have two hospitals for State cluster 2, while the other seven blocks must have one each.

After these patterns were formed for each primary stratum, they were joined together within the regions much as in III-4 above and then over the regions as in III-5 to yield a final product of eight patterns of 75 hospitals each. A random ordering was carried out to assign the labels block 3, block 4...block 10.

APPENDIX IV

ESTIMATING EQUATIONS AND SAMPLING VARIANCES

The Estimator, $\hat{X}$

The HDS estimator for a patient statistic, $\hat{x}$, is secured in the following manner. The value x_{ghijkl} is abstracted from the

- l^{th} sample record in the
- k^{th} sample hospital from the
- j^{th} block of sample hospitals,
- i^{th} hospital-size class,
- h^{th} geographic region, and
- g^{th} month,

and carried through seven stages of adjustment before it makes its final contribution to the value of $\hat{x}$. Three of these adjustments bring sample observations to hospital levels, x'_{ghijk}, while three others bring the hospital levels up to primary stratum levels, x''_{ghi}. Then a ratio estimation control takes x''_{ghi} to $\hat{x}_{ghi}$, the final stratum estimate.

The estimating steps account for

(1) second stage inflation (i.e., within-hospital sampling)
(2) patient nonresponse
(3) second stage ratio estimator control
(4) first stage inflation (sampling among hospitals)
(5) hospital nonresponse
(6) block adjustment (weighting for number of reporting panels)
(7) first stage ratio estimator control

Adjustments to hospital levels, x'_{ghijk}.—The value x_{ghijkl} is obtained for a discharge selected through a two stage probability sample. Therefore the sample observations must be inflated by the reciprocals of the sampling probabilities. The first inflation is

$$ {}_1x'_{ghijkl} = \left(\frac{1}{{}_2P_{hijkl}}\right) x_{ghijkl} \qquad (1) $$

where ${}_2P_{hijkl}$ is the probability of selecting the l^{th} record from the k^{th} hospital, given the selection of the k^{th} hospital.

When patient records which should be abstracted are missing from the file of records, data are imputed for them from information recorded for the abstracted discharges in that hospital and survey month.

$$ {}_2x'_{ghijkl} = \left(\frac{n''_{ghijk}}{n'_{ghijk}}\right) {}_1x'_{ghijkl} \qquad (2) $$

where n''_{ghijk} is the number of patient records that should be abstracted,

and n'_{ghijk} is the number actually abstracted for the HDS survey.

To minimize the effects of systematic sampling of patient records in the hospital (the universe is not always an integral number of times the sample size), an adjustment is made to the known and separately reported number of discharges from the hospital. The effect of this adjustment is to substitute the operational ratio of sample size to universe size within the hospital for the formal probability of selection.

$$ {}_3x'_{ghijkl} = \left(\frac{N_{ghijk}}{n_{ghijk}} \cdot {}_2P_{hijkl}\right) {}_2x'_{ghijkl} \qquad (3) $$

and $$ x'_{ghijk} = \sum_l {}_3x'_{ghijkl} \qquad (3a) $$

where N_{ghijk} is the number of patient discharges reported by the hospital as a control number,

and n_{ghijk} is the systematic sample number of patient discharges

Adjustments to stratum levels, x'_{ghijk}.—The first adjustment to bring the hospital level estimates in equation 3a up to stratum levels in the inflation of x'_{ghijk} by the reciprocal of the probability of choosing the k^{th} hospital.

$$ {}_1x''_{ghijk} = \left(\frac{1}{{}_1P_{hijk}}\right) x'_{ghijk} \qquad (4) $$

where ${}_1P_{hijk}$ is the probability of selecting the k^{th} hospital in block j and stratum hi.

