DESIGN OF THE NATIONWIDE INPATIENT SAMPLE (NIS), 2004

August 8, 2006

Healthcare Cost and Utilization Project
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850

TABLE OF CONTENTS

INDEX OF TABLES

INDEX OF FIGURES

EXECUTIVE SUMMARY

Introduction

The Nationwide Inpatient Sample (NIS) is one of a family of databases and software tools developed as part of the Healthcare Cost and Utilization Project (HCUP), a Federal-State-Industry partnership sponsored by the Agency for Healthcare Research and Quality (AHRQ). The NIS is the largest nationwide all-payer hospital inpatient care database in the U.S. Each year the NIS contains data from approximately seven to eight million hospital stays – all discharge records from approximately 1,000 hospitals selected from HCUP State Inpatient Databases (SID) data.

The HCUP NIS team developed the NIS to provide analyses of hospital utilization, charges, and quality of care across the United States. This report describes the NIS sample and weights, summarizes the contents of the 2004 NIS, and discusses data analysis issues. Previous NIS releases covered 1988 through 2003. This document highlights cumulative information for all previous years to provide a longitudinal view of the database. The 2004 NIS includes data from 37 states, the same number included in the 2003 NIS. Compared with 2003, one state was added (Arkansas) and one was dropped (Pennsylvania).

Hospital Sample Design

The NIS sampling frame included all community, non-rehabilitation hospitals in the SID that could be matched to the corresponding American Hospital Association (AHA) Annual Survey Database data. Based on data from 37 states, there were 3,705 hospitals in the 2004 sampling frame, a 1.5% decrease from the 2003 NIS. The target universe includes all acute care discharges from non-rehabilitation, community hospitals in the United States. In 2004, the target universe contained 4,906 hospitals.

The NIS is a stratified probability sample of hospitals in the frame, with sampling probabilities calculated to select 20% of the universe contained in each stratum. The overall objective was to select a sample of hospitals representative of the target universe. With this objective in mind, we defined NIS sampling strata based on the following 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.

After stratifying the universe of hospitals, we randomly selected up to 20% of the total number of U.S. hospitals within each stratum. If a stratum contained too few frame hospitals, then all were selected for the NIS, subject to sampling restrictions specified by states. The resulting sample for 2004 included 1,004 hospitals, representing 20.5% of the total hospital universe of 4,906 hospitals.

Changes to Sampling and Weighting Strategy

Given the increase in the number of contributing states, the NIS team evaluated and revised the sampling and weighting strategy for 1998 and subsequent data years in order to best represent the U.S. These changes included:

- Revising definitions of the strata variables
- Excluding rehabilitation hospitals from the NIS hospital universe
- Changing the calculation of hospital universe discharges for the weights.

Also, beginning with the 1998 NIS sampling procedures, all frame hospitals within a stratum have an equal probability of selection, regardless of whether they had appeared in prior NIS samples. This deviates from the procedure used for earlier samples, which maximized the longitudinal component of the NIS series. A full description of the evaluation and revision of the NIS sampling strategy for 1998 and subsequent data years can be found in the special report, *Changes in NIS Sampling and Weighting Strategy for 1998.* This document is available on the NIS Documentation CD-ROM and on the HCUP User Support Website at http://www.hcup-us.ahrq.gov/db/nation/nis/nisrelatedreports.jsp.

Beginning with the 2004 NIS, we changed the classification of urban or rural hospital location for the sampling strata to use the newer Core Based Statistical Area (CBSA) codes rather than the older Metropolitan Statistical Area (MSA) codes. The CBSA groups are based on 2000 Census data, whereas the MSA groups were based 1990 Census data. Also, the criteria for classifying the counties differ. For more information on the difference between CBSAs and MSAs, refer to the U.S. Census Bureau Website (http://www.census.gov/population/www/estimates/metroarea.html).

Previously, we classified hospitals in an MSA as urban hospitals, while we classified hospitals outside an MSA as rural hospitals. Beginning with the 2004 NIS, we classified hospitals with a CBSA type of Metropolitan or Division as urban, while we classified hospitals with a CBSA type of Micropolitan or Rural as rural. This change contributed to a slight decline in the number of hospitals that were classified as rural and a corresponding increase in the number of hospitals that were classified as urban. For the 2003 NIS, 44.9% of hospitals in the AHA Universe were classified as rural hospitals; while for 2004, only 41.3% of AHA Universe hospitals were classified as rural.

Hospital Sampling Frame

The 2004 NIS sampling frame included data provided by 37 HCUP State Partners. On average, 97% of the hospital universe is included in the sampling frame for all but six of these states. (Restrictions from other states did not have an appreciable effect on the percentage of hospitals in the sampling frame.) Three State Partners – Hawaii, South Carolina, and South Dakota – limited the number of state hospitals included in the frame to between 70 and 86 percent. Texas, supplied data from only 73% of the state's hospitals because some Texas hospitals, mostly small rural facilities, are exempt from statutory reporting requirements. Minnesota supplied data from only 89% of the state's hospitals to HCUP because a few Minnesota hospitals do not participate in the project. There are no apparent significant differences between the characteristics of participating and non-participating Minnesota hospitals. We omitted 45 Michigan hospitals that did not report total charges from the sampling frame, leaving 69% of Michigan hospitals in the frame.

While 20% of the hospitals in each region are selected for the NIS, the comprehensiveness of the sampling frame varies by region. In the Midwest, 89.7% of hospitals were included in the sampling frame, compared with 78.2% in the West, 67.5% in the South, and 64.8% in the Northeast. Because the NIS sampling frame has a disproportionate representation of the more populous states and includes hospitals with more annual discharges, its comprehensiveness in terms of discharges is higher. The states in the NIS sampling frame contained 99.0% of the population in the Midwest, 92.0% in the West, 84.1% in the South, and 74.9% in the Northeast. Overall, the 2004 NIS sampling frame comprised 75.5% of all U.S. hospitals and encompassed 87.5% of the U.S. population.

Final Hospital Sample

The final 2004 sample included 8,004,571 discharges from 1,004 hospitals selected from all 37 frame states. Hospitals were sampled throughout each region of the United States. Generally, in the Midwest and West, where a higher proportion of hospitals were represented, relatively fewer hospitals were sampled from each state than in the Northeast and South, where the proportion of hospitals in the NIS is lower. Since the inception of the original 1988 NIS, its scope has expanded across several dimensions:

- The number of states has increased from 8 to 37.
- The number of hospitals has increased from 759 to 1,004.
- The number of discharges has increased from 5.2 million to more than 8 million.

The 2004 NIS includes data from 37 states – 29 more states than the original 1988 NIS. With the loss of Pennsylvania from the NIS, the percentage of Northeast population represented in the NIS decreased from 98% in 2003 to 75% in 2004. However, with the addition of Arkansas, the percentage of Southern population represented in the NIS increased from 81% in 2003 to 84% for 2004. The percentage of the Western and Midwestern population represented in the NIS remained unchanged at 92 and 99 percent, respectively.

Ideally, relationships among outcomes and their correlates estimated from the NIS should accurately represent all U.S. hospitals. However, when creating nationwide estimates, it is advisable to check these estimates against other data sources, if available. For example, the National Hospital Discharge Survey (http://www.cdc.gov/nchs/about/major/hdasd/nhds.htm) can provide benchmarks against which to verify national estimates for hospitalizations with more than 5,000 cases.

The *NIS Comparison Report* assesses the accuracy of NIS estimates. The most recent report is available on the NIS Documentation CD-ROM and provides a comparison of a previous year's NIS with other data sources. The updated report for the current NIS will be posted on the HCUP User Support Website (http://www.hcup-us.ahrq.gov/db/nation/nis/nisrelatedreports.jsp) as soon as it is completed.

Subsamples

Two non-overlapping 10% subsamples of discharges were drawn from the NIS file for several data analysis purposes. 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.

Sample Weights

It is necessary to incorporate sample weights to obtain nationwide 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 (HOSPWT) are usually near five. The calculations for discharge-level sampling weights (DISCWT) are similar to the calculations for hospital-level sampling weights. In the 10% subsamples, each discharge has a 10% chance of being drawn. Therefore, the discharge weights (DISCWT10) are multiplied by 10 for each of the subsamples. Because the 10% subsamples are based on samples of discharges, each hospital is represented in the subsamples. Thus, no adjustment is required for the hospital weight when using the subsamples.

Weight Data Elements

To produce nationwide estimates, use the discharge weights to extrapolate sampled discharges in the Core file to the discharges from all U.S. community, non-rehabilitation hospitals. For the 2000 NIS, use DISCWT to create nationwide estimates for all analyses except those that involve total charges, and use DISCWTCHARGE to create nationwide estimates of total charges. For all other years of the NIS, DISCWTCHARGE is not required, and DISCWT (DISCWT_U prior to the 1998 NIS) should be used to create all estimates. For a 10% subsample file, use the corresponding subsample discharge weight, DISCWT10 (D10CWT_U prior to the 1998 NIS) or DISCWTCHARGE10.

Data Analysis

Missing Values

Missing data values can compromise the quality of estimates. If the outcome for discharges with missing values is different from the outcome for discharges with valid values, then sample estimates for that outcome will be biased and will not accurately represent the discharge population. Also, when estimating totals for non-negative variables with missing values, sums would tend to be underestimated because the cases with missing values would be omitted from the calculations. Several techniques are available to help overcome this bias. One strategy is to impute acceptable values to replace missing values. Another strategy is to use sample weight adjustments to compensate for missing values. Descriptions of such data preparation and adjustment are outside the scope of this report; however, it is recommended that researchers evaluate and adjust for missing data, if necessary.

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 sample without replacement may be used to calculate statistics and their variances in most applications.

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*. This report is available on the NIS Documentation CD-ROM and on the HCUP User Support Website at http://www.hcup-us.ahrq.gov/db/nation/nis/nisrelatedreports.jsp.

Longitudinal Analyses

All frame hospitals within a stratum have an equal probability of being selected for the sample, regardless of whether they have appeared 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 NIS hospitals 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.

