TECHNICAL SUPPLEMENT 5:
DESIGN OF THE HCUP NATIONWIDE INPATIENT SAMPLE, RELEASE 1

INTRODUCTION

The Nationwide Inpatient Sample (NIS) of the Healthcare Cost and Utilization Project (HCUP) was established to provide analyses of hospital utilization across the United States. Release 1 covers calendar years 1988 through 1992. The target universe includes all acute-care discharges from all community hospitals in the United States; the NIS comprises all discharges from a sample of hospitals in this target universe.

For each calendar year (1988 through 1992), this first release of the NIS contains 5.2 to 6.2 million discharges from a sample of 758 to 875 hospitals in 11 states (8 states for 1988). Future releases will add data to the NIS file from more states and more years. Thus, the NIS supports both cross-sectional and longitudinal analyses.

Potential research issues focus on both discharge- and hospital-level outcomes. Discharge outcomes of interest include trends in inpatient treatments with respect to:

- frequency,
- costs,
- lengths of stay,
- effectiveness,
- appropriateness, and
- access to hospital care.

Hospital outcomes of interest include:

- mortality rates,
- complication rates,
- patterns of care,
- diffusion of technology, and
- trends toward specialization.

These and other outcomes are of interest for the nation as a whole and for policy-relevant inpatient subgroups defined by geographic regions, patient demographics, hospital characteristics, physician characteristics, and pay sources.

This report provides a detailed description of the NIS sample design, as well as a summary of the resultant hospital sample. Sample weights were developed to obtain national estimates of hospital and inpatient parameters. These weights and other special-use weights are described in detail.

THE NIS HOSPITAL UNIVERSE

For each calendar year, the hospital universe is defined by all hospitals that were open during any part of that calendar year and were designated as community hospitals in the American Hospital

Association (AHA) Annual Survey of Hospitals. For purposes of the NIS, the definition of a community hospital is that used by the AHA: "all nonfederal short-term general and other specialty hospitals, excluding hospital units of institutions." Consequently, Veterans Hospitals and other federal hospitals are excluded.

Table 5 shows the number of universe hospitals for each calendar year based on HCUP's calendar-year conforming version of the AHA survey-year files. Survey responses were put on a calendar-year basis for 1988-1991 by merging data from adjacent survey years. However, 1992 AHA survey data remain in the original reporting-year form because HCUP received the 1993 AHA files too late to convert fiscal-year responses to calendar-year files for 1992.

Table a. Hospital Universe

Calendar Year	Number of Hospitals
1988	5,607
1989	5,548
1990	5,468
1991	5,412
1992	5,334

Hospital Merges, Splits, and Closures

All hospital entities that were designated community hospitals in the AHA hospital file were included in the hospital universe. Therefore, if two or more community hospitals merged to create a new community hospital, the original hospitals and the newly-formed hospital were all considered separate hospital entities in the universe for the year of the merge. Likewise, if a community hospital split, the original hospital and all newly created community hospitals were separate entities in the universe for the year of the split. Finally, community hospitals that closed during a year were included as long as they were in operation during some part of the calendar year.

Stratification Variables

To help ensure representativeness, sampling strata were defined based on five hospital characteristics contained in the AHA hospital files. The stratification variables were as follows:

1) *Geographic Region – Northeast, North Central, West, and South.* This is an important stratifier because practice patterns have been shown to vary substantially by region. For example, lengths of stay tend to be longer in East Coast hospitals than in West Coast hospitals.

2) *Ownership – public, private not-for-profit, and private investor-owned.* These types of hospitals tend to have different missions and different responses to government regulations and policies.

3) *Location – urban or rural.* Government payment policies often differ according to this designation. Also, rural hospitals are generally smaller and offer fewer services than urban hospitals.

4) *Teaching Status – teaching or nonteaching.* The missions of teaching hospitals differ from nonteaching hospitals. In addition, financial considerations differ between these two hospital groups. Currently, the Medicare DRG payments are uniformly higher to teaching hospitals than to nonteaching hospitals. A hospital is considered to be a teaching hospital if it has an AMA-approved residency program or is a member of the Council of Teaching Hospitals (COTH).

5) *Bedsize – small, medium, and large.* Bedsize categories are based on hospital beds, and are specific to the hospital's location and teaching status, as shown in Table 6.

Table b. Bedsize Categories

Location and Teaching Status	Hospital Bedsize		
	Small	**Medium**	**Large**
Rural	1-49	50-99	100+
Urban, nonteaching	1-99	100-199	200+
Urban, teaching	1-299	300-499	500+

Rural hospitals were not split according to teaching status, because rural teaching hospitals were rare. For example, in 1988 there were only 20 rural teaching hospitals. The bedsize categories were defined within location and teaching status because they would otherwise have been redundant. Rural hospitals tend to be small; urban nonteaching hospitals tend to be medium-sized; and urban teaching hospitals tend to be large. Yet it was important to recognize gradations of size within these types of hospitals.

