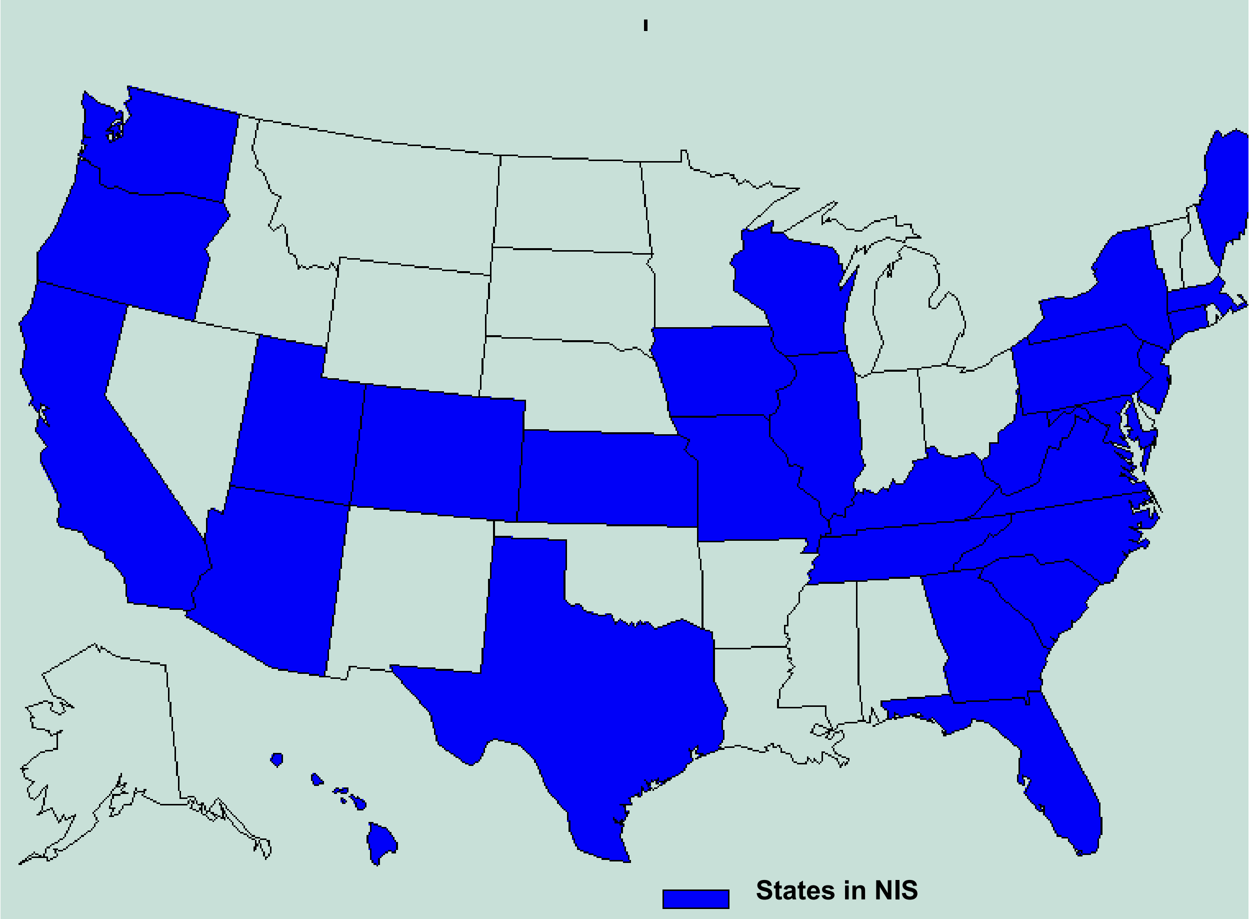


HCUP Nationwide Inpatient Sample

Design of the HCUP Nationwide Inpatient Sample, 2000

May 24, 2002

Table of Contents

Index of Tables

EXECUTIVE SUMMARY

The Nationwide Inpatient Sample (NIS) of the Healthcare Cost and Utilization Project (HCUP) was developed to provide analyses of hospital utilization, charges and quality of care across the United States. The target universe includes all discharges from non-rehabilitation, community hospitals in the United States that were open during any part of the calendar year 2000. There were 4,839 hospitals in the hospital universe in 2000. The 2000 NIS comprises all discharges from a sample of hospitals in this target universe.

This report provides a detailed description of the 2000 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. Previous NIS releases covered 1988 through 1999. Cumulative information is provided for all previous years to provide a longitudinal view of the database.