Studying Trends

When studying trends over time using the NIS, be aware that the sampling frame for the NIS changes over time. Because more states have been added, estimates from earlier years of the NIS may be subject to more sampling bias than later years of the NIS. In order to facilitate analysis of trends using multiple years of NIS data, an alternate set of NIS discharge and hospital weights for the 1988-1997 HCUP NIS was developed. These alternative weights were calculated in the same way as the weights for the 1998 and later years of the NIS. The special report, *Using the HCUP Nationwide Inpatient Sample to Estimate Trends*, includes details regarding the alternate weights and other recommendations for trends analysis. Both the NIS Trends Report and the alternative weights are available on the HCUP User Support Website under Methods Series (http://www.hcup-us.ahrq.gov/reports/methods.jsp). The NIS Trends Report is also available on the NIS Documentation CD-ROM.

To ease the burden on researchers conducting analyses that span multiple years, NIS trends supplemental files (NIS-Trends) are available through the HCUP Central Distributor. The NIS-Trends annual files contain the alternative trend weights for data prior to 1997 in addition to renamed, recoded, and new data elements consistent with the later years of the NIS. More information on these files is available on the HCUP-US Website under NIS database documentation (http://www.hcup-us.ahrq.gov/db/nation/nis/nisdbdocumentation.jsp).

INTRODUCTION

The Nationwide Inpatient Sample (NIS) is one of a family of databases and software tools developed as part of the Healthcare Cost and Utilization Project (HCUP), a Federal-State-Industry partnership sponsored by the Agency for Healthcare Research and Quality (AHRQ). The NIS is the largest nationwide all-payer hospital inpatient care database in the U.S. Each year the NIS contains data from approximately seven to eight million hospital stays – all discharge records from approximately 1,000 hospitals selected from HCUP State Inpatient Databases (SID) data.

The HCUP NIS team developed the NIS to facilitate analyses of hospital utilization, charges, and quality of care across the United States. Potential research issues focus on both discharge- and hospital-level outcomes. Discharge outcomes of interest include trends in inpatient treatment with respect to:

- Frequency
- Charges
- Lengths of stay
- Effectiveness
- Quality of care
- Appropriateness
- Access to hospital care.

Hospital-level outcomes of interest include:

- Mortality rates
- Complication rates
- Patterns of care
- Diffusion of technology
- Trends toward specialization.

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

This report focuses on the NIS sample and weights, summarizes the contents of the 2004 NIS, and discusses data analysis issues. The 2004 NIS includes data for calendar year 2004, while previous NIS releases covered 1988 through 2003. This document highlights cumulative information for all previous years to provide a longitudinal view of the database.

Table 1 displays the number of states, hospitals, and discharges in each year and reveals the increase in the number of participating states over time.

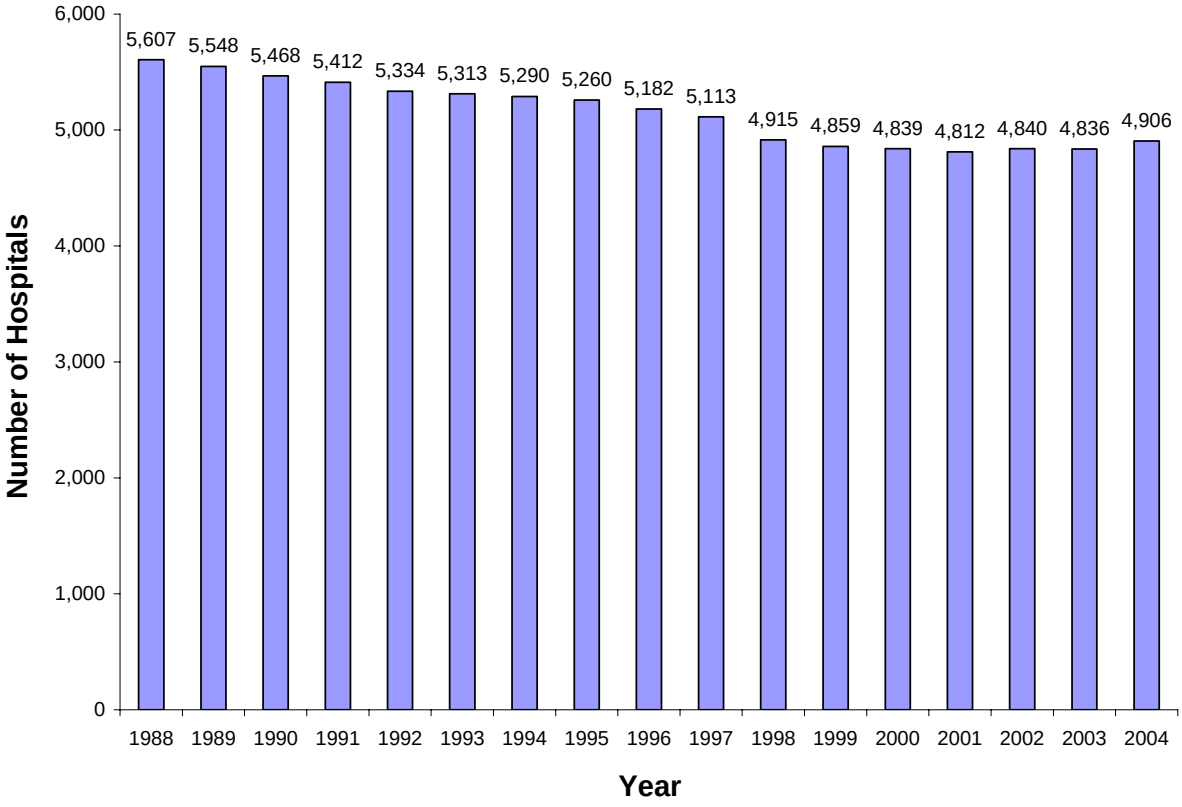
Table 1: Number of NIS States, Hospitals, and Discharges, by Year

Calendar Year	States in the Frame	Number of States	Sample Hospitals	Sample Discharges
1988	California, Colorado, Florida, Iowa, Illinois, Massachusetts, New Jersey, and Washington	8	758	5,265,756
1989	Added Arizona, Pennsylvania, and Wisconsin	11	875	6,110,064
1990	No new additions	11	861	6,268,515
1991	No new additions	11	847	6,156,188
1992	No new additions	11	838	6,195,744
1993	Added Connecticut, Kansas, Maryland, New York, Oregon, and South Carolina	17	913	6,538,976
1994	No new additions	17	904	6,385,011
1995	Added Missouri and Tennessee	19	938	6,714,935
1996	No new additions	19	906	6,542,069
1997	Added Georgia, Hawaii, and Utah	22	1012	7,148,420
1998	No new additions	22	984	6,827,350
1999	Added Maine and Virginia	24	984	7,198,929
2000	Added Kentucky, North Carolina, Texas, and West Virginia	28	994	7,450,992
2001	Added Michigan, Minnesota, Nebraska, Rhode Island, and Vermont	33	986	7,452,727
2002	Added Nevada, Ohio, and South Dakota; Dropped Arizona	35	995	7,853,982
2003	Added Arizona, Indiana, and New Hampshire; Dropped Maine	37	994	7,977,728
2004	Added Arkansas; Dropped Pennsylvania	37	1,004	8,004,571

THE NIS HOSPITAL UNIVERSE

The hospital universe is defined as all hospitals located in the U.S. that are open during any part of the calendar year and designated as community hospitals in the American Hospital Association (AHA) Annual Survey Database. The AHA defines community hospitals as follows: "All nonfederal short-term general and other specialty hospitals, excluding hospital units of institutions." Consequently, Veterans Hospitals and other Federal facilities (Department of Defense and Indian Health Service) are excluded. Beginning with the 1998 NIS, we excluded short-term rehabilitation hospitals from the universe because the type of care provided and the characteristics of the discharges from these facilities were markedly different from other short-term hospitals. Figure 1 displays the number of universe hospitals for each year based on the AHA Annual Survey. Between the years 1988-2001, a steady decline in the number of hospitals is evident. However, beginning in 2002, the number of universe hospitals has stabilized.

Figure 1: Hospital Universe, by Year[1]



Hospital Merges, Splits, and Closures

All U.S. hospital entities designated as community hospitals in the AHA hospital file, except short-term rehabilitation hospitals, were included in the hospital universe. Therefore, when 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 during the year they merged. Similarly, if a community hospital split, the original hospital and all newly-created community hospitals were treated as separate entities in the universe during the year this occurred. Finally, community hospitals that closed during a given year were included in the hospital universe, as long as they were in operation during some part of the calendar year.

Stratification Variables

Given the increase in the number of contributing states, the NIS team evaluated and revised the sampling and weighting strategy for 1998 and subsequent data years, in order to best represent the U.S. This included changes to the definitions of the strata variables, the exclusion of rehabilitation hospitals from the NIS hospital universe, and a change to the calculation of hospital universe discharges for the weights. 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 NIS Documentation CD-ROM and on the HCUP User Support Website at http://www.hcup-us.ahrq.gov/db/nation/nis/nisrelatedreports.jsp. (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, 1997.* This report is available on the 1997 NIS Documentation CD-ROM and on the HCUP User Support Website.)

The NIS sampling strata were defined based on five hospital characteristics contained in the AHA hospital files. 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. Figure 2 highlights the NIS states in gray, and Table 2 lists the states that comprise each region.

2. *Control – government non-Federal (public), private not-for-profit (voluntary), and private investor-owned (proprietary)*. Depending on their control, 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, we stratified hospitals as public, voluntary, and proprietary. We used this stratification for Southern rural, Southern urban non-teaching, and Western urban non-teaching hospitals. For smaller strata – the Midwestern rural and Western rural hospitals – we used a collapsed stratification of public versus private, 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 non-teaching*. The missions of teaching hospitals differ from non-teaching hospitals. In addition, financial considerations differ between these two hospital groups. Currently, the Medicare Diagnosis Related Group (DRG) payments are uniformly higher to teaching hospitals. We considered a hospital to be a teaching hospital if it has residency training approval by the Accreditation Council for Graduate Medical Education (ACGME), 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.[2]

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

Beginning with the 2004 NIS, we changed the classification of urban or rural hospital location for the sampling strata to use the newer Core Based Statistical Area (CBSA) codes rather than the older Metropolitan Statistical Area (MSA) codes. The CBSA groups are based on 2000 Census data, whereas the MSA groups were based 1990 Census data. Also, the criteria for classifying the counties differ. For more information on the difference between CBSAs and MSAs, refer to the U.S. Census Bureau Website (http://www.census.gov/population/www/estimates/metroarea.html).