For example, in serving rural discharges, the role of "large" rural hospitals (particularly rural referral centers) often differs from the role of "small" rural hospitals. The cut-off points for the bedsize categories are consistent with those used in *Hospital Statistics,* published annually by the AHA.

To further ensure geographic representativeness, implicit stratification variables included state and three-digit zip code (the first three digits of the hospital's five-digit zip code). The hospitals were sorted according to these variables prior to systematic sampling.

HOSPITAL SAMPLING FRAME

For each calendar year, the *universe* of hospitals was established as all community hospitals located in the U.S. However, it was not feasible to obtain and process all-payer discharge data from a random sample of the entire universe of hospitals for at least two reasons. First, all-payer discharge data were not available from all hospitals for research purposes. Second, based on the experience of prior hospital discharge data collections, it would have been too costly to obtain data from individual hospitals, and it would have been too burdensome to process each hospital's unique data structure.

Therefore, the NIS *sampling frame* was constructed from the subset of universe hospitals that released their discharge data for research use. Two sources for all-payer discharge data were state agencies and private data organizations, primarily state hospital associations. Currently, the Agency for Health Care Policy and Research (AHCPR) has agreements with 22 data sources that maintain statewide, all-payer discharge data files to include their data in the HCUP database. However, only 8 states in 1988 and 11 states in 1989-1992 could be included in this first release, as shown in Table 7. Future releases of the NIS will include more states.

Table c. States in the Frame for the NIS, Release 1

Calendar Years	States in the Frame
1988	California, Colorado, Florida, Iowa, Illinois, Massachusetts, New Jersey, and Washington
1989-1992	Add Arizona, Pennsylvania, and Wisconsin

The Illinois Health Care Cost Containment Council stipulated that no more than 40 percent of Illinois data could be included in the database for any calendar quarter. As a result, approximately 40 percent of the Illinois community hospital universe was randomly selected for the frame each year using the same methodology used to select the NIS hospital sample. That is, Illinois hospitals were stratified on the stratification variables described above, and a systematic random sample of hospitals was drawn for the frame.

Therefore, the list of the entire frame of hospitals was composed of the 40 percent sample of community hospitals for Illinois and all AHA community hospitals in each of the other frame states *that could be matched to the discharge data provided to HCUP*. If an AHA community hospital could not be matched to the discharge data provided by the data source, it was eliminated from the sampling frame (but not from the universe). Unfortunately, only Florida community hospitals are included in the frame for the South region. It is expected that additional southern states will be included in future releases.

The number of frame hospitals for each year is shown in Table 8.

Table d. Hospital Frame

Calendar Year	Number of Hospitals
1988	1,247
1989	1,658
1990	1,620
1991	1,604
1992	1,591

HOSPITAL SAMPLE DESIGN

Design Requirements

The NIS is a stratified probability sample of hospitals in the frame, with sampling probabilities calculated to select 20 percent of the universe contained in each stratum. The overall objective was to select a sample of hospitals "generalizable" to the target universe, including hospitals outside the frame, which have a zero probability of selection. Moreover, this sample was to be geographically dispersed, yet drawn from the subset of states with inpatient discharge data that agreed to provide such data to the project.

It should be possible, for example, to estimate DRG-specific average lengths of stay over all U.S. hospitals using weighted average lengths of stay, based on averages or regression estimates from the NIS. Ideally, relationships among outcomes and their correlates estimated from the NIS should generally hold across all U.S. hospitals. However, since only 11 states contributed data to this first release, some estimates may be biased. When possible, estimates based on the NIS should be checked against national benchmarks, such as Medicare data or data from the National Hospital Discharge Survey.

The target sample size was 20 percent of the total number of community hospitals in the U.S. for each year in the study period, 1988-1992. This sample size was determined by AHCPR based on their experience with similar research databases.

Alternative stratified sampling allocation schemes were considered. However, allocation proportional to the number of hospitals seemed best for several reasons:

- Fewer than 10 percent of government-planned database applications will produce nationwide estimates. The major government applications will investigate relationships among variables. For example, government researchers will do a substantial amount of regression modeling with these data.

- The HCUP-2 sample[1] used the same stratification and allocation scheme, and it has served AHCPR analysts well. Moreover, the large number of sample hospitals and discharges seemingly reduced the need for variance-reducing allocation schemes.

- AHCPR researchers wanted a simple, easily understood sampling methodology. It was an appealing idea that the NIS sample could be a "miniaturization" of the universe of hospitals (with the obvious geographical limitations imposed by data availability).

- AHCPR statisticians considered other optimal allocation schemes, including sampling hospitals with probabilities proportional to size (number of discharges), and they concluded that sampling with probability proportional to the number of hospitals was preferable. Even though it was recognized that the approach chosen would not be as efficient, the extremely large sample sizes would still yield good estimates. Furthermore, because the data would also be used for purposes other than producing national estimates, it was critical that all hospital types (including small hospitals) be adequately represented.

