TECHNICAL SUPPLEMENT 7:
DESIGN OF THE HCUP NATIONWIDE INPATIENT SAMPLE, RELEASE 3

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. The NIS, Release 1 covers calendar years 1988-1992. The NIS, Release 2 covers calendar year 1993, and the NIS, Release 3 covers calendar year 1994. 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.

This third release of the NIS contains 6.4 million discharges from a sample of 904 hospitals in 17 states. The first release (1988 through 1992) contains 5.2 to 6.2 million discharges per year from a sample of 758 to 875 hospitals per year in 11 states (8 states for 1988). The second release of the NIS contains 6.5 million discharges from a sample of 913 hospitals in 17 states. 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, Release 3 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. Tables include cumulative information for NIS, Release 1 (1988 through 1992); NIS, Release 2 (1993); and NIS, Release 3 (1994) to provide a longitudinal view of the database.

THE NIS HOSPITAL UNIVERSE

The hospital universe is defined by all hospitals that were open during any part of the 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 21 shows the number of universe hospitals for each year based on the AHA Annual Survey.

Table q. Hospital Universe[1]

Year	Number of Hospitals
1988	5,607
1989	5,548
1990	5,468
1991	5,412
1992	5,334
1993	5,313
1994	5,290

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, Midwest, 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) *Control – government nonfederal, 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 22.

Table r. 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 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. At the time when the sample was drawn, the Agency for Health Care Policy and Research (AHCPR) had 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 the first release of the NIS, and an additional 6 states have been included in the second and the third release of the NIS, as shown in Table 23.

Table s. States in the Frame for the NIS, Release 1, NIS, Release 2, and NIS, Release 3

Years	States in the Frame
NIS, Release 1	
1988	California, Colorado, Florida, Iowa, Illinois, Massachusetts, New Jersey, and Washington
1989-1992	Add Arizona, Pennsylvania, and Wisconsin
NIS, Release 2	
1993	Add Connecticut, Kansas, Maryland, New York, Oregon, South Carolina
NIS, Release 3	
1994	Add Connecticut, Kansas, Maryland, New York, Oregon, South Carolina

The list of the entire frame of hospitals was composed of all AHA community hospitals in each of the frame states *that could be matched to the discharge data provided to HCUP*, with restrictions on the hospitals that could be included from Illinois and South Carolina. 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).

The Illinois Health Care Cost Containment Council stipulated that no more than 40 percent of the data provided by Illinois could be included in the database for any calendar quarter. As a result, the number of Illinois community hospitals in the frame was restricted, and 104 of the 208 Illinois community hospitals in the universe (50 percent of hospitals) were randomly selected 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. This prevented the sample from including more than 40 percent of Illinois discharges.

South Carolina stipulated that only hospitals that appear in sampling strata with two or more hospitals were to be included in the NIS. Four South Carolina hospitals were excluded from the

frame since there were fewer than two South Carolina hospitals in four sampling frame strata. The remaining 59 South Carolina community hospitals are included in the frame.

The number of frame hospitals for each year is shown in Table 24.

Table t. Hospital Frame

Year	Number of Hospitals
1988	1,247
1989	1,658
1990	1,620
1991	1,604
1992	1,591
1993	2,168
1994	2,135

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, which includes hospitals outside the frame (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 17 states contributed data to this third release, some estimates may differ from estimates from comparative data sources. 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 to determine the appropriateness of the NIS for specific analyses.

The target sample size was 20 percent of the total number of community hospitals in the U.S. for 1994. 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 is preferred 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[2] 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 yield good estimates. Furthermore, because the data are to be used for purposes other than producing national estimates, it is critical that all hospital types (including small hospitals) are 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, subject to sampling restrictions specified by states. 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.

A systematic random sample was drawn 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.

1994 NIS Hospital Sampling Procedure

The 1994 sample was drawn by a procedure that retained most of the 1993 hospitals, while allowing hospitals new to the frame an opportunity to enter the 1994 NIS.

Even in frame states that were present in the 1993 sample, hospitals that opened in 1994 needed a chance to enter the sample. Also, hospitals that changed strata between 1993 and 1994 were considered new to the 1994 frame.

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 following procedure provides rules for creating a "year 2" sample, given that a "year 1" sample had already been drawn. In this example, year 1 would be 1993 and year 2 would be 1994. 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 (see Technical Supplement: *Design of the HCUP Nationwide Inpatient Sample, Release 1*, section "1988 NIS Hospital Sampling Procedure.")

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 in Technical Supplement: *Design of the HCUP Nationwide Inpatient Sample, Release 1.*

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 $(P_2-P_1) \div (1-P_1)$.

The justification for this sampling rate, $(P_2-P_1) \div (1-P_1)$, 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 $(1-P_1)$. 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 $(1-P_2)$.

Since $P_2 > P_1$, then $(1-P_2) < (1-P_1)$. 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 $S = (1-P_2) \div (1-P_1)$. This gave them the desired overall *exclusion* rate of $(1-P_1) \times (1-P_2) \div (1-P_1) = (1-P_2)$. Consequently, the *inclusion* rate for these hospitals was set at $1-S = (P_2-P_1) \div (1-P_1)$.