Previously, we classified hospitals in an MSA as urban hospitals, while we classified hospitals outside an MSA as rural hospitals. Beginning with the 2004 NIS, we classified hospitals with a CBSA type of Metropolitan or Division as urban, while we classified hospitals with a CBSA type of Micropolitan or Rural as rural. This change contributed to a slight decline in the number of hospitals that were classified as rural and a corresponding increase in the number of hospitals that were classified as urban. For the 2003 NIS, 44.9% of hospitals in the AHA Universe were classified as rural hospitals; while for 2004, only 41.3% of AHA Universe hospitals were classified as rural.

We chose the bed size cutoff points so that approximately one-third of the hospitals in a given region, location, and teaching status combination would fall within each bed size category (small, medium or large). We used different cutoff points for rural, urban non-teaching, and urban teaching hospitals 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 containing small numbers of hospitals.

We did not split rural hospitals according to teaching status, because rural teaching hospitals were rare. For example, in 2004, rural teaching hospitals comprised less than one percent of the total hospital universe. We defined the bed size categories 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 accurate geographic representation, implicit stratification variables included state and three-digit ZIP Code (the first three digits of the hospital's five-digit ZIP Code). Within each stratum, we sorted hospitals by three-digit ZIP Code prior to systematic random sampling.

Figure 2: NIS States, by Region

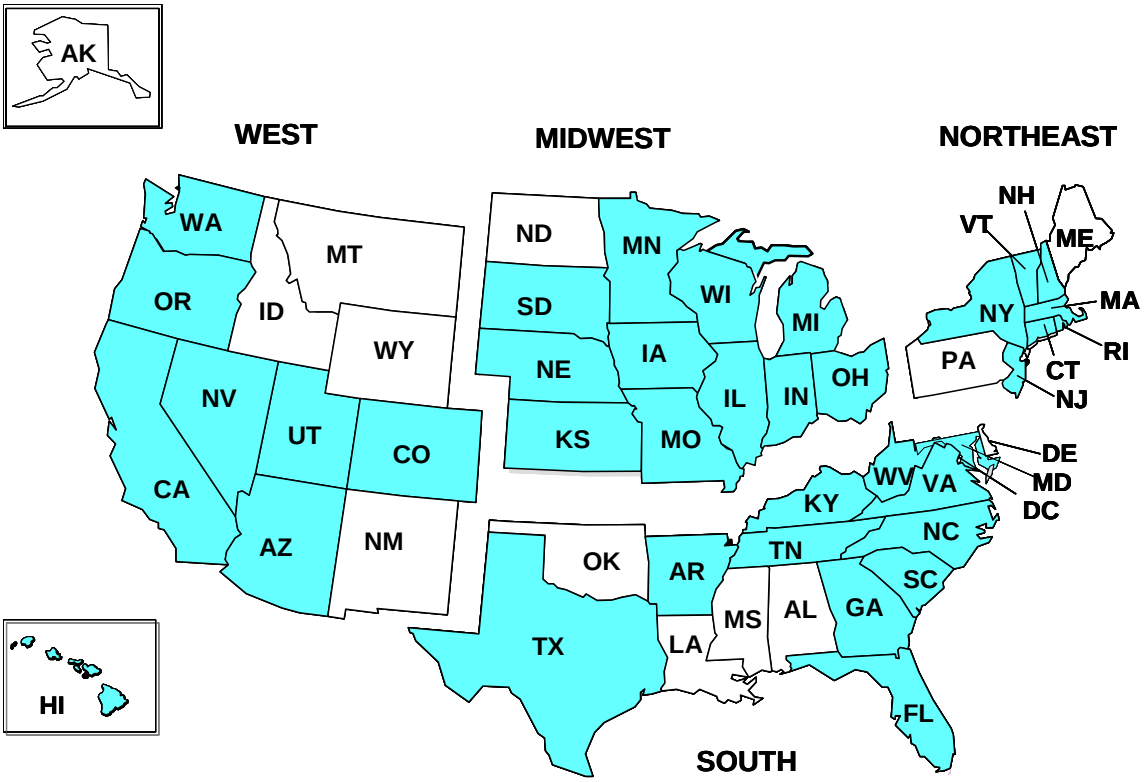


Table 2: All States, by Region

Region	States
1: Northeast	Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont.
2: Midwest	Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin.
3: South	Alabama, Arkansas, Delaware, District of Columbia, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, West Virginia.
4: West	Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, Wyoming.

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, non-teaching	1-124	125-199	200+
Urban, teaching	1-249	250-424	425+
MIDWEST			
Rural	1-29	30-49	50+
Urban, non-teaching	1-74	75-174	175+
Urban, teaching	1-249	250-374	375+
SOUTH			
Rural	1-39	40-74	75+
Urban, non-teaching	1-99	100-199	200+
Urban, teaching	1-249	250-449	450+
WEST			
Rural	1-24	25-44	45+
Urban, non-teaching	1-99	100-174	175+
Urban, teaching	1-199	200-324	325+

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 short-term rehabilitation hospitals. However, some hospitals do not supply data to HCUP. Therefore, we constructed the NIS *sampling frame* from the subset of universe hospitals that released their discharge data to AHRQ for research use. When the 2004 sample was drawn, AHRQ had agreements with 37 HCUP State Partner organizations to include their data in the NIS. The number of State Partners contributing data to the NIS has expanded over the years, as shown in Table 1. As a result, the number of hospitals included in the NIS sampling frame has also increased over the years, as displayed in Figure 3.

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).

Figure 3: NIS Hospital Sampling Frame, by Year

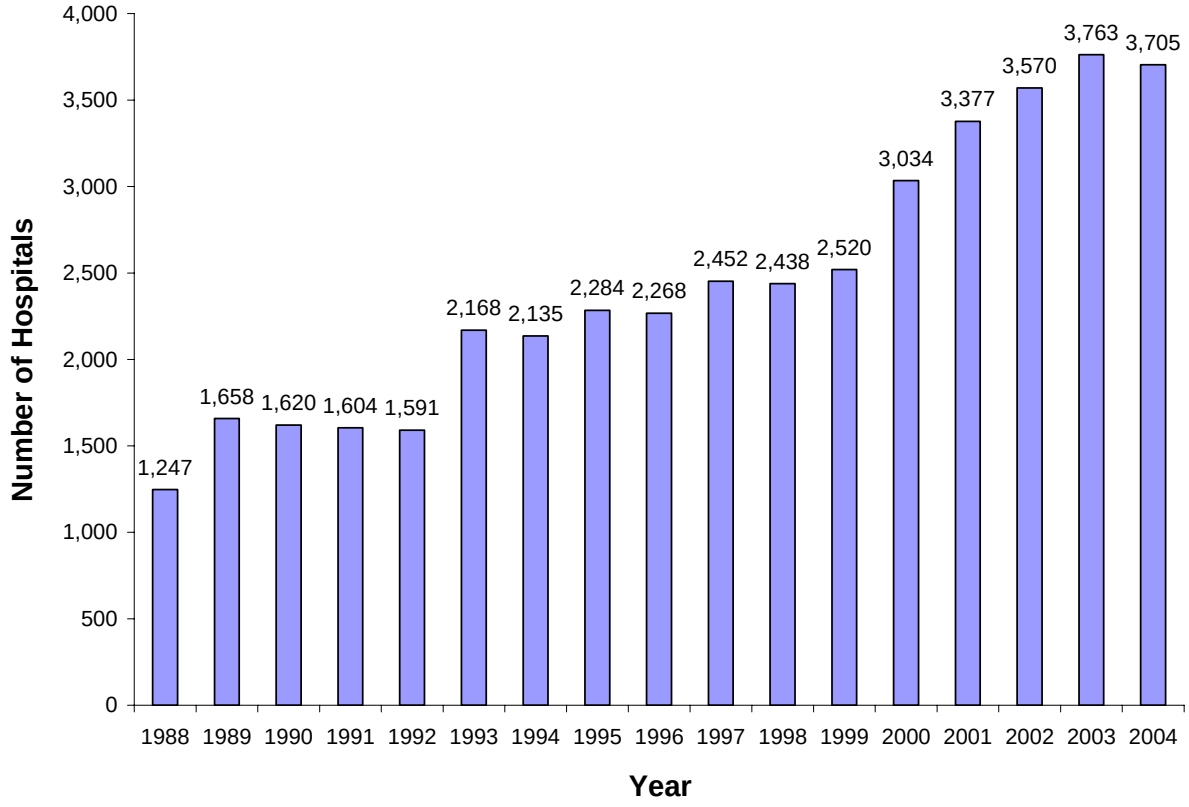


Figure 4 illustrates the number of hospitals in the universe, frame, and sample and the percentage of universe hospitals in the frame for each state in the sampling frame for 2004. In most cases, the difference between the universe and the frame represents the difference in the number of community, non-rehabilitation hospitals in the 2004 AHA Annual Survey of Hospitals and the hospitals for which data were supplied to HCUP that could be matched to the AHA data.

The largest discrepancy between HCUP data and AHA data is in Texas. As is evident in Figure 4, only 323 out of 442 Texas community, non-rehabilitation hospitals supplied data to HCUP for 2004. Certain Texas state-licensed hospitals are exempt from statutory reporting requirements. Exempt hospitals include:

- Hospitals that do not seek insurance payment or government reimbursement
- Rural providers.