Hospital Sampling Procedure

Once the universe of hospitals was stratified, up to 20 percent of the total number of U.S. hospitals was randomly selected within each stratum. If too few frame hospitals were in the stratum, then all frame hospitals were selected for the NIS. To simplify variance calculations, at least two hospitals were drawn from each stratum. If fewer than two frame hospitals were contained in a stratum, then that stratum was merged with an "adjacent" stratum containing hospitals with similar characteristics.

We drew a systematic random sample from each stratum, after sorting hospitals by state within each stratum, then by the three-digit zip code (the first three digits of the hospital's five-digit zip code) within each state, and then by a random number within each three-digit zip code. These sorts ensured further geographic generalizability of hospitals within the frame states, and random ordering of hospitals within three-digit zip codes.

Generally, three-digit zip codes that are near in value are geographically near within a state. Furthermore, the U.S. Postal Service locates regional mail distribution centers at the three-digit level. Thus, the boundaries tend to be a compromise between geographic size and population size.

1988 NIS Hospital Sampling Procedure

The 1988 hospital sample was selected according to the following steps:

1. The universe of hospitals was stratified on region, ownership, location, teaching status, and bedsize category.

2. The number of universe and frame hospitals were counted in each stratum.

3. If any stratum had fewer than two hospitals in the *frame,* it was combined with an adjacent stratum to ensure at least two frame hospitals. In all cases where this was required, it was necessary only to collapse the ownership categories. For all cases in which strata were

collapsed, private not-for-profit was combined with public, or private investor-owned was combined with public.

4. Within each stratum, the frame hospitals were sorted by state, by three-digit zip code within each state, and by a random number within each three-digit zip code.

5. For each stratum:

   a. The stratum-specific sampling rate (probability) was calculated:

   $$P = \min(1, N/F)$$

   where N = max (2, .20*U), and U = total number of universe hospitals in the stratum. Therefore, N was the number of hospitals "needed" in each stratum (at least two hospitals and at most 20 percent of the universe).

   F = total number of frame hospitals in the stratum.

   If $F \leq N$ (the number of frame hospitals was less than the number needed), then P = 1 and all hospitals were selected in the stratum.

   b. The skip interval for the systematic sample was calculated:

   $$S = 1 \div P$$

   Every Sth hospital on the list was sampled. For example, if P = .5, then every second hospital was sampled. However, S need *not* have been an integer for this procedure.

   c. A random starting point was calculated for the sample:

   $$R = \text{random number in the interval } (0,S)$$

   d. Every Sth hospital was drawn for the sample from the list of sorted frame hospitals. Let INT(x) be the integer part of x. The first hospital drawn was number INT(1 + R). The second hospital drawn was number INT(1 + R + S). The third hospital drawn was number INT(1 + R + 2S), and so on.

Frame hospitals within a given stratum all had an equal chance of entering the sample. Also, on average, the correct number of hospitals (20 percent of the universe) was drawn for each stratum that had a sufficient number of hospitals.

A total of 758 hospitals was drawn for the 1988 NIS. This number fell short of the overall target of 1,121 hospitals (20 percent of the universe), because several strata contained too few frame hospitals to meet the 20 percent target. More details on the final sample are described later in this report.

1989-1992 NIS Hospital Sampling Procedure

Once the 1988 hospital sample was drawn, it was necessary to draw the 1989 sample by a procedure that "reselected" most of the 1988 hospitals, while allowing hospitals new to the frame an opportunity to enter the 1989 NIS. In particular, hospitals in three states (AZ, PA, and WI) that were not in the 1988 frame entered the 1989 frame.

Even in other frame states, hospitals that opened in 1989 needed a chance to enter the sample. Also, hospitals that changed strata between 1988 and 1989 were considered new to the 1989 frame.

Likewise, once the 1989 hospital sample was drawn, it was necessary to draw the 1990 sample in a way that retained most of the 1989 sample hospitals, while allowing new frame hospitals a chance of selection in 1990.

Consequently, a recursive procedure was developed to update the sample from year to year in a way that properly accounted for changes in stratum size, composition, and sampling rate. The goal of this procedure was to maximize the year-to-year overlap among sample hospitals, yet keep the sampling rate constant for all hospitals *within a stratum*.

The 1988 sampling procedure determined the probability of selection for available frame hospitals within each stratum (probability P). It also gave a procedure for selecting a systematic sample of frame elements with this probability. This procedure was taken as a starting point.

The following procedure provides rules for creating a "year 2" sample, given that a "year 1" sample had already been drawn. For example, year 1 could be 1988 and year 2 could be 1989, or year 1 could be 1989 and year 2 could be 1990. All notation is assumed to refer to sizes and probabilities within a particular stratum.