Hospital Sample Design

The sampling 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. There were 3,034 hospitals in the 2000 sampling frame, a 20% increase from the 1999 NIS. 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. With this objective in mind, NIS sampling strata were defined based on five hospital characteristics contained in the AHA hospital files.

1. Geographic Region - Northeast, Midwest, West, and South;
2. Control – public, private not-for-profit and proprietary;
3. Location – urban or rural;
4. Teaching Status – teaching or non-teaching;
5. Bed size – small, medium, and large.

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. The resulting sample included 994 hospitals, 20.5% of the total hospital universe.

At the time the 2000 sample was drawn, the Agency for Healthcare Research and Quality (AHRQ) had agreements with data sources from twenty-eight states. Over 90% of the hospital universe is included in the sampling frame for all but five of these states. For four states, Hawaii, Illinois, Missouri and South Carolina, sampling restrictions were dictated by the data source with the result that 61% to 82% of state hospitals were included in the sampling frame. (Restrictions from other states did not have an appreciable effect on the percentage of hospitals in the sampling frame.)

For one state, Texas, only 70% of the hospitals supplied data to HCUP. Certain Texas state-licensed hospitals, primarily the smaller hospitals, are exempt from statutory reporting requirements. As a result, small hospitals are substantially less likely to be included in the sampling frame, while larger hospitals are more likely to be included. Although the number of hospitals omitted appears sizable, these missing hospitals contain only 6% of Texas discharges.

While 20% of the hospitals from each region are selected for the NIS, the comprehensiveness of the sampling frame varies by region. In the Northeast, 90% of hospitals are included in the sampling frame, while 77% of those hospitals in the West are included, with 63% in the South, and 40% in the Midwest.

Because the NIS sampling frame has a disproportionate representation of the more populous states, and hospitals with more annual discharges, its comprehensiveness in terms of discharges is higher. The proportion of the regional population in the NIS states ranges from 95% in the Northeast to 45% in the Midwest.

NIS Sample

The final NIS sample included 7,450,992 discharges from 994 hospitals selected from all 28 frame states. Hospitals were sampled throughout each region of the United States. In the Northeast and West, where a higher proportion of states are represented, relatively fewer hospitals are sampled from each state than in the South and Midwest, where the proportion of states in the NIS is lower.

Ideally, relationships among outcomes and their correlates estimated from the NIS should generally hold across all U.S. hospitals. However, since only 28 states contributed data to this 2000 release, some estimates may differ from estimates derived from comparative data sources. See the *1999 HCUP Nationwide Inpatient Sample (NIS) Comparison Report* for a comparison of statistics calculated from the 1999 NIS with estimates from databases with similar populations to assess the comparability of estimates. This special report is available on the 2000 NIS Documentation CD-ROM and on the HCUP Web site at http://www.ahrq.gov/data/hcup/.

Ten Percent Subsamples

Two non-overlapping 10 percent subsamples of discharges are drawn from the NIS file 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. The subsamples were selected by drawing every tenth discharge starting with two different randomly selected starting points. Having a different starting point for each of the two subsamples guaranteed that the resulting subsamples would not overlap.

Sampling Weights

It is necessary to incorporate sample weights to obtain national estimates. Therefore, sample weights were developed separately for hospital- and discharge-level analyses. Within a stratum, each NIS sample hospital's universe weight is equal to the number of universe hospitals it represents during the year. Since 20% of the AHA universe hospitals in each stratum are sampled when possible, the hospital weights are usually around five. The calculations for discharge-level sampling weights are similar to the calculations of hospital-level sampling weights. In the 10 percent subsamples, each discharge has a 10 percent chance of being drawn. Therefore, the discharge weights are multiplied by 10 for each of the subsamples.

In the 2000 NIS files, the discharge weight data elements are named DISCWT and DISCWTCHARGE.

- Prior to the 2000 NIS, DISCWTCHARGE is not available, and DISCWT should be used to create all national estimates.
- For the 2000 NIS, DISCWT should be used to create national estimates for all analyses except those that involve total charges, and DISCWTCHARGE should be used to create national estimates of total charges. Texas discharges were not included in the calculation of DISCWTCHARGE, and DISCWTCHARGE was set to zero for all Texas discharges because total charges were not available for the first half of the year from that state. Consequently, DISCWTCHARGE differs from DISCWT for NIS hospitals in the South region.