The Texas statute that exempts rural providers from the requirement 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 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 smaller rural public hospitals and, as a result, these facilities are less likely to be included in the sampling frame than other Texas hospitals. While the number of hospitals omitted appears sizable, those available for the NIS include 92.1% of inpatient discharges from Texas universe hospitals because excluded hospitals tended to have relatively few discharges.

The Minnesota frame contains 14 fewer hospitals than the state universe because a few of the state's hospitals do not participate in HCUP. There are no apparent significant differences between the characteristics of participating and non-participating Minnesota hospitals.

The Ohio frame contains 13 fewer hospitals than the state universe, including three hospitals that could not be matched to the AHA data because the Partner masked their identities in the data.

For Hawaii, Indiana, Michigan, Nebraska, South Carolina, and South Dakota, we had to exclude several HCUP hospitals from the frame, as described below:

- The Hawaii frame contains seven fewer hospitals than the state universe. Three hospitals were excluded because of sampling restrictions stipulated by the State Partner, and four hospitals identified in AHA data were not included in the data supplied to HCUP.

- Similarly, the Indiana frame contains seven fewer hospitals than the state universe. Two hospitals were excluded because of sampling restrictions stipulated by the State Partner, and five hospitals identified in AHA data were not included in the data supplied to HCUP.

- The Michigan frame contains 45 fewer hospitals than the state universe. Because charges represent a critical outcome variable in the NIS, we decided to omit 33 hospitals from the frame that did not provide total charges. By excluding these hospitals, we avoid having to adjust the weights or create another weighting variable specifically for total charges. These hospitals are fairly evenly distributed by hospital type. There are no sampling strata in the state containing only hospitals without charges. The total charge data reported for Michigan is similar to total charge data reported by other Midwestern states. Thus, there does not seem to be an obvious bias in the type of cases for which charges are reported. The stratification and weighting scheme should adjust for the hospitals that are being excluded. In addition, one hospital was excluded because of sampling restrictions stipulated by Michigan, and 12 hospitals identified in AHA data were not included in the data supplied to HCUP.

- The Nebraska frame contains three fewer hospitals than the state universe. We dropped two hospitals from the sampling frame because they had incomplete data and were missing a high percentage of Medicare Discharges. In addition, one hospital identified in AHA data was not included in the data supplied to HCUP.

- The South Carolina frame contains eight fewer hospitals than the state universe. Five hospitals were excluded because of sampling restrictions stipulated by South Carolina, and three hospitals identified in AHA data were not included in the data supplied to HCUP.

- Likewise, the South Dakota frame contains six fewer hospitals than the South Dakota universe. Three hospitals were excluded because of sampling restrictions stipulated by South Dakota, while three hospitals identified in AHA data were not included in the data supplied to HCUP.

Figure 4: Number of Hospitals in the 2004 Universe, Frame, and Sample for Frame States

Part A: Arkansas – North Carolina

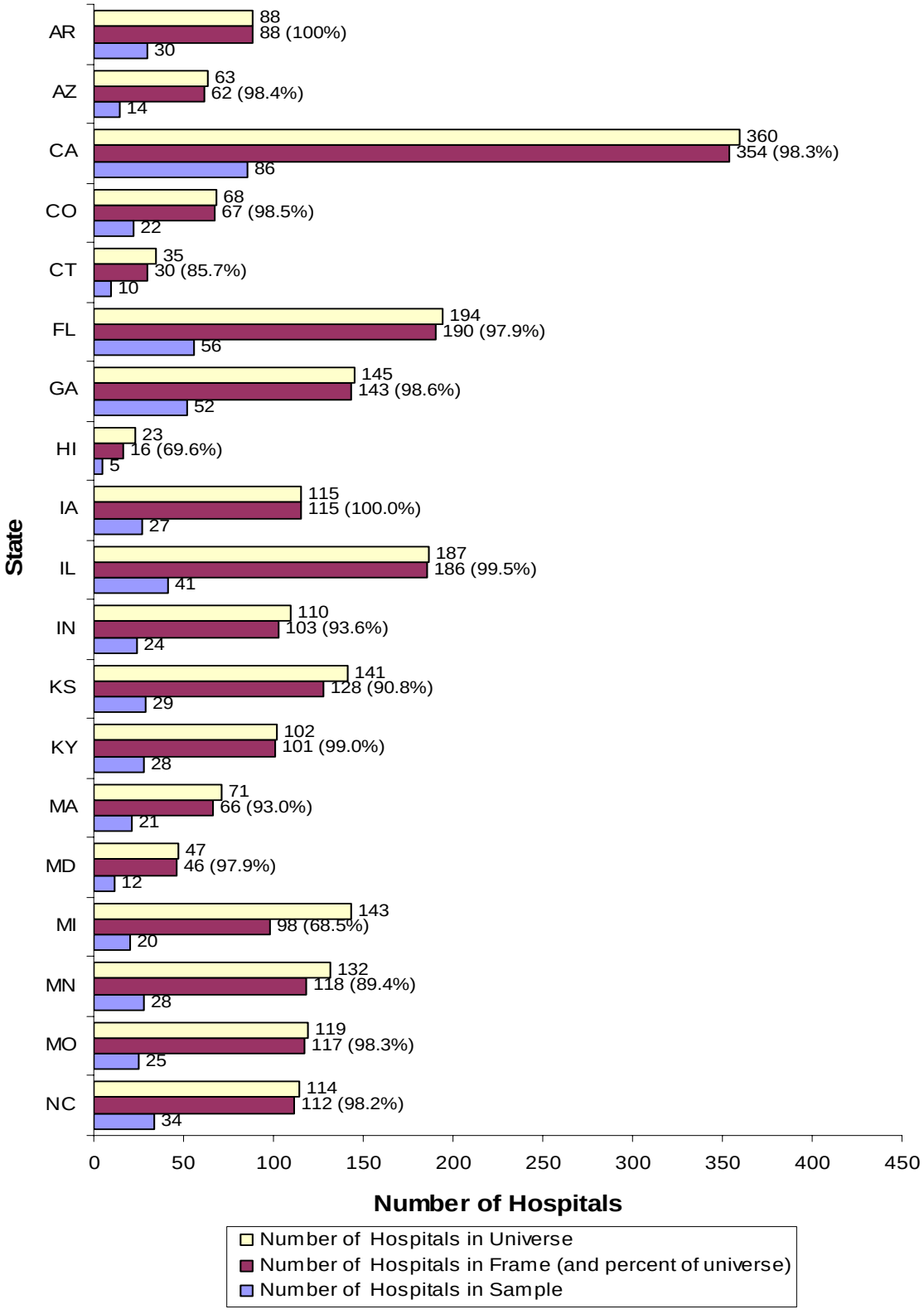
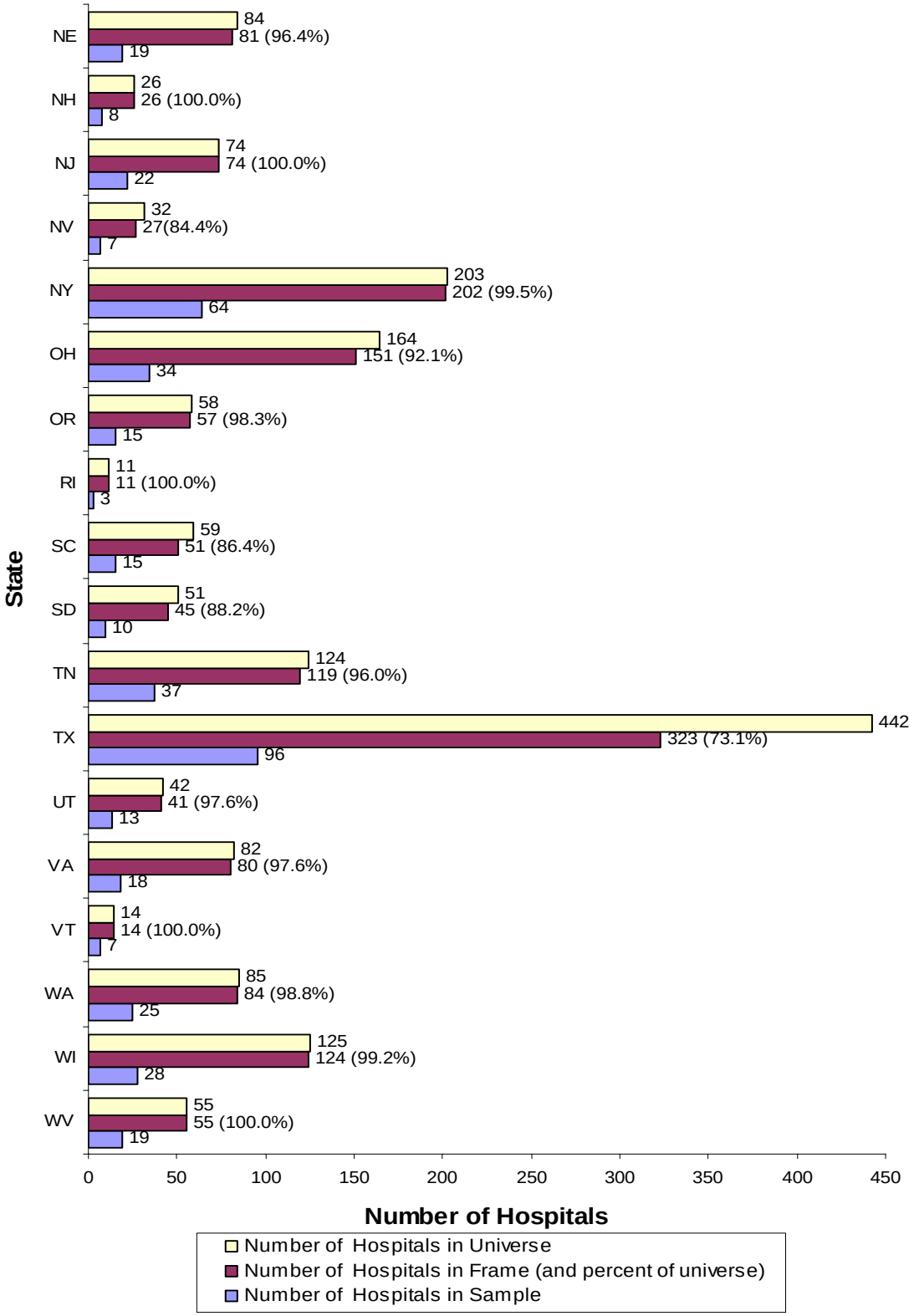


Figure 4: Number of Hospitals in the 2004 Universe, Frame, and Sample for Frame States

Part B: Nebraska – West Virginia



HOSPITAL SAMPLE DESIGN

Design Considerations

The NIS is a stratified probability sample of hospitals in the frame, with sampling probabilities calculated to select 20% of the universe of U.S. community, non-rehabilitation hospitals contained in each stratum. This sample size was determined by AHRQ based on their experience with similar research databases. The overall design objective was to select a sample of hospitals that accurately represents 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 only from data supplied by HCUP Partners.