Probabilities P_1 and P_2 were calculated for sampling hospitals from the frame within the stratum for year 1 and year 2, respectively, based on the frame and universe for year 1 and year 2, respectively. These probabilities were set by the same algorithm used to calculate P for the 1988 hospital sample (step 5a for selecting the 1988 sample).

Now consider the three possibilities associated with changes between years 1 and 2 in the stratum-specific hospital sampling probabilities:

1. $P_2 = P_1$: The target probability was unchanged.

2. $P_2 < P_1$: The target probability decreased.

3. $P_2 > P_1$: The target probability increased.

Below is the procedure used for each of these three cases with one exception: if the stratum-specific probability of selection P_2 was equal to 1, then all frame hospitals were selected for the year 2 sample, regardless of the value of P_1.

Stratum-Specific Sampling Rates the Same ($P_2 = P_1$). If the probability P_2 was the same as P_1, all hospitals in the year 1 sample that remained in the year 2 frame were retained for the year 2 sample. Any new frame hospitals (those in the year 2 frame but not in the year 1 frame) were selected at the rate P_2, using the systematic sampling method described for the 1988 sample selection.

Stratum-Specific Sampling Rate Decreased ($P_2 < P_1$). Now consider the case where the probability of selection decreased between years 1 and 2. First, hospitals new to the frame were sampled with probability P_2. Second, hospitals previously selected for the year 1 sample (that remained in the year 2 frame) were selected for the year 2 sample with probability $P_2 \div P_1$.

The justification for this second procedure was straightforward. For the year 1 sample hospitals that stayed in the frame, the year 1 sample was viewed as the first stage of a two-stage sampling process. The first stage was carried out at the sampling rate of P_1. The second stage was carried out at the sampling rate of $P_2 \div P_1$. Consequently, the "overall" probability of selection was $P_1 \times P_2 \div P_1 = P_2$.

Stratum-Specific Sampling Rate Increased ($P_2 > P_1$). The procedures associated with the case in which the probability of selection was increased between year 1 and year 2 were equally straightforward. First, hospitals new to the frame were sampled with probability P_2. Second, hospitals that were selected in year 1 (that remained in the year 2 frame) were selected for the year 2 sample. Third, hospitals that were in the frame for both years 1 and 2, but not selected for the year 1 sample, were selected for the year 2 sample with probability $(P_2\text{-}P_1) \div (1\text{-}P_1)$.

The justification for this sampling rate, $(P_2\text{-}P_1) \div (1\text{-}P_1)$, is somewhat complex. In year 1 certain frame hospitals were included in the sample at the rate P_1. This can also be viewed as having excluded a set of hospitals at the rate $(1\text{-}P_1)$. Likewise, in year 2 it was imperative that each hospital excluded from the year 1 sample be excluded from the year 2 sample at an overall rate of $(1\text{-}P_2)$.

Since $P_2 > P_1$, then $(1\text{-}P_2) < (1\text{-}P_1)$. Therefore, just as was done for the case of $P_2 < P_1$, multistage selection was implemented. However, it was implemented for exclusion rather than inclusion.

Therefore, those hospitals excluded from the year 1 sample were also excluded from the year 2 sample at the rate $S = (1\text{-}P_2) \div (1\text{-}P_1)$. This gave them the desired overall *exclusion* rate of $(1\text{-}P_1) \times (1\text{-}P_2) \div (1\text{-}P_1) = (1\text{-}P_2)$. Consequently, the *inclusion* rate for these hospitals was set at $1\text{-}S = (P_2\text{-}P_1) \div (1\text{-}P_1)$.

Zero-Weight Hospitals

To enhance researchers' ability to study the effects of hospital splits and merges, if a hospital was the result of either a split or a merge involving one or more NIS sample hospitals, it was added to the NIS file. However, unless it was selected as a part of the regular NIS sample, it was assigned a sampling weight of zero. Also, any NIS hospital that closed (according to the AHA) was retained in the NIS file and assigned sample weights of zero, if it was not selected for the regular NIS sample in the year it closed. These zero-weight hospitals were included in all following years if inpatient data were available. However, no attempt was made to include these zero-weight hospitals in previous years. For example, if a hospital first appeared in 1990 as a zero-weight hospital, then the hospital would also be added to the 1991-1992 NIS files, but not the 1988-1989 NIS files.

Ten Percent Subsamples

Two non-overlapping 10 percent subsamples of discharges were drawn from the NIS file for each year. The subsamples were selected by drawing every tenth discharge starting with two different starting points (randomly selected between 1 and 10). Having a different starting point for each of the two subsamples guaranteed that they would not overlap. Discharges were sampled so that 10 percent of each hospital's discharges in each quarter were selected for each of the subsamples. The two samples can be combined to form a single, generalizable 20 percent subsample of discharges.

FINAL HOSPITAL SAMPLE

The annual numbers of hospitals and discharges in the NIS, Release 1 are shown in Table 9, for both the regular NIS sample and the total sample (which includes zero-weight hospitals).