Comparability with Prior NIS Releases

There has been an upward trend since the first NIS release across several dimensions: there are more states, more hospitals, more discharges and greater representativeness of the national population. The latter is a natural result of the major increase in states included in the sampling frame. The twenty-eight 2000 NIS states include four more states than the 1999 NIS and 20 more states than the original 1988 NIS. The four Southern states added to the 2000 NIS have substantially increased the percentage of the regional population included, from 46% in the 1999 NIS to 81% in the 2000 NIS. The annual number of hospitals in the NIS has grown from 758 for 1988 to the present 994, while the number of annual discharges has increased from 5.2 million to the current 7.5 million.

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. Standard formulas for a stratified, single-stage cluster sampling without replacement may be used to calculate statistics and their variances in most applications.

The NIS database includes a Hospital Weights file with variables required by statistical software to calculate finite population statistics. In addition to the sample weights described earlier, hospital identifiers (Primary Sampling Units or 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.

Longitudinal Analyses

All frame hospitals within a stratum have an equal probability of selection for the sample, regardless of whether they had been in prior NIS samples. This deviates from the procedure used for earlier samples, prior to data year 1998, which maximized the longitudinal component of the NIS series. 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. 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.

INTRODUCTION

The Nationwide Inpatient Sample (NIS) of the Healthcare Cost and Utilization Project (HCUP) was developed to provide analyses of hospital utilization, charges and quality of care across the United States. 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. The 2000 NIS includes data for calendar year 2000. Previous releases covered 1988 through 1999. Table 1 shows the number of states, hospitals and discharges in each year, and shows that prior years of HCUP included fewer states.

Table 1. Number of NIS States, Hospitals and Discharges, by Year

Calendar Year	States in the Frame	Number of States	Sample Hospitals	Sample Discharges (Millions)
1988–1992	California, Colorado, Florida, Iowa, Illinois, Massachusetts, New Jersey, Washington, Arizona, Pennsylvania, and Wisconsin	8–11	758–875	5.2–6.2
1993	Add Connecticut, Kansas, Maryland, New York, Oregon, South Carolina	17	913	6.5
1994	No new additions	17	904	6.4
1995	Add Missouri, Tennessee	19	938	6.7
1996	No new additions	19	906	6.5
1997	Add Georgia, Hawaii, and Utah	22	1012	7.1
1998	No new additions	22	984	6.8
1999	Add Maine, Virginia	24	984	7.2
2000	Add Kentucky, North Carolina, Texas, and West Virginia	28	994	7.5

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

- frequency,
- charges,
- lengths of stay,
- effectiveness,
- quality of care,
- 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 2000 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 are described in detail. Tables include cumulative information for all previous NIS releases to provide a longitudinal view of the database.

THE NIS HOSPITAL UNIVERSE

The hospital universe is defined as all hospitals located in the U.S. that were open during any part of the calendar year and that 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 (Department of Defense and Indian Health Service) are excluded. Beginning with the 1998 NIS, rehabilitation hospitals were excluded from the NIS hospital universe because the type of care provided and the characteristics of the discharges from these hospitals were markedly different from other short-term hospitals. Table 2 shows the number of universe hospitals for each year based on the AHA Annual Survey.

Table 2. 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
1995	5,260
1996	5,182
1997	5,113
1998	4,915
1999	4,859
2000	4,839

Hospital Merges, Splits, and Closures

All U.S. hospital entities that were designated community hospitals in the AHA hospital file, except rehabilitation hospitals, 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

The NIS sampling strata were defined based on five hospital characteristics contained in the AHA hospital files. In order to improve the representativeness of the NIS, given the expansion of the number of contributing states, the sampling and weighting strategy was evaluated and revised for 1998 and subsequent data years. This included changes to the definitions of the strata variables. A full description of this process can be found in the special report on *Changes in NIS Sampling and Weighting Strategy for 1998.* This report is available on the 2000 NIS Documentation CD-ROM and on the HCUP Web site at http://www.ahrq.gov/data/hcup/. A description of the sampling procedures and definitions of strata variables used from 1988 through 1997 can be found in the special report: *Design of the HCUP Nationwide Inpatient Sample, Release 6.* This report is available on the 1997 NIS Documentation CD-ROM and on the HCUP Web site. Beginning with the 1998 NIS, the stratification variables were defined as follows:

1. *Geographic Region – Northeast, Midwest, West, and South*. This is an important stratification variable 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 (public), private not-for-profit (voluntary) and private investor-owned (proprietary)*. These types of hospitals tend to have different missions and different responses to government regulations and policies. When there were enough hospitals of each type to allow it, hospitals were stratified as public, voluntary, and proprietary. This stratification was used for Southern rural, Southern urban nonteaching, and Western urban nonteaching hospitals. For smaller strata – the Midwestern rural and Western rural hospitals – a collapsed stratification of public versus private was used, with the voluntary and proprietary hospitals combined to form a single ‘private’ category. For all other combinations of region, location and teaching status, no stratification based on control was advisable given the number of hospitals in these cells.