Adjustments are made when sample hospitals do not respond for a survey month. Data for the nonresponding hospitals are imputed from the responding ones. The imputation formula is

$$ {}_2x''_{ghijk} = \left[\frac{\sum_j^{\nu} \sum_k m_{hij} B_{hijk}}{\sum_j^{\nu} \sum_k m'_{ghij} B_{hijk}}\right] {}_1x''_{ghijk} \qquad (5) $$

where ν is the number of noncertainty sample blocks in the survey in the g^{th} month,

m_{hij} is the number of sample hospitals in the j^{th} block, h^{th} stratum,

m'_{ghij} is the number of sample hospitals in the j^{th} block, hi^{th} stratum, which respond during month g,

and B_{hijk} is the MFI number of beds for inpatient use in the k^{th} hospital, j^{th} block and hi^{th} stratum. (MFI number indicates the frame from which the samples were drawn, and not necessarily the current number of beds in the hospital.)

Each block of noncertainty hospitals is a probability sample and hence yields estimates for this noncertainty universe. Therefore an average of all the blocks in a survey month is taken to produce an estimate with smaller variance. This is accomplished operationally by dividing each inflated original observation by ν—the number of noncertainty blocks in the estimate.

$$3x''_{ghijk} = \left(\frac{1}{\nu}\right) 2x''_{ghijk} \tag{6}$$

and then adding these quantities:

$$x''_{ghi} = \sum_j \sum_k 3x''_{ghijk} \tag{6a}$$

The final adjustment, $\hat{x}_{ghi}$. This is the first stage ratio estimator control. This adjustment to the known total of MFI beds for inpatient use is taken to reduce further the variance of the estimate:

$$\hat{x}_{ghi} = \left(\frac{B_{hi}}{B''_{hi}}\right) x''_{ghi} \tag{7}$$

where B_{hi} is the MFI number of beds for patient use in stratum hi,

and B''_{hi} is the sample estimate of the number of beds for patient use and it is calculated from

$$B''_{hi} = \frac{1}{\nu} \sum_j^{\nu} \sum_k m_{hij} \left(\frac{1}{1^P_{hijk}}\right) B_{hijk} \tag{8}$$

No adjustments are necessary for nonresponse or missing data since the frame is by definition complete. (A sample of births of new hospitals coming into being since the effective date of the frame is to be surveyed in an entirely separate operation.)

The complete estimator.—These six steps combine then to produce an estimate $\hat{x}_{ghi}$ for the h^{th} geographic region, i^{th} hospital-size class for the g^{th} month:

$$\hat{x}_{ghi} = \left[\left(\frac{\sum_j^{\nu} \sum_k m_{hij} B_{hijk}}{\sum_j^{\nu} \sum_k m'_{ghij} B_{hijk}}\right) \sum_j^{\nu} \sum_k m'_{ghij} \left(\frac{N_{ghijk}}{n_{ghijk}} \cdot \frac{n''_{ghijk}}{n'_{ghijk}}\right) \left(\frac{1}{1^P_{hijk}}\right) \sum_l^{n_{ghijk}} x_{ghijkl}\right] \left[\frac{B_{hi}}{\sum_j^{\nu} \sum_k m_{hij} \left(\frac{1}{1^P_{hijk}}\right) B_{hijk}}\right] \tag{9}$$

The reader will note that the formulas given are those which yield noncertainty strata estimates. In forming estimates based on data from the certainty hospitals, some of the equations and parts of the equations given do not apply. In particular, equation (9) for certainty strata estimates is:

$$\hat{x}_{ghi} = \left(\frac{\sum_k^{M_{ghil}} B_{hilk}}{\sum_k^{M'_{ghil}} B_{hilk}}\right) M'_{ghil} \left(\frac{N_{ghilk}}{n_{ghilk}} \cdot \frac{n''_{ghilk}}{n'_{ghilk}}\right) \sum_l^{n'_{ghilk}} x_{ghilkl} \tag{10}$$

where M_{ghil} is the universe number of certainty hospitals in stratum hi

and M'_{ghil} is the number of responding certainty hospitals. (Note that M_{ghil} is a fixed constant over all g, being the frame value.) Summing over all strata, including both certainty and noncertainty, and months yields the final estimator $\hat{\bar{x}}$

$$\hat{\bar{x}} = \sum_g \sum_{hi} \hat{x}_{ghi} \tag{10a}$$

Sampling Variance

The HDS variance estimator is an approximation of a rigorously unbiased estimator, but experimental calculations have shown the biases of the approximation to be trivial. The statistics produced by the HDS pass through two phases: first, monthly estimates based on data collected from all responding hospitals are computed, and, second, these estimates are summed over the months to produce the published statistics. An exact variance estimator would be based on the same sequence of events, i.e., each month's statistic would have a variance calculated and the variance of a yearly estimate would be the sum of these monthly variances plus terms accounting for the covariances existing between the monthly estimates.