Table e. NIS Hospital Sample

Calendar Year	Regular Sample		Total Sample	
	Number of Hospitals	Number of Discharges	Number of Hospitals	Number of Discharges
1988	758	5,242,904	759	5,265,756
1989	875	6,067,667	882	6,110,064
1990	861	6,156,638	871	6,268,515
1991	847	5,984,270	859	6,156,188
1992	838	6,008,001	856	6,195,744
Total		29,459,480		29,996,267

A more detailed breakdown of the regular NIS hospital sample (excluding zero-weight hospitals), by calendar year and geographic region is shown in Table 10. For each calendar year and each geographic region, Table 10 shows the number of:

- universe hospitals (Universe),
- frame hospitals (Frame),
- sampled hospitals (Sample),
- target hospitals (Target = 20 percent of the universe), and
- shortfall hospitals (Shortfall = Sample - Target).

For example, in 1988 the Northeast region contained 825 hospitals in the universe. It also contained 193 hospitals in the frame, of which 141 hospitals were drawn for the sample. This was 24 hospitals short of the overall target sample size of 165.

From Table 10 it is clear that most of the 1988 shortfall occurred in the North Central and Southern regions. The addition of Wisconsin to the frame in 1989 significantly reduced the shortfall in the North Central region. However, the large shortfall of over 200 hospitals in the Southern region persisted throughout the study period 1988-1992, because only Florida hospitals were in the frame for this release of the NIS.

Table 11 shows the number of hospitals in the universe, frame, and regular sample for each state in the sampling frame for 1988 and 1992. In all states except Illinois, the difference between the universe and the frame represents the number of AHA community hospitals for which no data were received from that state's data source. As explained earlier, the number of hospitals in the Illinois frame is approximately 40 percent of the number in the Illinois universe, as stipulated in agreements with the data source.

The number of hospitals in the NIS hospital sample that continue across multiple sample years is shown in Table 12. From Table 12 it is clear that longitudinal cohorts that include 1988 are the smallest, because the total number of sample hospitals was smallest for 1988 (758 hospitals). However, if 1989 is taken as a starting year, it can then be seen that 93.1 percent of the 1989 hospital sample continued in the 1990 sample (815 of 875). Likewise, the 87.2 percent and 81.0 percent of the 1989 sample hospitals continued on through 1991 and 1992, respectively.

Table f. Number of Hospitals: Universe, Frame, Regular Sample, Target, and Shortfall By Year and Region

Calendar						
Year	**Region**	**Universe**	**Frame**	**Sample**	**Target**	**Shortfall**
1988	NE	825	193	141	165	-24
	NC	1,600	208	206	320	-114
	S	2,132	224	200	426	-226
	W	1,050	622	211	210	1
	Total	5,607	1,247	758	1,121	-363
1989	**Region**					
	NE	813	423	165	163	2
	NC	1,582	340	305	316	-11
	S	2,114	222	199	423	-224
	W	1,039	673	206	208	-2
	Total	5,548	1,658	875	1,110	-235
1990	**Region**					
	NE	806	412	166	161	5
	NC	1,574	338	304	315	-11
	S	2,076	218	197	415	-218
	W	1,012	652	194	202	-8
	Total	5,468	1,620	861	1,094	-233
1991	**Region**					
	NE	798	406	162	160	2
	NC	1,560	337	297	312	-15
	S	2,056	215	195	411	-216
	W	998	646	193	200	-7
	Total	5,412	1,604	847	1,082	-235
1992	**Region**					
	NE	790	408	167	158	9
	NC	1,543	334	293	309	-16
	S	2,018	209	192	404	-212
	W	983	640	186	197	-11
	Total	5,334	1,591	838	1,067	-229

Due to rounding, values for regions may not sum to total.

Table g. Number of Hospitals in the Universe, Frame, and Regular Sample for Each State in the Sampling Frame: 1988 and 1992

Calendar Year	State	Universe	Frame	Sample
1988	CA	471	463	140
	CO	80	62	29
	FL	238	224	200
	IA	127	121	119
	IL	221	87	87
	MA	110	103	83
	NJ	92	90	58
	WA	99	97	42
	Total	1,438	1,247	758
1992	**State**			
	AZ	60	47	15
	CA	437	434	114
	CO	71	69	29
	FL	224	209	192
	IA	121	119	106
	IL	211	87	79
	MA	102	92	43
	NJ	97	89	32
	PA	232	227	92
	WA	91	90	28
	WI	128	128	108
	Total	1,774	1,591	838

Table h. Number of Hospitals and Discharges in Longitudinal Cohort

Number of Years	Calendar Years	Longitudinal Regular Sample Hospitals	% of Base Year Sample	Longitudinal Regular Sample Discharges
2	1988-1989	610	80.5	8,492,039
	1989-1990	815	93.1	11,525,749
	1990-1991	802	93.1	11,297,175
	1991-1992	781	92.2	11,272,981
3	1988-1990	573	75.6	12,168,677
	1989-1991	763	87.2	16,074,381
	1990-1992	745	86.5	16,085,651
4	1988-1991	542	71.5	15,096,807
	1989-1992	709	81.0	20,340,970
5	1988-1992	502	66.2	18,106,098

SAMPLING WEIGHTS

Although the sampling design was simple and straightforward, it is necessary to incorporate sample weights to obtain state and national estimates. Therefore, sample weights were developed separately for hospital- and discharge-level analyses for each year from 1988 to 1992. Three hospital-level weights were developed to weight NIS sample hospitals to the state, frame, and universe. Similarly, three discharge-level weights were developed to weight NIS sample discharges to the state, frame, and universe.