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, is a member of the Council of Teaching Hospitals (COTH) or has a ratio of full-time equivalent interns and residents to beds of .25 or higher.

5. *Bed size – small, medium, and large.* Bed size categories are based on hospital beds, and are specific to the hospital's region, location and teaching status, as shown in Table 3.

Table 3. Bed Size Categories, by Region

Location and Teaching Status	Hospital Bed size		
	Small	Medium	Large
NORTHEAST			
Rural	1-49	50-99	100+
Urban, nonteaching	1-124	125-199	200+
Urban, teaching	1-249	250-424	425+
MIDWEST			
Rural	1-29	30-49	50+
Urban, nonteaching	1-74	75-174	175+
Urban, teaching	1-249	250-374	375+
SOUTH			
Rural	1-39	40-74	75+
Urban, nonteaching	1-99	100-199	200+
Urban, teaching	1-249	250-449	450+
WEST			
Rural	1-24	25-44	45+
Urban, nonteaching	1-99	100-174	175+
Urban, teaching	1-199	200-324	325+

The bed size cutoff points were chosen so that approximately one-third of the hospitals in a given region, location and teaching status combination would be in each bed size category (small, medium or large). Different cutoff points for rural, urban nonteaching, and urban teaching hospitals were used because hospitals in those categories tend to be small, medium, and large, respectively. For example, a medium-sized teaching hospital would be considered a rather large rural hospital. Further, the size distribution is different among regions for each of the urban/teaching categories. For example, teaching hospitals tend to be smaller in the West than they are in the South. Using differing cutoff points in this manner avoids strata with small numbers of hospitals in them.

Rural hospitals were not split according to teaching status, because rural teaching hospitals were rare. For example, in 2000 rural teaching hospitals were less than 1% of the total hospital universe. The bed size categories were defined within location and teaching status because they would otherwise have been redundant. Rural hospitals tend to be small; urban non-teaching 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.

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 random sampling.

HOSPITAL SAMPLING FRAME

The *universe* of hospitals was established as all community hospitals located in the U.S. with the exception, beginning in 1998, of rehabilitation hospitals. However, it was not feasible to obtain and process all-payer discharge data from a random sample of the entire universe of hospitals because 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 the 2000 sample was drawn, the Agency for Healthcare Research and Quality (AHRQ) had agreements with 28 data sources that maintain statewide, all-payer discharge data files to include their data in the HCUP databases. Prior years of HCUP included fewer states, as shown in Table 1.

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*. 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 target universe). Further restrictions were put on the sampling frames for Georgia, Hawaii, Illinois, South Carolina, Missouri, and Tennessee, as described below.

The Illinois Health Care Cost Containment Council stipulated that no more than 40 percent of the discharges provided by Illinois could be included in the database for any calendar quarter. Consequently, it was necessary to reduce the number of Illinois hospitals in the NIS sampling frame by drawing a systematic random sample of Illinois frame hospitals prior to drawing the NIS sample. By trial and error, it was determined that a sample of 67 percent of Illinois frame hospitals ultimately yielded just under 40% of the discharges supplied by Illinois in each calendar quarter of the NIS.

Georgia, Hawaii, South Carolina and Tennessee stipulated that only hospitals that appear in sampling strata with two or more hospitals from the state were to be included in the NIS. Due to this restriction, two Georgia hospitals, five Hawaii hospitals, seven South Carolina hospitals and one Tennessee hospital were excluded from the 2000 NIS sampling frame. Two additional South Carolina hospitals, although in sampling strata with other hospitals, were removed from the sampling frame due to unique characteristics that would make them identifiable.

Missouri stipulated that only hospitals that had signed releases for public use of the data should be included in the NIS. For 2000, 32 Missouri hospitals did not sign releases for public use of the data. These hospitals were excluded from the sampling frame, leaving 73 hospitals in the 2000 frame.

The number of frame hospitals for each year is shown in Table 4.