It should be possible, for example, to estimate DRG-specific average lengths of stay across all U.S. hospitals using weighted average lengths of stay, based on averages or regression coefficients calculated from the NIS. Ideally, relationships among outcomes and their correlates estimated from the NIS should accurately represent all U.S. hospitals. However, the 2004 NIS includes data from only 37 states. Therefore, it is advisable to verify your estimates against other data sources, if available. For example, the National Hospital Discharge Survey (http://www.cdc.gov/nchs/about/major/hdasd/nhds.htm) can provide benchmarks against which to check your national estimates for hospitalizations with more than 5,000 cases.
The *NIS Comparison Report* assesses the accuracy of NIS estimates. The most recent report is available on the NIS Documentation CD-ROM and provides a comparison of a previous year's NIS with other data sources. The updated report for the current NIS will be posted on the HCUP User Support Website (http://www.hcup-us.ahrq.gov/db/nation/nis/nisrelatedreports.jsp) as soon as it is completed.

The NIS team considered alternative stratified sampling allocation schemes. However, allocation proportional to the number of hospitals was preferred for several reasons:

- AHRQ researchers wanted a simple, easily understood sampling methodology. The concept that the NIS sample could represent a "miniaturization" of the hospital universe was appealing. There were, however, obvious geographic limitations imposed by data availability.

- AHRQ statisticians considered other optimal allocation schemes, including sampling hospitals with probabilities proportional to size (number of discharges). They ultimately concluded that sampling with probability proportional to the number of hospitals was preferable. While this approach was admittedly less efficient, the extremely large sample sizes yield good estimates. Furthermore, because the data are to be used for purposes other than producing nationwide estimates, (e.g., regression modeling), it is critical that all hospital types, including small hospitals, are adequately represented.

Overview of the Sampling Procedure

After stratifying the universe of hospitals, we randomly selected up to 20% of the total number of U.S. hospitals within each stratum. If too few frame hospitals appeared in a cell, we selected all frame hospitals for the NIS, subject to sampling restrictions specified by states. To simplify variance calculations, we drew at least two hospitals from each stratum. If fewer than two frame hospitals were available in a stratum, we merged it with an "adjacent" cell containing hospitals with similar characteristics.

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

Generally, three-digit ZIP Codes that are proximal in value are geographically near one another 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.

Subsamples

We drew two non-overlapping 10% subsamples of discharges 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% 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% 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 appeared 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, 1997*. This report is available on the 1997 NIS Documentation CD-ROM and on the HCUP User Support Website at http://www.hcup-us.ahrq.gov/db/nation/nis/nisrelatedreports.jsp. 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 NIS Documentation CD-ROM and on the HCUP User Support Website.

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

Figure 5 depicts the numbers of hospitals sampled each year, and Figure 6 presents the numbers of discharges in each year of the NIS. For the 1988-1992 NIS, zero-weight hospitals were maintained to provide a longitudinal sample. Therefore, two figures exist for each of these years: one number for the regular NIS sample and another number for the total sample.

Figure 7 displays the weighted number of discharges sampled each year. Note that this number decreased from 35,408,207 in 1997 to 34,874,001 in 1998, a difference of 534,206 (1.5%). This slight decline is associated with two changes to the 1998 NIS design: the exclusion of community, rehabilitation hospitals from the hospital universe, and a change to the calculation of hospital universe discharges for the weights. Prior to 1998, we calculated discharges as the sum of total facility admissions (AHA data element ADMTOT), which includes long-term-care admissions, plus births (AHA data element BIRTHS) reported for each U.S. community hospital in the AHA Annual Survey.

Beginning in 1998, we calculate discharges as the sum of hospital admissions (AHA data element ADMH) plus births for each U.S. community, non-rehabilitation hospital. This number is more consistent with the number of discharges we receive from the state data sources. We also substitute total facility admissions, if the number of hospital admissions is missing. Without these changes, the weighted number of discharges for 1998 would have been 35,622,743. The exclusion of community, rehabilitation hospitals reduced the number of universe hospitals by 177 and the number of weighted discharges by 214,490. The change in the calculation of discharges reduced the weighted number of discharges by 534,252.