Hospital-Level Sampling Weights

Universe Hospital Weights. Hospital weights to the universe were calculated by post-stratification. For each calendar year, hospitals were stratified on the same variables that were used for sampling: geographic region, urban/rural location, teaching status, bedsize, and ownership. The strata that were collapsed for sampling were also collapsed for sample weight calculations. Within stratum s, each NIS sample hospital's universe weight was calculated as:

$$W_s(\text{universe}) = N_s(\text{universe}) \div N_s(\text{sample}),$$

where N_s(universe) and N_s(sample) were the number of community hospitals within stratum s in the universe and sample, respectively. Thus, each hospital's universe weight is equal to the number of universe hospitals it represented during that calendar year.

Frame Hospital Weights. Hospital-level sampling weights were also calculated to represent the entire collection of states in the frame using the same post-stratification scheme as described above for the weights to represent the universe. For each year, within stratum s, each NIS sample hospital's frame weight was calculated as:

$$W_s(\text{frame}) = N_s(\text{frame}) \div N_s(\text{sample}).$$

N_s(frame) was the total number of universe community hospitals within stratum s in the states that contributed data to the frame. N_s(sample) was the number of sample hospitals selected for the NIS in stratum s. Thus, each hospital's frame weight is equal to the number of universe hospitals it represented in the frame states during that calendar year.

State Hospital Weights. For each year, a hospital's weight to its state was calculated in a similar fashion. Within each state, strata often had to be collapsed after sample selection for development of weights to ensure a minimum of two sample hospitals within each stratum. For each state and each year, within stratum s, each NIS sample hospital's state weight was calculated as:

$$W_s(\text{state}) = N_s(\text{state}) \div N_s(\text{state sample}).$$

N_s(state) was the number of universe community hospitals in the state within stratum s. N_s(state sample) was the number of hospitals selected for the NIS from that state in stratum s. Thus, each hospital's state weight is equal to the number of hospitals that it represented in its state during that calendar year.

All of these hospital weights can be rescaled if necessary for selected analyses, to sum to the NIS hospital sample size each year.

Discharge-Level Sampling Weights

The calculations for discharge-level sampling weights were very similar to the calculations of hospital-level sampling weights. The discharge weights usually are constant for all discharges within a stratum.

The only exceptions were for strata with sample hospitals that, according to the AHA files, were open for the entire calendar year but contributed less than their full year of data to the NIS. For those hospitals, we *adjusted* the number of observed discharges by a factor 4 ÷ Q, where Q was the number of calendar quarters that the hospital contributed discharges to the NIS. For example, when a sample hospital contributed only two quarters of discharge data to the NIS, the *adjusted* number of discharges was double the observed number.

With that minor adjustment, each discharge weight is essentially equal to the number of reference (universe, frame, or state) discharges that each sampled discharge represented in its stratum. This calculation was possible because the number of total discharges was available for every hospital in the universe from the AHA files. Each universe hospital's AHA discharge total was calculated as the sum of newborns and total facility discharges.

Universe Discharge Weights. Discharge weights to the universe were calculated by post-stratification. For each calendar year, hospitals were stratified just as they were for universe hospital weight calculations. Within stratum s, for hospital i, each NIS sample discharge's universe weight was calculated as:

$$DW_{is}(\text{universe}) = [DN_s(\text{universe}) \div ADN_s(\text{sample})] * (4 \div Q_i),$$

where DN_s(universe) was the number of discharges from community hospitals in the universe within stratum s; ADN_s(sample) was the number of *adjusted* discharges from sample hospitals selected for the NIS; and Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's weight is equal to the number of universe discharges it represented in stratum s during that calendar year.

Frame Discharge Weights. Discharge-level sampling weights were also calculated to represent all discharges from the entire collection of states in the frame using the same post-stratification scheme described above for the discharge weights to represent the universe. For each year, within stratum s, for hospital i, each NIS sample discharge's frame weight was calculated as:

$$W_{is}(\text{frame}) = [DN_s(\text{frame}) \div ADN_s(\text{sample})] * (4 \div Q_i),$$

DN_s(frame) was the number of discharges from all community hospitals in the states that contributed to the frame within stratum s. ADN_s(sample) was the number of *adjusted* discharges from sample hospitals selected for the NIS in stratum s. Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharges's frame weight is equal to the number of discharges it represented in the frame states during that calendar year.