Table 4. 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
1995	2,284
1996	2,268
1997	2,452
1998	2,438
1999	2,520
2000	3,034

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 (i.e., having 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 28 states contributed data to this 2000 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 2000. This sample size was determined by AHRQ 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:

- AHRQ 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).

- AHRQ 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, (e.g., regression modeling), it is critical that all hospital types, including small hospitals, are adequately represented.

Overview of the 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.

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.

Change to Hospital Sampling Procedure beginning with the 1998 NIS

Beginning with the 1998 NIS sampling procedures, all frame hospitals within a stratum have an equal probability of selection for the sample, regardless of whether they had been in prior NIS samples. This deviates from the procedure used for earlier samples, which maximized the longitudinal component of the NIS series.

Further description of the sampling procedures for earlier releases of the NIS can be found in the special report: *Design of the HCUP Nationwide Inpatient Sample, Release 6*. This report is available on the 1997 NIS Documentation CD-ROM and on the HCUP Web site. For a description of the development of the new sample design for 1998 and subsequent data years, see the special report: *Changes in NIS Sampling and Weighting Strategy for 1998.* This report is available on the 2000 NIS Documentation CD-ROM and on the HCUP Web site.

Zero-Weight Hospitals

Beginning with the 1993 NIS, the NIS samples no longer contain zero-weight hospitals. For a description of zero-weight hospitals in the 1988-1992 samples, see the special report: *Design of the HCUP Nationwide Inpatient Sample, Release 1.* This report is available on the 1988-1992 NIS Documentation CD-ROM.

FINAL HOSPITAL SAMPLE

Table 5 shows the annual numbers of hospitals and discharges in each year of the NIS. For the 1988-1992 NIS, zero-weight hospitals were maintained to provide a longitudinal sample, so figures are presented for both the regular NIS sample and the total sample.

Table 5. Number of Sample Hospitals and Discharges, by Year

	Regular Sample		Total Sample	
Year	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
1993	913	6,538,976	-	-
1994	904	6,385,011	-	-
1995	938	6,714,935	-	-
1996	906	6,542,069	-	-
1997	1,012	7,148,420	-	-
1998	984	6,827,350	-	-
1999	984	7,198,929	-	-
2000	994	7,450,992	-	-

Table 6 shows a summary of the 2000 NIS hospital sample by geographic region and the number of:

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

Table 6. Number of Hospitals in the 2000 Universe, Frame, Target, Sample and Surplus by Region

Hospital Region	Universe	Frame	% of Universe in Frame	Target	Sample	Surplus
Northeast	675	610	90.4%	135	139	4
Midwest	1,398	558	39.9%	280	286	6
South	1,860	1,173	63.1%	372	381	9
West	9,06	693	76.5%	181	188	7
Total	4,839	3,034	62.7%	968	994	26

For example, in 2000 the Northeast region contained 675 hospitals in the universe. It also contained 610 hospitals in the frame, of which 139 hospitals were drawn for the sample. This was four more than the target sample size of 139 hospitals, resulting in a surplus of four hospitals over the target. The total sample exceeded the target by 26 hospitals, with a resulting sample of 20.5% of the total hospital universe. There was a sample surplus in each region because the number of hospitals sampled in each stratum was rounded up to the next highest integer whenever the sample target of 20 percent of the universe was not an integer number of hospitals.

Table 7 shows a summary of the estimated population of the U.S. on July 1, 2000, and of states in the 2000 NIS, by geographic region, based on U.S. Census Bureau population estimates released on the Internet on December 29, 2000. For each geographic region Table 7 shows:

- the estimated U.S. population on July 1, 2000,
- the estimated population on July 1, 2000, of states in the 2000 NIS, and
- the percentage of estimated U.S. population included in states in the 2000 NIS.

Table 7. Percentage of U.S. Population in 2000 NIS States, by Region

Region	Estimated U.S. Population	Population of 2000 NIS States	Percent of U.S. Population in NIS States
Northeast	53,644,868	50,745,042	94.6
Midwest	64,473,034	29,031,025	45.0
South	99,991,010	81,302,933	81.3
West	63,444,653	56,280,631	88.7

For example, the estimated population of the Northeast region on July 1, 2000 was 53,644,868. The estimated population on July 1, 2000, of states in the Northeast region that were included in the 2000 NIS was 50,745,042. This represents 94.6% of the total Northeast region population. The percentage of estimated U.S population included in states in the 2000 NIS was almost as high in the West (88.7%), but was much lower in the Midwest (45.0%). The four Southern states added to the 2000 NIS have substantially increased the percentage of the regional population included, from 45.9% in the 1999 NIS to 81.3% in the 2000 NIS.