Zero-Weight Hospitals

The 1994 sample contains no zero-weight hospitals. For a description of zero-weight hospitals in the 1988-1992 sample, see the Technical Supplement: *Design of the HCUP Nationwide Inpatient Sample, Release 1.*

Ten Percent Subsamples

Two nonoverlapping 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 NIS, Release 1; NIS, Release 2; and NIS Release 3 are shown in Table 25, for both the regular NIS sample and the total sample (which includes zero-weight hospitals for 1988-1992).

Table u. NIS Hospital Sample

Year	Regular Sample		Total Sample	
	Number of Hospitals	Number of Discharges	Number of Hospitals	Number of Discharges
NIS, Release 1				
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
NIS, Release 2				
1993	913	6,538,976	913	6,538,976
NIS, Release 3				
1994	904	6,385,011	904	6,385,011
Total		42,383,467		42,920,254

A more detailed breakdown of the 1994 NIS hospital sample by geographic region is shown in Table 26. For each geographic region, Table 26 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).

Table v. Number of Hospitals in Universe, Frame, Regular Sample, Target, and Shortfall By Region, 1994

Region	Universe	Frame	Sample	Target	Shortfall
NE	780	654	168	156	12
MW	1,527	473	304	305	-1
S	2,010	313	256	403	-147
W	973	695	176	195	-19

Total	5,290	2,135	904	1,059	-155

For example, in 1994 the Northeast region contained 780 hospitals in the universe. It also contained 654 hospitals in the frame, of which 168 hospitals were drawn for the sample. This was 12 hospitals more than the target sample size of 156.

Table 27 shows the number of hospitals in the universe, frame, and regular sample for each state in the sampling frame for 1994. In all states except Illinois and South Carolina, the difference between the universe and the frame represents the difference in the number of community hospitals in the 1994 AHA Annual Survey of Hospitals and the number of community hospitals for which data were supplied to HCUP. As explained earlier, the number of hospitals in the Illinois frame is approximately 50 percent of the hospitals in the Illinois universe in order to comply with the agreement with the data source concerning the restriction on the number of Illinois discharges. The number of hospitals in the South Carolina frame is eight fewer than the South Carolina universe. Four hospitals were excluded because of sampling restrictions stipulated by South Carolina, and four hospitals were not included in the data supplied to HCUP.

The number of hospitals in the NIS hospital samples that continue across multiple sample years is shown in Table 28. This table will be of interest to those who may combine Release 1, 2, and 3 of the NIS. Table 28 shows that longitudinal cohorts that span several years and include 1988 and 1993 are the lowest in number of continuing sample hospitals. For example, if 1988 is taken as a starting year, only 44.2 percent of the 1988 hospital sample continued in the 1994 sample (335 of 758).

Table w. Number of Hospitals in the Universe, Frame, and Regular Sample for States in the Sampling Frame: 1994

State	Universe	Frame	Sample
AZ	61	49	12
CA	430	428	102
CO	69	68	22
CT	37	32	7
FL	220	204	163
IA	116	116	64
IL	208	104	77
KS	137	126	71
MA	95	84	27
MD	50	50	42
NJ	94	86	19
NY	230	228	62
OR	63	62	19
PA	230	224	53
SC	67	59	51
WA	90	88	21
WI	127	127	92
Total	2324	2135	904

Table x. 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
	1992-1993	609	72.7	8,804,638
	1993-1994	693	75.9	10,271,404
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
	1991-1993	570	67.3	12,559,421
	1992-1994	540	64.4	11,279,667
4	1988-1991	542	71.5	15,096,807
	1989-1992	709	81.0	20,340,970
	1990-1993	548	63.6	16,023,500
	1991-1994	508	60.0	14,481,319
5	1988-1992	502	66.2	18,106,098
	1989-1993	523	59.8	19,000,777
	1990-1994	490	56.9	17,437,229
6	1988-1993	378	49.9	16,906,818
	1989-1994	471	53.8	19,987,910
7	1988-1994	335	44.2	17,128,064

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. 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 year, hospitals were stratified on the same variables that were used for sampling: geographic region, urban/rural location, teaching status, bedsize, and control. 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 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 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 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 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. 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 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 discharge's frame weight is equal to the number of discharges it represented in the frame states during that year.

State Discharge Weights. 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. 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 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) were 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, and SUDAAN, developed at the Research Triangle Institute, do calculations for numerous statistics arising from the stratified, single-stage cluster sampling design. An example of using SUDAAN to calculate variances in the NIS is presented in Technical Supplement: *Calculating Variances Using Data from the HCUP Nationwide Inpatient Sample.*[3]

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 1994 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[4]) 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.[5] 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 28, 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 nonoverlapping 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. Most AHA surveys do not cover a January-to-December calendar year. The number of hospitals for 1988-1991 are based on the HCUP calendar-year version of the AHA Annual Survey files. To create a calendar-year reporting period, data from the AHA surveys must be apportioned in some manner across calendar years. Survey responses were converted to calendar-year periods for 1988-1991 by merging data from adjacent survey years. The number of hospitals for 1992-1994 are based on the AHA Annual Survey files.

2. 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.

3. Duffy, S.Q. and J.P. Sommers (1996, March). *Calculating Variances Using Data from the HCUP Nationwide Inpatient Sample.* Rockville, MD: Agency for Health Care Policy and Research.

4. 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.

5. 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.