Figure 5: Number of Hospitals Sampled, by Year

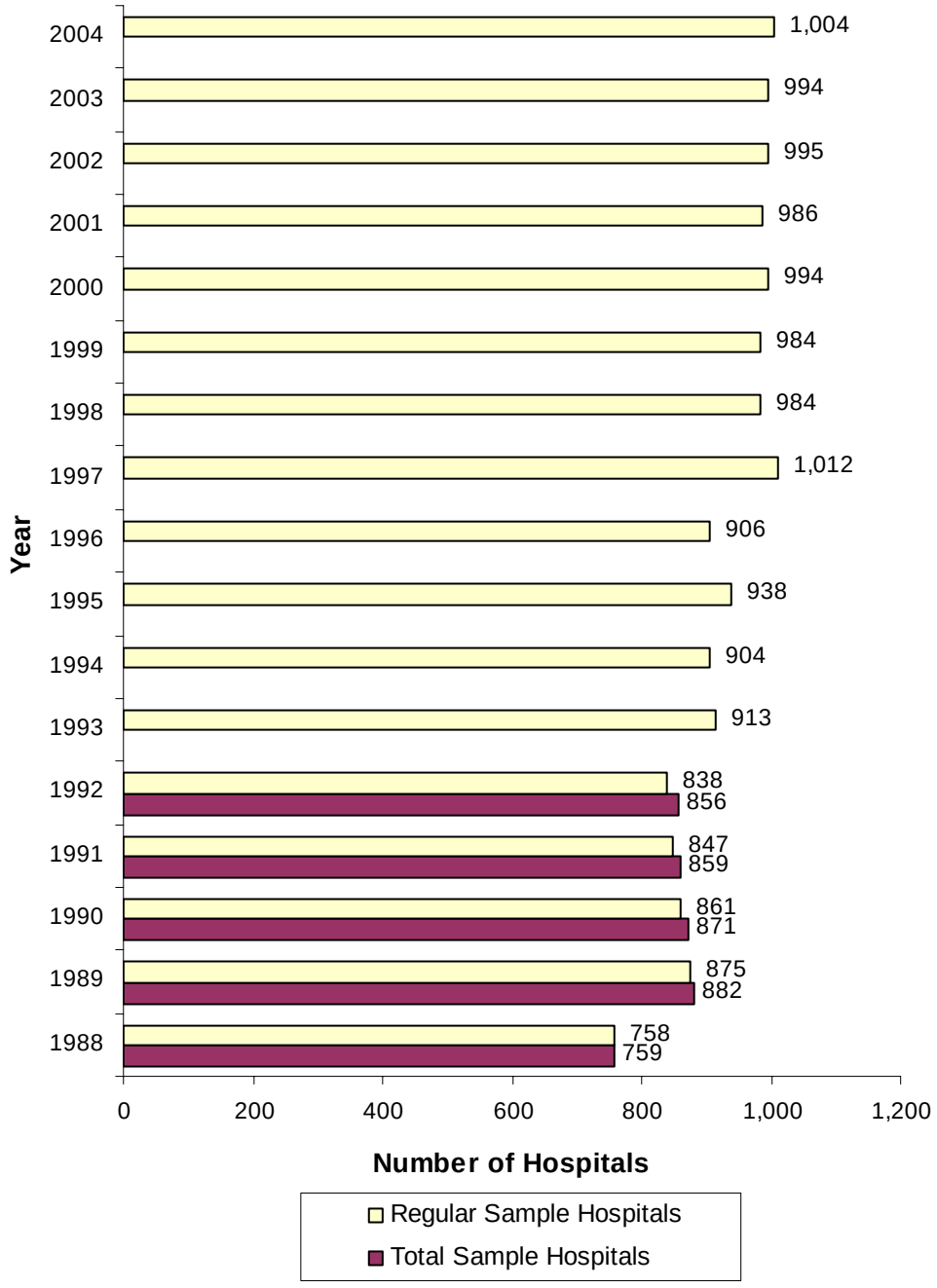


Figure 6: Number of NIS Discharges, Unweighted, by Year

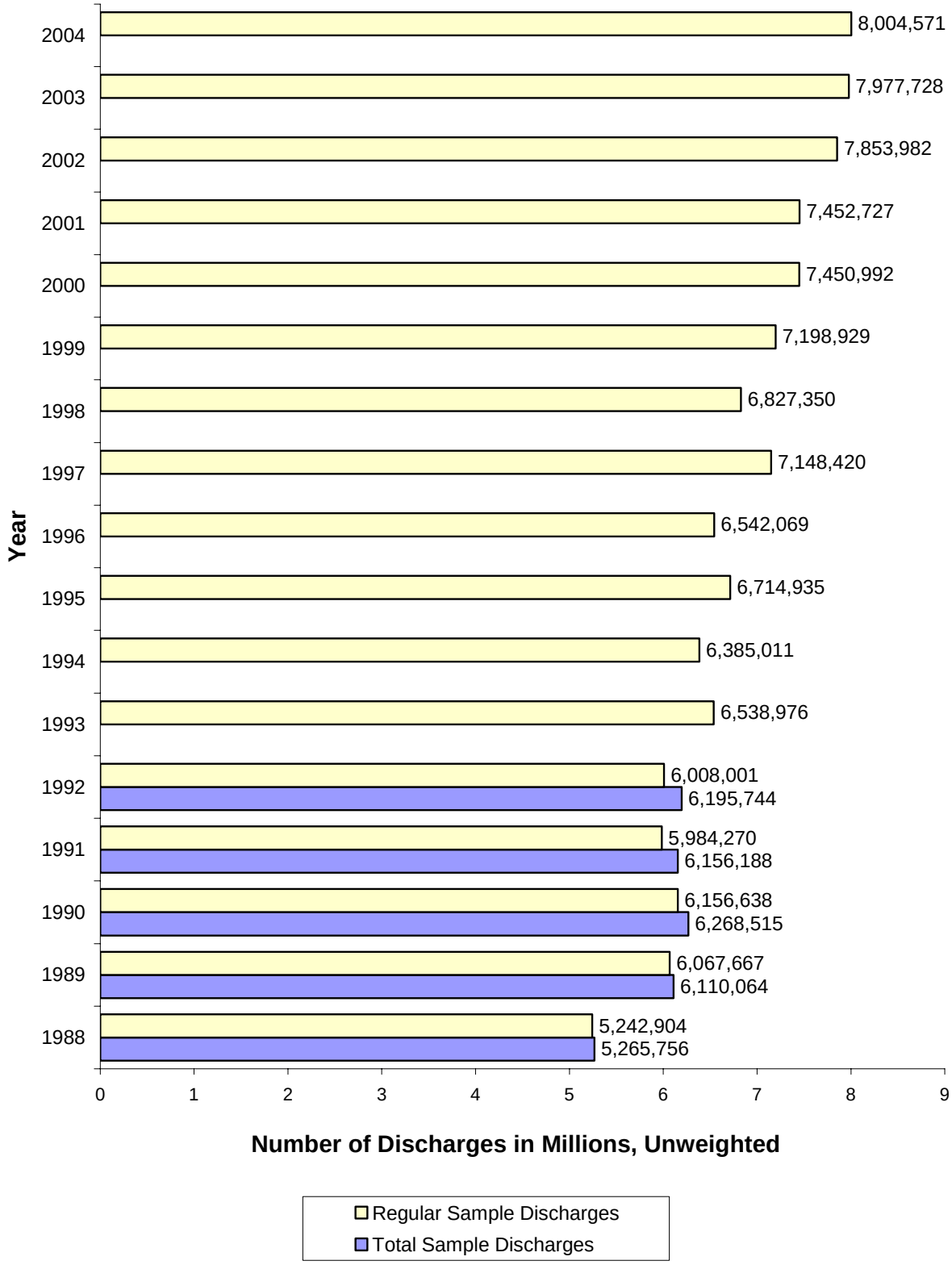


Figure 7: Number of NIS Discharges, Weighted, by Year

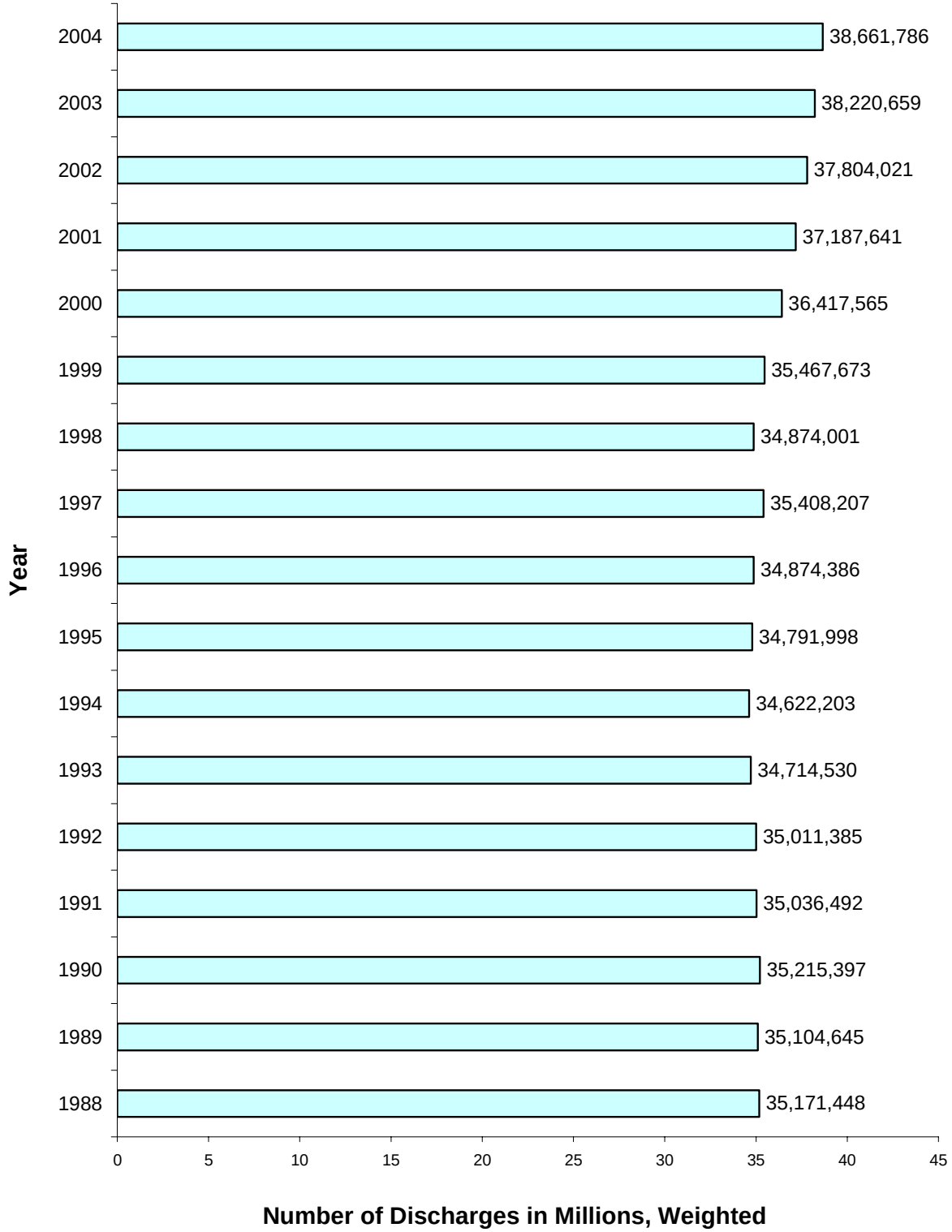


Figure 8 presents a summary of the 2004 NIS hospital sample by geographic region and the number of:

- Universe hospitals (Universe)
- Frame hospitals (Frame)
- Sampled hospitals (Sample)
- Target hospitals (Target = 20% of the universe)
- Surplus hospitals (Surplus = Sample – Target).

For example, in 2004, the Northeast region contained 653 hospitals in the universe. It also included 423 hospitals in the frame, of which 135 were drawn for the sample. This was four more than the target sample size of 131 hospitals, resulting in a surplus. The total sample exceeded the target by 23 hospitals, with a resulting sample of 20.5% of the total hospital universe. We sampled more than the target number of hospitals in each region because we rounded the target sample size for each stratum up to the next highest integer whenever it was not an integer.

Figure 9 summarizes the estimated U.S. population by geographic region on July 1, 2004.[3] For each region, the figure reveals:

- The estimated U.S. population
- The estimated population of states in the 2004 NIS
- The percentage of estimated U.S. population included in NIS states.

For example, the estimated population of the Midwest region on July 1, 2004 was 65,693,747. On that same date, the estimated population of states in the Midwest region that were included in the 2004 NIS was 65,057,439. This represents 99.0% of the total Midwest region's population. The percentage of estimated U.S. population included in states in the 2004 NIS was lower in the West (92.0%), South (84.1%), and Northeast (74.9%). As a result of the loss of Pennsylvania from the NIS, the percentage of Northeast population represented decreased from 97.6 to 74.9 percent. However, with the addition of Arkansas, the Southern population represented in the NIS grew from 81.3% in 2003 to 84.1% in 2004 – an increase of 2.8 percentage points. Overall, the states in the 2004 NIS included an estimated 87.5% of the entire U.S population, representing a decrease of 3.3 percentage points from 2003.

Figure 10 depicts the number of discharges in the 2004 sample for each state. The number of discharges sampled ranged from 31,194 in Hawaii to 829,399 in California.

Figure 8: Number of Hospitals in 2004 Universe, Frame, Sample, Target, and Surplus, by Region