Table 8 shows the number of hospitals and discharges in the universe, frame, and sample for each state in the sampling frame for 2000. The difference between the universe and the frame represents the difference in the number of community hospitals in the 2000 AHA Annual Survey of Hospitals and the number of community hospitals for which data were supplied to HCUP in all states except Georgia, Hawaii, Illinois, Missouri and South Carolina.

- The number of hospitals in the Georgia frame is two less than the Georgia universe. Two hospitals were excluded because of sampling restrictions stipulated by Georgia.

- The number of hospitals in the Hawaii frame is eight less than the Hawaii universe. Five hospitals were excluded because of sampling restrictions stipulated by Hawaii, and three hospitals identified in AHA data were not included in the data supplied to HCUP.

- The number of hospitals in the Illinois frame was randomly reduced to approximately 67 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 Missouri frame is forty-six less than the Missouri universe. Thirty-two hospitals were excluded because they signed release for confidential use only, and fourteen hospitals identified in AHA data were not included in the data supplied to HCUP.

- The number of hospitals in the South Carolina frame is eleven less than the South Carolina universe. Nine hospitals were excluded because of sampling restrictions stipulated by South Carolina, and two hospitals identified in AHA data were not included in the data supplied to HCUP.

- The number of hospitals in the Tennessee frame is eight less than the Tennessee universe. One hospital was excluded because of sampling restrictions stipulated by Tennessee, and seven hospitals identified in AHA data were not included in the data supplied to HCUP.

Table 8. Number of Hospitals and Discharges in the 2000 Universe, Frame, and Sample for States in the Sampling Frame

State	Number of Hospitals in Universe	Number of Hospitals in Frame	Number of Hospitals in Sample	Number of Discharges in Sample
AZ	59	55	14	186,060
CA	384	379	93	815,757
CO	67	66	22	131,639
CT	34	31	6	67,516
FL	193	190	55	578,074
GA	150	148	60	314,302
HI	21	13	3	19,619
IA	115	115	54	208,064
IL	194	130	69	634,799
KS	133	121	57	185,410
KY	100	95	31	152,427
MA	73	68	10	35,261
MD	48	47	13	182,428
ME	36	36	16	205,671
MO	119	73	40	322,345
NC	112	109	36	343,893
NJ	76	75	16	221,254
NY	213	213	49	545,429
OR	59	58	18	94,703
PA	192	187	42	371,138
SC	62	51	20	118,436
TN	116	108	31	162,266
TX	408	287	93	723,081
UT	41	40	14	22,629
VA	86	84	21	227,599
WA	83	82	24	169,811
WI	120	119	66	356,523
WV	54	54	21	54,858
TOTAL	**3,348**	**3,034**	**994**	**7,450,992**

It can be seen in Table 9 that only 70% of Texas hospitals were supplied to HCUP. Certain Texas state-licensed hospitals are exempt from statutory reporting requirements. Exempt hospitals include:

1. Hospitals that do not seek insurance payment or government reimbursement, and
2. Rural providers.

   The Texas statute that exempts rural providers from being required to submit data defines a hospital as a rural provider if it:

   (I) is located in a county that:

   (A) has a population estimated by the United States Bureau of the Census to be not more than 35,000 as of July 1 of the most recent year for which county population estimates have been published; or

   (B) has a population of more than 35,000, but that does not have more than 100 licensed hospital beds and is not located in an area that is delineated as an urbanized area by the United States Bureau of the Census; and

   (II) is not a state-owned hospital or a hospital that is managed or directly or indirectly owned by an individual, association, partnership, corporation, or other legal entity that owns or manages one or more other hospitals.

These exemptions apply primarily to the smaller hospitals, and as can be seen from Table 9 below, small hospitals are substantially less likely to be included in the sampling frame, while larger hospitals are more likely to be included.