State Discharge Weights. For each year, a discharge's weight to its state was similarly calculated. Strata were collapsed in the same way as they were for the state hospital weights to ensure a minimum of two sample hospitals within each stratum. For each year, within stratum s, for hospital i, each NIS sample discharge's state weight was calculated as:

$$W_{is}(\text{state}) = [DN_s(\text{state}) \div ADN_s(\text{state sample})] * (4 \div Q_i),$$

DN_s(state) was the number of discharges from all community hospitals in the state within stratum s. ADN_s(state sample) was the *adjusted* number of discharges from hospitals selected for the NIS from that state in stratum s. Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's state weight is equal to the number of discharges that it represented in its state during that calendar year.

All of these discharge weights can be rescaled if necessary for selected analyses, to sum to the NIS discharge sample size each year.

Discharge Weights for 10 Percent Subsamples

In the 10 percent subsamples, each discharge had a 10 percent chance of being drawn. Therefore, the discharge weights contained in the Hospital Weights file can be multiplied by 10 for each of the subsamples, or multiplied by 5 for the two subsamples combined.

DATA ANALYSIS

Variance Calculations

It may be important for researchers to calculate a measure of precision for some estimates based on the NIS sample data. Variance estimates must take into account both the sampling design and the form of the statistic. The sampling design was a stratified, single-stage cluster sample. A stratified random sample of hospitals (clusters) was drawn and then *all* discharges were included from each selected hospital.

If hospitals inside the frame were similar to hospitals outside the frame, the sample hospitals can be treated as if they were randomly selected from the entire universe of hospitals within each stratum. Standard formulas for a stratified, single-stage cluster sampling without replacement could be used to calculate statistics and their variances in most applications.

A multitude of statistics can be estimated from the NIS data. Several computer programs are listed below that calculate statistics and their variances from sample survey data. Some of these programs use general methods of variance calculations (e.g., the jackknife and balanced half-sample replications) that take into account the sampling design. However, it may be desirable to calculate variances using formulas specifically developed for some statistics.

In most cases, computer programs are readily available to perform these calculations. For instance, OSIRIS IV, developed at the University of Michigan, does calculations for numerous statistics arising from the stratified, single-stage cluster sampling design.

These variance calculations are based on finite-sample theory, which is an appropriate method for obtaining cross-sectional, nationwide estimates of outcomes. According to finite-sample theory, the intent of the estimation process is to obtain estimates that are precise representations of the nationwide population at a specific point in time. In the context of the NIS, any estimates that attempt to accurately describe characteristics (such as expenditure and utilization patterns or hospital market factors) and interrelationships among characteristics of hospitals and discharges during a specific year from 1988 to 1992 should be governed by finite-sample theory.

Alternatively, in the study of hypothetical population outcomes not limited to a specific point in time, analysts may be less interested in specific characteristics from the finite population (and time period) from which the *sample* was drawn, than they are in hypothetical characteristics of a conceptual "superpopulation" from which any particular finite *population* in a given year might have been drawn. According to this superpopulation model, the nationwide population in a given year is only a snapshot in time of the possible interrelationships among hospital, market, and discharge characteristics. In a given year, all possible interactions between such characteristics may not have been observed, but analysts may wish to predict or simulate interrelationships that may occur in the future.

Under the finite-population model, the variances of estimates approach zero as the sampling fraction approaches one, since the population is defined at that point in time, and because the estimate is for a characteristic as it existed at the time of sampling. This is in contrast to the superpopulation model, which adopts a stochastic viewpoint rather than a deterministic viewpoint. That is, the nationwide population in a particular year is viewed as a random sample of some underlying superpopulation over time.

Different methods are used for calculating variances under the two sample theories. Under the superpopulation (stochastic) model, procedures (such as those described by Potthoff, Woodbury, and Manton[2]) have been developed to draw inferences using weights from complex samples. In this context, the survey weights are not used to weight the sampled cases to the universe, because the universe is conceptually infinite in size. Instead, these weights are used to produce unbiased estimates of parameters that govern the superpopulation.

In summary, the choice of an appropriate method for calculating variances for nationwide estimates depends on the type of measure and the intent of the estimation process.

Computer Software for Variance Calculations

The hospital weights will be useful for producing hospital-level statistics for analyses that use the *hospital* as the unit of analysis, and the discharge weights will be useful for producing discharge-level statistics for analyses that use the *discharge* as the unit of analysis. These would be used to weight the sample data in estimating population statistics.

Several statistical programming packages allow weighted analyses.[3] For example, nearly all SAS (Statistical Analysis System) procedures incorporate weights.

In addition, several publicly available subroutines have been developed specifically for calculating statistics and their standard errors from survey data:

- OSIRIS IV was developed by L. Kish, N. Van Eck, and M. Frankel at the Survey Research Center, University of Michigan. It consists of two main programs for estimating variances from complex survey designs.