An exact variance formula is not difficult to derive but consequent programming and computational problems are sufficiently costly, and the improvement over approximate methods so slight that the more precise method is not justified.[a] Three rules were adopted to facilitate the computations of the variance:

(1) All data appearing in the variance calculations are at the level of a year instead of the month.

(2) All data appearing in the variance calculations are from hospitals which responded for the 12 months in the year.

(3) All hospitals in blocks 3 through j (j = 4, 5... 10) are consolidated into a superblock 3; thus for the variance calculations there are two noncertainty parts: block 2 and superblock 3.

[a] To check the expressions programmed for HDS, an exact expression was developed to obtain variances for several items. These were used to verify the derivations and programming of the HDS estimators.

The certainty strata.—The sampling variance formulas for statistics from the certainty block (c) are:

$$s^2_{\hat{\bar{x}}_c} = \sum_{hi}^{c} s^2_{\hat{\bar{x}}_{hil}} \tag{11}$$

where

$$s^2_{\hat{\bar{x}}_{hil}} = \left(\frac{\sum_k^{M_{hil}} B_{hilk}}{\sum_k^{M'_{hil}} B_{hilk}}\right)^2 \left[\sum_k^{M'_{hil}} \left(N''_{hilk}\right)^2 \left(1-\frac{n'_{hilk}}{N''_{hilk}}\right)\right. \tag{12}$$

$$\left.\left(\frac{1}{n'_{hilk}}\right)\left(\frac{1}{n'_{hilk}}-1\right)\left\{\sum_l x^2_{hilkl} - \frac{\left(\sum_l x_{hilkl}\right)^2}{n'_{hilk}}\right\}\right]$$

and M_{hil} is the number of certainty hospitals,

M'_{hil} is the number of responding certainty hospitals,

$$N''_{hilk} = \left(\sum_g N_{ghilk}\right)\left(\frac{\sum_g n''_{ghilk}}{\sum_g n'_{ghilk}}\right)$$

and $$x_{hilkl} = \sum_g x_{ghilkl}.$$

The noncertainty strata.—The sampling variance formulas for statistics from the noncertainty strata (NC) are

$$s^2_{\hat{\bar{x}}_{NC}} = \sum_{hi}^{NC} s^2_{\hat{\bar{x}}_{hi,NC}} \tag{13}$$

where

$$s^2_{\hat{\bar{x}}_{hi,NC}} = \left(\frac{\sum_j^{\nu}\sum_k^{m_{hij}} B_{hijk}}{\sum_j^{\nu}\sum_k^{m'_{hij}} B_{hijk}}\right)^2 \left(\frac{1}{\nu}\right)\left\{\left(\frac{1}{\nu}\right)s^2_{\hat{\bar{x}}_{hi(2)}} + \left(\frac{\nu-1}{\nu}\right)s^2_{\hat{\bar{x}}_{hi(3)}}\right\} \tag{14}$$

and all other terms have been defined previously. The first term on the right-hand side of equation (14) arises because of equation (5), the second term because of equation (6), while the term in braces yields a composite estimate with weights assigned appropriately for the number of blocks used in making the variance calculations.