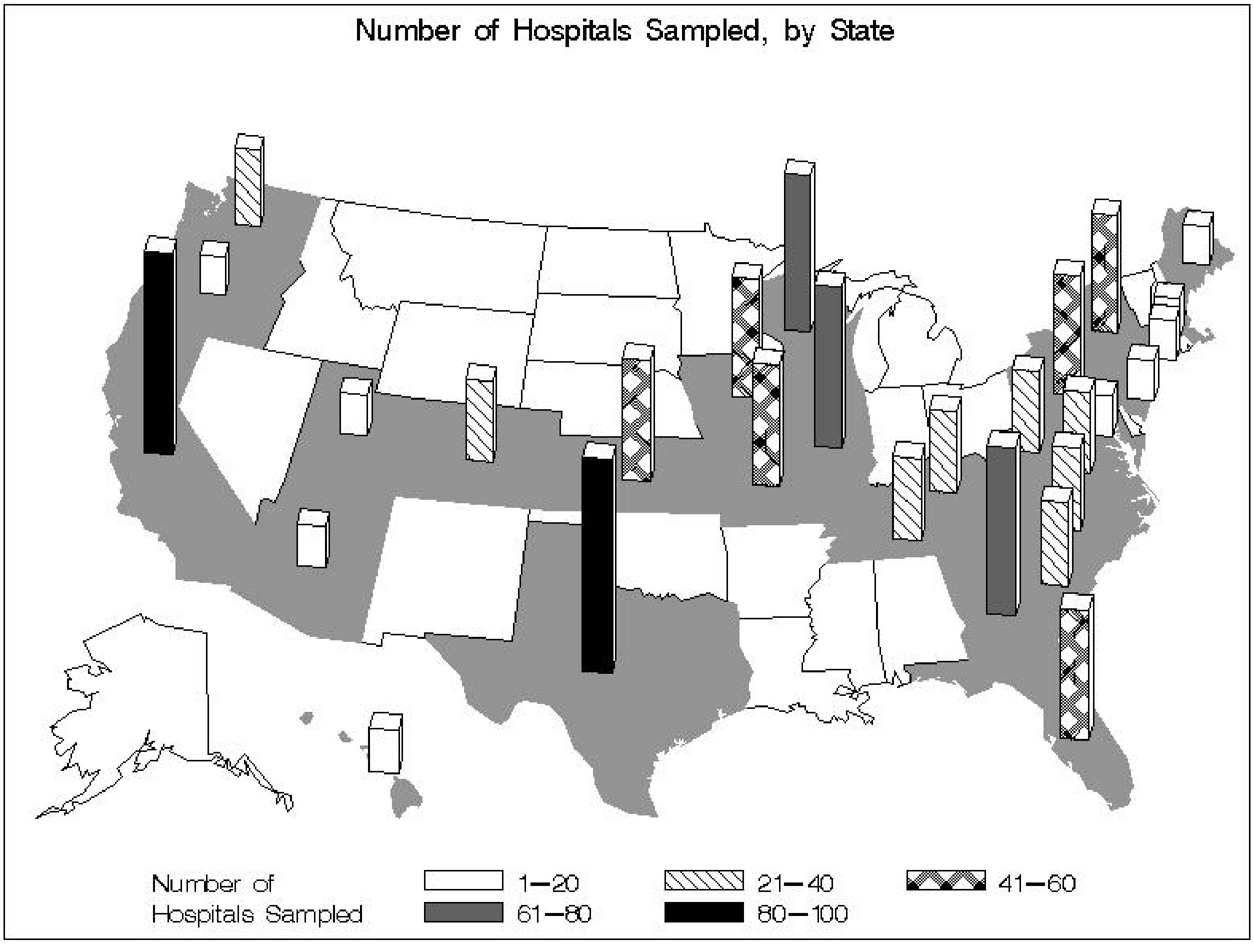
Table 9. Texas Hospitals Included and Excluded from Sampling Frame by Bed Size Category

Bed Size Category	Number of non-rehab AHA Community Hospitals	Number of AHA Discharges	Number of Hospitals In Sampling Frame	Number of Hospitals Not Included in Sampling Frame	Percent of Hospitals Included in Sampling Frame	Percent of Discharges Included in Sampling Frame
Small	199	350,660	109	90	55%	80%
Medium	127	935,381	99	28	78%	92%
Large	82	1,388,588	79	3	96%	99%
TOTAL	**408**	**2,674,629**	**121**	**287**	**70%**	**94%**

While the number of hospitals omitted appears sizable, it can be seen in Table 10 below that the hospitals available to the NIS includes 94% of AHA inpatient discharges from Texas hospitals. At the hospital level, over 85% of non-profit hospitals and 95% of proprietary hospitals are included in the NIS sampling frame.