- SUDAAN, a set of SAS subroutines, was developed at the Research Triangle Institute by B. V. Shah. It is adequate for handling most survey designs with stratification. The procedures can handle estimation and variance estimation for means, proportions, ratios, and regression coefficients.

- SUPER CARP (Cluster Analysis and Regression Program) was developed at Iowa State University by W. Fuller, M. Hidiroglou, and R. Hickman. This program computes estimates and variance estimates for multistage, stratified sampling designs with arbitrary probabilities of selection. It can handle estimated totals, means, ratios, and regression estimates.

The NIS database includes a Hospital Weights file with variables required by these programs to calculate finite population statistics. In addition to the sample weights described earlier, hospital identifiers (PSUs), stratification variables, and stratum-specific totals for the numbers of discharges and hospitals are included so that finite-population corrections (FPCs) can be applied to variance estimates.

In addition to these subroutines, standard errors can be estimated by validation and cross-validation techniques. Given that a very large number of observations will be available for most analyses, it may be feasible to set aside a part of the data for validation purposes. Standard errors and confidence intervals can then be calculated from the validation data. If the analytical file is too small to set aside a large validation sample, cross-validation techniques may be used.

For example, tenfold cross-validation would split the data into ten equal-sized subsets. The estimation would take place in ten iterations. At each iteration, the outcome of interest is predicted for one-tenth of the observations by an estimate based on a model fit to the other nine-tenths of the observations. Unbiased estimates of error variance are then obtained by comparing the actual values to the predicted values obtained in this manner.

Finally, it should be noted that a large array of hospital-level variables are available for the entire universe of hospitals, including those outside the sampling frame. For instance, the variables from the AHA surveys and from the Medicare Cost Reports are available for nearly all hospitals. To the extent that hospital-level outcomes correlate with these variables, they may be used to sharpen regional and nationwide estimates.

As a simple example, each hospital's number of C-sections would be correlated with their total number of deliveries. The number of C-sections must be obtained from discharge data, but the number of deliveries is available from AHA data. Thus, if a regression can be fit predicting C-sections from deliveries based on the NIS data, that regression can then be used to obtain hospital-specific estimates of the number of C-sections for all hospitals in the universe.

Longitudinal Analyses

As previously shown in Table 12, hospitals that continue in the NIS for multiple consecutive years are a subset of the hospitals in the NIS for any one of those years. Consequently, longitudinal analyses of hospital-level outcomes may be biased if they are based on any subset of NIS hospitals limited to continuous NIS membership. In particular, such subsets would tend to contain fewer hospitals that opened, closed, split, merged, or changed strata. Further, the sample weights were developed as annual, cross-sectional weights rather than longitudinal weights. Therefore, different weights might be required, depending on the statistical methods employed by the analyst.

One approach to consider in hospital-level longitudinal analyses is to use repeated-measure models that allow hospitals to have missing values for some years. However, the data are not actually missing for some hospitals, such as those that closed during the study period. In any case, the analyses may be more efficient (e.g., produce more precise estimates) if they account for the potential correlation between repeated measures on the same hospital over time, yet incorporate data from all hospitals in the sample during the study period.

Discharge Subsamples

The two non-overlapping 10 percent subsamples of discharges were drawn from the NIS file for each year for several reasons pertaining to data analysis. One reason for creating the subsamples was to reduce processing costs for selected studies that will not require the entire NIS. Another reason is that the two subsamples may be used to validate models and obtain unbiased estimates of standard errors. That is, one subsample may be used to estimate statistical models, and the other subsample may be used to test the fit of those models on new data. This is a very important analytical step, particularly in exploratory studies, where one runs the risk of fitting noise.

For example, it is well known that the percentage of variance explained by a regression, R^2, is generally overestimated by the data used to fit a model. The regression model could be estimated from the first subsample and then applied to the second subsample. The squared correlation between the actual and predicted value in the second subsample is an unbiased estimate of the model's true explanatory power when applied to new data.

ENDNOTES

1. Coffey, R. and D. Farley (1988, July). *HCUP-2 Project Overview,* (DHHS Publication No. (PHS) 88-3428. Hospital Studies Program Research Note 10, National Center for Health Services Research and Health Care Technology Assessment, Rockville, MD: Public Health Service.

2. Potthoff, R.F., M.A. Woodbury, and K.G. Manton (1992). "Equivalent Sample Size" and "Equivalent Degrees of Freedom" Refinements for Inference Using Survey Weights Under Superpopulation Models. *Journal of the American Statistical Association*, Vol. 87, 383-396.

3. Carlson, B.L., A.E. Johnson, and S.B. Cohen (1993). An Evaluation of the Use of Personal Computers for Variance Estimation with Complex Survey Data. *Journal of Official Statistics*, Vol. 9, No. 4, 795-814.