The variance from block 2 is

$$S^2_{\hat{\bar{x}}_{hi(2)}} = \left(M_{hi}\right)^2\left(\frac{1}{m'_{hi2}}\right)\left[\left(1-\frac{m'_{hi2}}{M_{hi}}\right)\right.$$

$$\left\{\left(S^2_{1x_{hi(2)}} - \left(\frac{1}{m'_{hi2}}\right)S^2_{2x_{hi(2)}}\right)\right. \tag{15}$$

$$\left.+\left(\frac{X''_{hi}}{B''_{hi}}\right)^2 S^2_{B_{hi(2)}} - 2\left(\frac{X''_{hi}}{B''_{hi}}\right)S_{1x_{hi(2)}B_{hi(2)}}\right\}$$

$$\left.+\frac{1}{m'_{hi2}}S^2_{2x_{hi(2)}}\right]$$

where

$$S^2_{1x_{hi(2)}} = \left(\frac{1}{m'_{hi2}-1}\right)\left[\sum_k^{m'_{hi2}}\left(x'_{hi2k}\right)^2 - \frac{\left(\sum_k x'_{hi2k}\right)^2}{m'_{hi2}}\right] \tag{16}$$

$$S^2_{2x_{hi(2)}} = \sum_k^{m'_{hi2}}\left(N''_{hi2k}\right)^2 \frac{\left(1-\frac{n'_{hi2k}}{n''_{hi2k}}\right)}{n'_{hi2k}}\left(\frac{1}{n'_{hi2k}-1}\right)$$

$$\left[\sum_l^{n'_{hi2k}}\left(x_{hi2kl}\right)^2 - \frac{\left(\sum_l x_{hi2kl}\right)^2}{n'_{hi2k}}\right] \tag{17}$$

$$S^2_{B_{hi(2)}} = \left(\frac{1}{m'_{hi2}-1}\right)\left[\sum_k^{m'_{hi2}}\left(B_{hi2k}\right)^2 - \frac{\left(\sum_k B_{hi2k}\right)^2}{m'_{hi2}}\right] \tag{18}$$

$$S_{1x_{hi(2)}B_{hi(2)}} = \left(\frac{1}{m'_{hi2}-1}\right)\left[\sum_k^{m'_{hi2}}\left(x'_{hi2k}\right)\left(B_{hi2k}\right)\right. \tag{19}$$

$$\left. - \frac{\left(\sum_k x'_{hi2k}\right)\left(\sum_k B_{hi2k}\right)}{m'_{hi2}}\right]$$

and $x''_{hi} = \sum_g x''_{ghi}$

x''_{ghi} is given in equation (6a).

B''_{hi} is given in equation (8).

$x'_{hi2k} = \sum_g x'_{ghi2k}$

x'_{ghi2k} is given in equation (3a).

All other terms have been previously defined. The estimator $s^2_{\hat{\bar{x}}_{hi(3)}}$ can be given approximately and conceptually by the equation

$$s^2_{\hat{\bar{x}}_{hi(3)}} \doteq \left(\frac{1}{\nu-1}\right)\sum_{j=3}^{\nu+1} s^2_{\hat{\bar{x}}_{hi(j)}}$$

where $s^2_{\hat{\bar{x}}_{hi(j)}}$ is precisely the same as $s^2_{\hat{\bar{x}}_{hi(2)}}$ with (j) substituted for (2) in formula (15). (The actual calculations for $s^2_{\hat{\bar{x}}_{hi(3)}}$ consolidated all hospitals in blocks 3, 4... $(\nu+1)$ into the single superblock 3, as described earlier in this section.)

Published Variances

Although the analyst uses the statistics and variances produced by the formulations given above, the publications usually show "typical" values for the variances. There are many reasons for this, the most obvious being that the average publication contains hundreds of statistics; showing a variance for each would not only be costly but also unnecessary. Sta-

tistics of the same type and size from the same survey usually have substantially the same variance. Thus only some average value for statistics of a fixed type and size is needed. An adequate value can be secured from a fitted functional relationship of relvariances ($V^2_{x'_i}$) to size and type of estimate (x'_i), obtained from the initial calculations.