Table 10. Texas Hospitals and Discharges Included and Excluded from Sampling Frame By Control

Included in Sampling Frame	Control	Number of non-rehab AHA Community Hospitals	Number of AHA Discharges	Mean Bed Size	Percent of Hospitals	Percent of Discharges
NO	Public	95	103,987	31	73%	22%
	Non-Profit	19	39,193	54	13%	3%
	Proprietary	7	14,485	41	5%	2%
TOTAL HOSPITALS NOT IN FRAME		**121**	**157,665**	**35**	**30%**	**6%**
YES	Public	36	360,358	183	27%	78%
	Non-Profit	123	1,294,022	194	87%	97%
	Proprietary	128	862,584	141	95%	98%
TOTAL FRAME HOSPITALS		**287**	**2,516,964**	**169**	**70%**	**94%**
ALL		**408**	**2,674,629**	**130**	**100%**	**100%**


Number of Hospitals Sampled, by State
Number of Hospitals Sampled
1–20
21–40
41–60
61–80
80–100

The above map shows that hospitals were sampled throughout each region of the United States. It also shows the effect of stratification on the sample. In the Northeast and West, where a higher proportion of states are represented, relatively fewer hospitals are sampled from each state than in the South and Midwest, where the proportion of states in the NIS is lower.

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. Hospital-level weights were developed to weight NIS sample hospitals to the hospital universe. Similarly, discharge-level weights were developed to weight NIS sample discharges to the hospital 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, bed size, 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 W_s(universe) was the hospital universe weight, 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 (HOSPWT) is equal to the number of universe hospitals it represented during that year. Since 20% of the hospitals in each stratum were sampled when possible, the hospital weights are usually around five.

Discharge Weights

The calculations for discharge-level sampling weights were 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 of 4 ÷ Q, where Q was the number of calendar quarters for which 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. This adjustment was done only for weighting purposes. The NIS dataset only includes the actual (unadjusted) number of observed discharges.

With that minor adjustment, each discharge weight is essentially equal to the number of AHA universe 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 hospital discharges.

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 DW_{is}(universe) was the discharge weight, 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 (DISCWT) is equal to the number of universe discharges it represented in stratum s during that year. Since all discharges from 20% of the hospitals in each stratum were sampled when possible, the discharge weights are usually around five.
In the 2000 NIS files, the discharge weight data elements are named DISCWT and DISCWTCHARGE. To produce national estimates, use DISCWT or DISCWTCHARGE to weight sampled discharges in the Core file to the discharges from all non-rehabilitation community hospitals located in the U.S.

- Prior to the 2000 NIS, DISCWTCHARGE is not available, and DISCWT should be used to create all national estimates.
- For the 2000 NIS, DISCWT should be used to create national estimates for all analyses except those that involve total charges, and DISCWTCHARGE should be used to create national estimates of total charges. Texas discharges were not included in the calculation of DISCWTCHARGE, and DISCWTCHARGE was set to zero for all Texas discharges because total charges were not available for the first half of the year from that state. Consequently, DISCWTCHARGE differs from DISCWT for NIS hospitals in the South region.

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 were multiplied by 10 for each of the subsamples, and DISCWT or DISCWTCHARGE should be 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.

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 2000 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. 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. The discharge weights would be used to weight the sample data in estimating population statistics.

In most cases, computer programs are readily available to perform these calculations. Several statistical programming packages allow weighted analyses.[2] For example, nearly all SAS (Statistical Analysis System) procedures incorporate weights. In addition, several statistical analysis programs have been developed that specifically calculate statistics and their standard errors from survey data. Version 8 of SAS contains procedures (PROC SURVEYMEANS and PROC SURVEYREG) for calculating statistics based on specific sampling designs. STATA and SUDAAN are two other common statistical software packages that do calculations for numerous statistics arising from the stratified, single-stage cluster sampling design. Examples of the use of SAS, SUDAAN and STATA to calculate variances in the NIS are presented in the special report: *Calculating Nationwide Inpatient Sample Variances, 2000*. This report is available on the 2000 NIS Documentation CD-ROM and on the HCUP Web site. For an excellent review of programs to calculate statistics from survey data, visit the following web site: http://www.fas.harvard.edu/~stats/survey-soft/.

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 (Primary Sampling Units or 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. In 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 cesarean sections would be correlated with their total number of deliveries. The number of cesarean 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 cesarean sections from deliveries based on the NIS data, that regression can then be used to obtain hospital-specific estimates of the number of cesarean sections for all hospitals in the universe.

Longitudinal Analyses

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 for every hospital. The numbers 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 numbers of hospitals for 1992-1999 are based on the AHA Annual Survey files.

[2] 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.