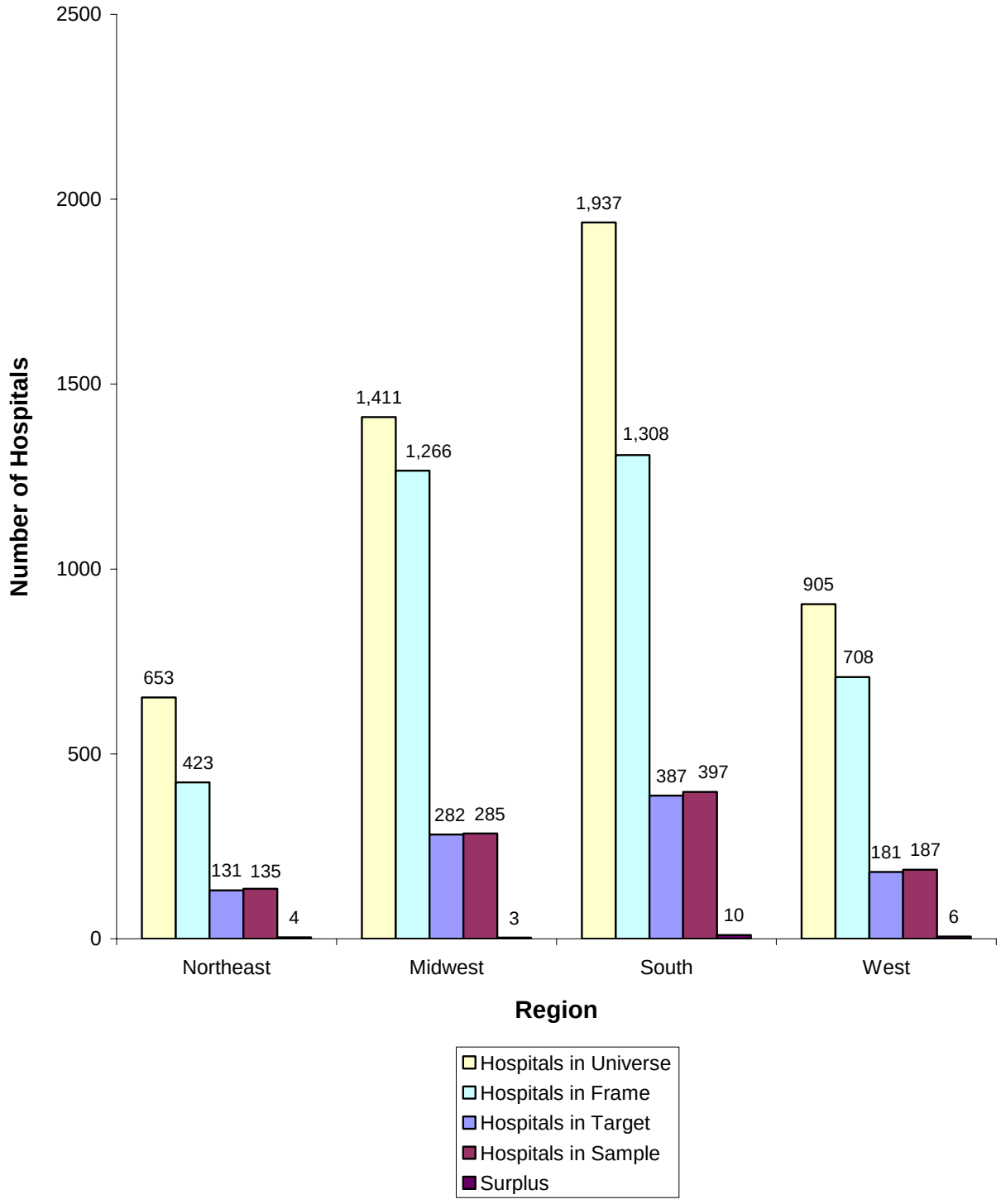


Figure 9: Percentage of U.S. Population in 2004 NIS States, by Region

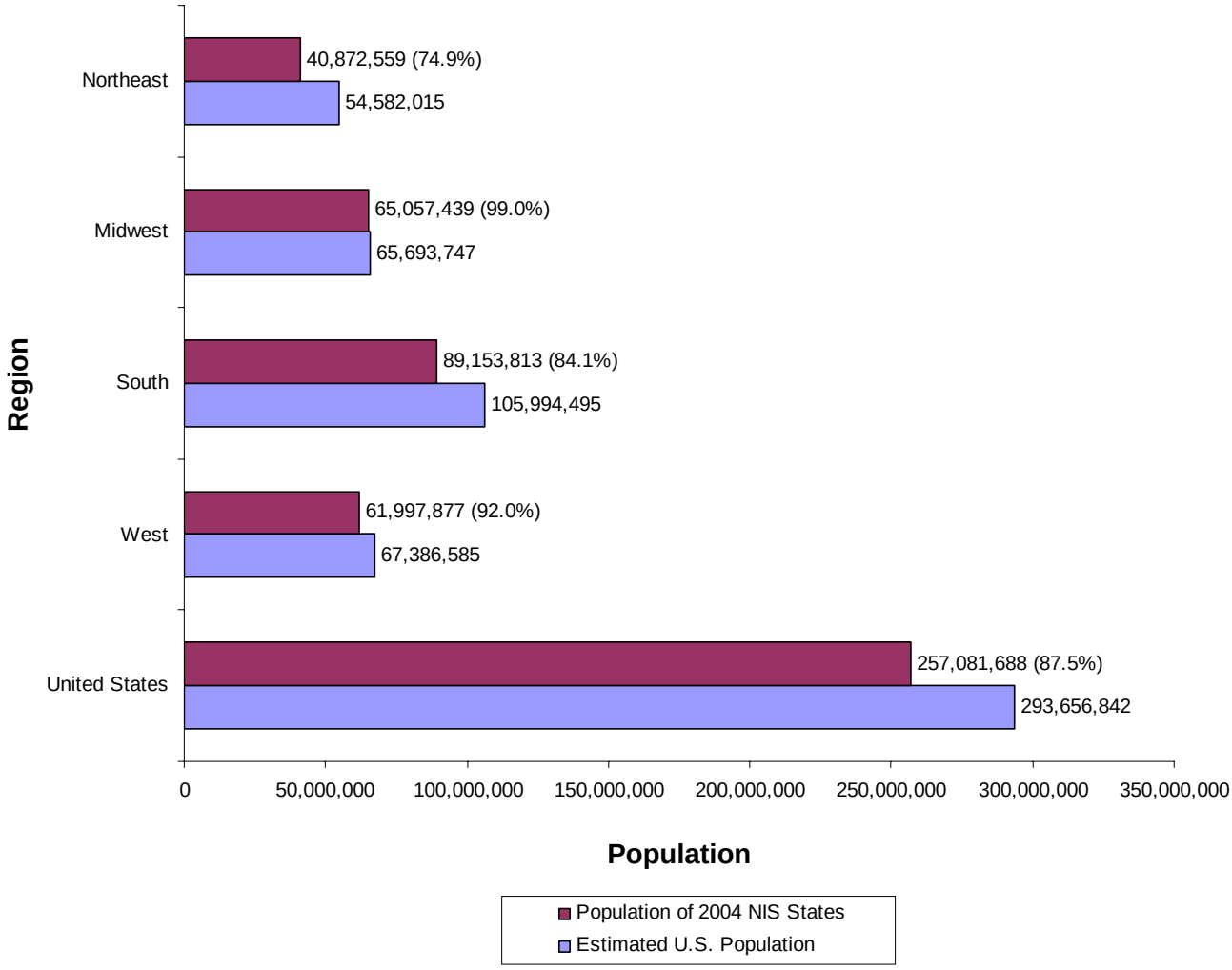
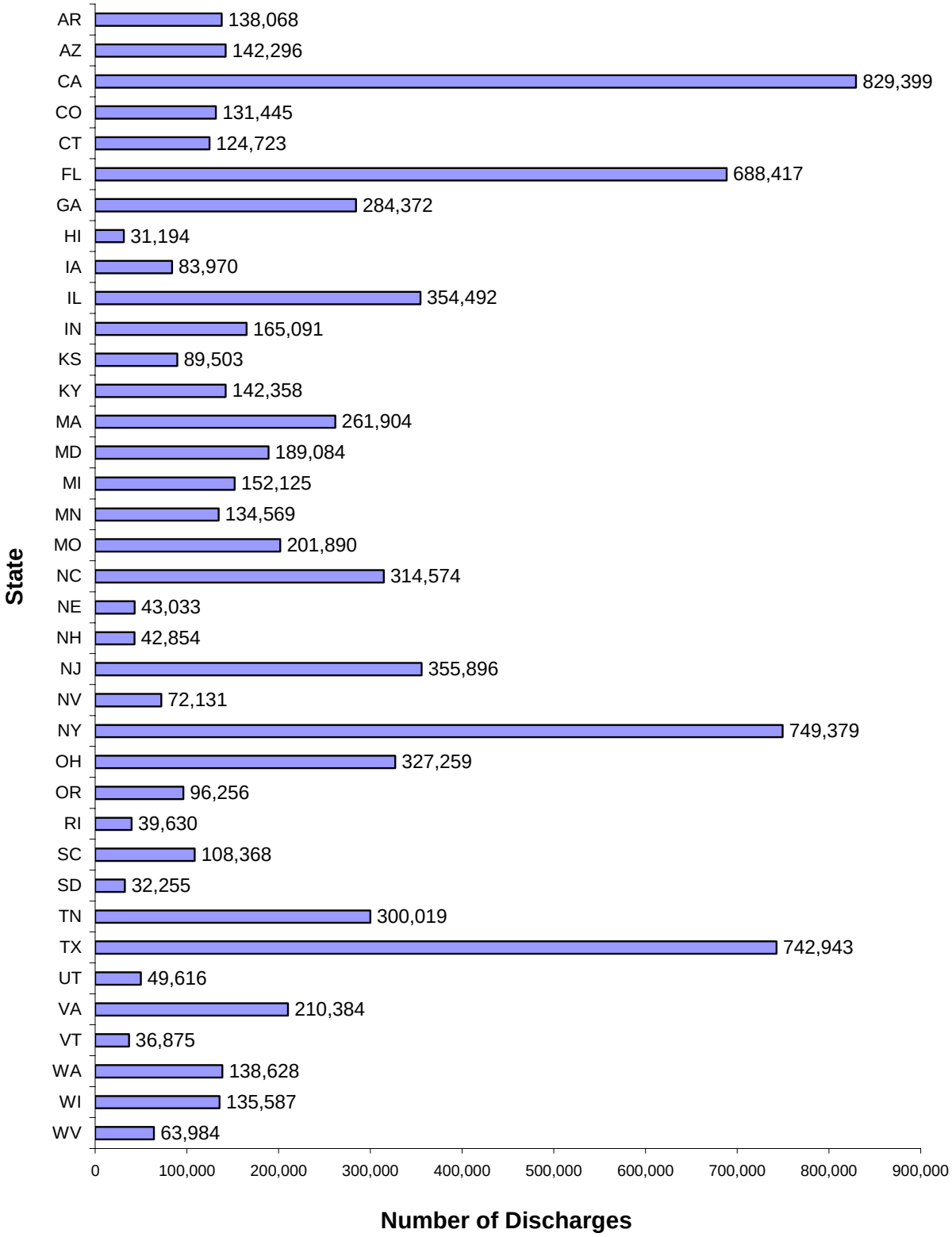


Figure 10: Number of Discharges in the 2004 NIS, by State



SAMPLE WEIGHTS

To obtain nationwide estimates, we developed discharge weights using the AHA universe as the standard. These were developed separately for hospital- and discharge-level analyses. Hospital-level weights were developed to extrapolate NIS sample hospitals to the hospital universe. Similarly, discharge-level weights were developed to extrapolate NIS sample discharges to the discharge 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 each 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, and 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 represents during that year. Because 20% of the hospitals in each stratum were sampled when possible, the hospital weights are usually near five.