The values of published variances in the Hospital Discharge Survey come from

$$\hat{V}^2_{x'_i} = a + b/x'_i \qquad (20)$$

The values for a and b are determined by a "least squares" technique where the normal equations are obtained from

$$\frac{\partial S}{\partial a} = \frac{\partial S}{\partial b} = 0 \qquad (21)$$

in which

$$S = \Sigma \left[\frac{V^2_{x'_i} - (a + b/x'_i)}{\hat{V}^2_{x'_i}} \right]^2 , \text{ and} \qquad (22)$$

$$V^2_{x'_i} = s^2_{x'_i} / (x'_i)^2 , \qquad (23)$$

where $s^2_{x_i}$ is the initially calculated sampling variance, and x_i is the value of the statistic. The formulation in (22) requires an iterative procedure to estimate a and b. In the first approximation, $\hat{V}^2_{x'_i}$ is set to equal $V^2_{x'_i}$. In subsequent calculations $\hat{V}^2_{x'_i}$ is obtained from equation (20). Calculations are continued until successive a's and successive b's are within 2 percent of each other, respectively.

———o o o———

☆U.S. GOVERNMENT PRINTING OFFICE: 1977— 241-180:12

VITAL AND HEALTH STATISTICS PUBLICATIONS SERIES

Formerly Public Health Service Publication No. 1000

Series 1. Programs and Collection Procedures.—Reports which describe the general programs of the National Center for Health Statistics and its offices and divisions, data collection methods used, definitions, and other material necessary for understanding the data.

Series 2. Data Evaluation and Methods Research.—Studies of new statistical methodology including experimental tests of new survey methods, studies of vital statistics collection methods, new analytical techniques, objective evaluations of reliability of collected data, contributions to statistical theory.

Series 3. Analytical Studies.—Reports presenting analytical or interpretive studies based on vital and health statistics, carrying the analysis further than the expository types of reports in the other series.

Series 4. Documents and Committee Reports.—Final reports of major committees concerned with vital and health statistics, and documents such as recommended model vital registration laws and revised birth and death certificates.

Series 10. Data from the Health Interview Survey.—Statistics on illness; accidental injuries; disability; use of hospital, medical, dental, and other services; and other health-related topics, based on data collected in a continuing national household interview survey.

Series 11. Data from the Health Examination Survey.—Data from direct examination, testing, and measurement of national samples of the civilian, noninstitutionalized population provide the basis for two types of reports: (1) estimates of the medically defined prevalence of specific diseases in the United States and the distributions of the population with respect to physical, physiological, and psychological characteristics; and (2) analysis of relationships among the various measurements without reference to an explicit finite universe of persons.

Series 12. Data from the Institutionalized Population Surveys.—Discontinued effective 1975. Future reports from these surveys will be in Series 13.

Series 13. Data on Health Resources Utilization.—Statistics on the utilization of health manpower and facilities providing long-term care, ambulatory care, hospital care, and family planning services.

Series 14. Data on Health Resources: Manpower and Facilities.—Statistics on the numbers, geographic distribution, and characteristics of health resources including physicians, dentists, nurses, other health occupations, hospitals, nursing homes, and outpatient facilities.

Series 20. Data on Mortality.—Various statistics on mortality other than as included in regular annual or monthly reports. Special analyses by cause of death, age, and other demographic variables; geographic and time series analyses; and statistics on characteristics of deaths not available from the vital records, based on sample surveys of those records.

Series 21. Data on Natality, Marriage, and Divorce.—Various statistics on natality, marriage, and divorce other than as included in regular annual or monthly reports. Special analyses by demographic variables; geographic and time series analyses; studies of fertility; and statistics on characteristics of births not available from the vital records, based on sample surveys of those records.

Series 22. Data from the National Mortality and Natality Surveys.—Discontinued effective 1975. Future reports from these sample surveys based on vital records will be included in Series 20 and 21, respectively.

Series 23. Data from the National Survey of Family Growth.—Statistics on fertility, family formation and dissolution, family planning, and related maternal and infant health topics derived from a biennial survey of a nationwide probability sample of ever-married women 15-44 years of age.

For a list of titles of reports published in these series, write to:

Scientific and Technical Information Branch
National Center for Health Statistics
Public Health Service, HRA
Rockville, Md. 20857

NCHS

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE
Public Health Service
Health Resources Administration
5600 Fishers Lane
Rockville, Md. 20857

For information about the *Vital and Health Statistics* Series call 301-443-NCHS.