Discharge Weights

The calculations for discharge-level sampling weights were similar to the calculations for hospital-level sampling weights. The discharge weights are usually constant for all discharges within a stratum. The only exceptions are for strata with sample hospitals that, according to the AHA files, were open for the entire year but contributed less than a 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 performed only for weighting purposes. The NIS data set includes only 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 represents 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) represented 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 represented 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 represents in stratum *s* during that year. Because all discharges from 20% of the hospitals in each stratum were sampled when possible, the discharge weights are usually near five.

Weight Data Elements

To produce nationwide estimates, use one of the following discharge weights to extrapolate discharges in the NIS Core file to the discharges from all U.S. community, non-rehabilitation hospitals. When using one of the 10% subsample files, use the subsample discharge weight (the discharge weight multiplied by 10). When using the hospital weights with the subsample files, there is no need to multiply the hospital weights, because all hospitals will be represented in the subsample files. Thus, the same hospital weight (HOSPWT) can be used for the full NIS and for the subsample files.

NIS Year	Name of Discharge Weight on the Core File to Use for Creating Nationwide Estimates	Name of Discharge Weight on the 10% Subsample File to Use for Creating Nationwide Estimates
2001-2004	• DISCWT for all analyses.	• DISCWT10 for all analyses.
2000	• DISCWT to create nationwide estimates for all analyses except those that involve total charges. • DISCWTCHARGE to create nationwide estimates of total charges.	• DISCWT10 to create nationwide estimates for all analyses except those that involve total charges. • DISCWTCHARGE10 to create nationwide estimates of total charges.
1998-1999	• DISCWT for all analyses.	• DISCWT10 for all analyses.
1988-1997	• DISCWT_U for all analyses.	• D10CWT_U for all analyses.

DATA ANALYSIS

Missing Values

Missing data values can compromise the quality of estimates. If the outcome for discharges with missing values is different from the outcome for discharges with valid values, then sample estimates for that outcome will be biased and inaccurately represent the discharge population. There are several techniques available to help overcome this bias. One strategy is to use imputation to replace missing values with acceptable values. Another strategy is to use sample weight adjustments to compensate for missing values.[4] Descriptions of such data preparation and adjustment are outside the scope of this report; however, it is recommended that researchers evaluate and adjust for missing data, if necessary.

On the other hand, if the cases with and without missing values are assumed to be similar with respect to their outcomes, no adjustment may be necessary for estimates of means and rates.

This is because the non-missing cases would be representative of the missing cases. However, some adjustment may still be necessary for the estimates of totals. Sums of data elements (such as aggregate charges) containing missing values would be incomplete because cases with missing values would be omitted from the calculations.

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 consisted of 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 are 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 sample 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 and interrelationships among hospitals and discharges during a specific year should be governed by finite-sample theory. Examples of this would be estimates of expenditure and utilization patterns or hospital market factors.

Alternatively, in the study of hypothetical population outcomes not limited to a specific point in time, the concept of a “superpopulation” may be useful. 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. This is the case because the population is defined at that point in time, and because the estimate is for a characteristic as it existed when sampled. 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 are useful for producing hospital-level statistics for analyses that use the *hospital* as the unit of analysis, while the discharge weights are useful for producing discharge-level statistics for analyses that use the *discharge* as the unit of analysis. The discharge weights may be used to estimate nationwide population statistics.

In most cases, computer programs are readily available to perform these calculations. Several statistical programming packages allow weighted analyses.[5] For example, nearly all Statistical Analysis System (SAS) procedures incorporate weights. In addition, several statistical analysis programs have been developed to specifically calculate statistics and their standard errors from survey data. Version eight or later 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 perform calculations for numerous statistics arising from the stratified, single-stage cluster sampling design. Examples of the use of SAS, SUDAAN, and STATA to calculate NIS variances are presented in the special report, *Calculating Nationwide Inpatient Sample Variances*. This report is available on the NIS Documentation CD-ROM and on the HCUP User Support Website at http://www.hcup-us.ahrq.gov/db/nation/nis/nisrelatedreports.jsp. For an excellent review of programs to calculate statistics from survey data, visit the following Website: http://www.hcp.med.harvard.edu/statistics/survey-soft/.

The NIS database includes a Hospital Weights file with variables required by these programs to calculate finite population statistics. The file includes hospital identifiers (Primary Sampling Units or PSUs), stratification variables, and stratum-specific totals for the numbers of discharges and hospitals so that finite-population corrections 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 analytic file is too small to set aside a large validation sample, cross-validation techniques may be used. For example, ten-fold cross-validation would split the data into ten subsets of equal size. 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 in the U.S., although hospital identifiers are suppressed in the NIS for a number of states. For these states it will not be possible to link to outside hospital-level data sources. 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, the number of Cesarean sections performed in each hospital would be correlated with their total number of deliveries. The figure for Cesarean sections must be obtained from discharge data, but the number of deliveries is available from AHA data. Thus, if a regression model can be fit predicting this procedure from deliveries based on the NIS data, that regression model can then be used to obtain hospital-specific estimates of the number of Cesarean sections for all hospitals in the AHA 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.

Studying Trends

When studying trends over time using the NIS, be aware that the sampling frame for the NIS changes over time. Because more states have been added, estimates from earlier years of the NIS may be subject to more sampling bias than later years of the NIS. In order to facilitate analysis of trends using multiple years of NIS data, an alternate set of NIS discharge and hospital weights for the 1988-1997 HCUP NIS were developed. These alternate weights were calculated in the same way as the weights for the 1998 and later years of the NIS. The special report, *Using the HCUP Nationwide Inpatient Sample to Estimate Trends*, includes details regarding the alternate weights and other recommendations for trends analysis. Both the *NIS Trends Report* and the alternate weights are available on the HCUP-US Website under Methods Series (http://www.hcup-us.ahrq.gov/reports/methods.jsp). The *NIS Trends Report* is also available on the NIS Documentation CD-ROM.

To ease the burden on researchers conducting analyses that span multiple years, NIS trends supplemental files (NIS-Trends) are available through the HCUP Central Distributor. The NIS-Trends annual files contain the alternative trend weights for data prior to 1997 in addition to renamed, recoded, and new data elements consistent with the later years of the NIS. More information on these files is available on the HCUP-US Website under NIS database documentation (http://www.hcup-us.ahrq.gov/db/nation/nis/nisdbdocumentation.jsp).

Discharge Subsamples

The two non-overlapping 10% subsamples of discharges were drawn from the NIS file for each year for several reasons pertinent 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 was 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, while 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 in the data.

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.

CONCLUSION

In this report, we have described the development and use of the NIS sample and weights and summarized the contents of the 2004 NIS. We have included cumulative information for all previous years to provide a longitudinal view of the database. We have also highlighted important considerations for data analysis and have provided references to detailed reports on this subject.

The 2004 NIS includes data from 37 states, the same number included in the 2003 NIS. For 2004, state participation has changed slightly, with the loss of Pennsylvania and the addition of Arkansas. The sampling frame is representative of the United States, comprising 75.5% of all U.S. hospitals and encompassing 87.5% of the U.S. population.

ENDNOTES

[1] Most AHA Annual Survey Database files do not cover a January-to-December period for every hospital. The numbers of hospitals for 1988-1991 are based on adjusted versions of the files which we created by apportioning the data from adjacent survey files across calendar years. The numbers of hospitals for later years are based on the unadjusted AHA Annual Survey Database files.

[2] We used the following AHA Annual Survey Database data elements to assign the NIS Teaching Hospital Indicator:

AHA Data Element Name = Description [HCUP Data Element Name].
BDH = Number of short-term hospital beds [B001H].
BDTOT = Number of total facility beds [B001].
FTRES = Number of full-time employees: interns & residents (medical & dental) [E125].
PTRES = Number of part-time employees: interns & residents (medical & dental) [E225].
MAPP8 = Council of Teaching Hospitals (COTH) indicator [A101].
MAPP3 = Residency training approval by the Accreditation Council for Graduate Medical Education (ACGME) [A102].

Prior to the 1998 NIS, we used the following SAS code to assign the NIS teaching hospital status indicator, H_TCH:

```
/* FIRST ESTABLISH SHORT-TERM BEDS DEFINITION */
IF BDH NE . THEN BEDTEMP = BDH ;      /* SHORT TERM BEDS  */
ELSE IF BDH =. THEN BEDTEMP=BDTOT ;   /* TOTAL BEDS PROXY */

/*******************************************************/
/* NEXT ESTABLISH TEACHING STATUS BASED ON F-T & P-T   */
/* RESIDENT/INTERN STATUS FOR HOSPITALS.               */
/*******************************************************/
RESINT = (FTRES + .5*PTRES)/BEDTEMP ;
IF (MAPP3  = . AND MAPP8  = .) THEN DO ;
     IF RESINT > .10 THEN ST_TEACH = 1 ;
     ELSE ST_TEACH = 0 ;
END ;
IF (MAPP3=1 OR MAPP8=1) THEN ST_TEACH=1 ; /* 1=TEACHING    */
ELSE ST_TEACH=0 ;                         /* 0=NONTEACHING */
/*******************************************************/
/* CREATE TEACHING CATEGORY VARIABLES TO FURTHER       */
/* REFINE TEACHING STATUS DEFINITION.                  */
/*******************************************************/
IF ST_TEACH = 1 THEN DO ;
     IF 0 < RESINT < .15 THEN TEACHCAT=0 ;   /* MINOR CATEGORY */
     ELSE IF RESINT GE .15 THEN TEACHCAT=1 ; /* MAJOR CATEGORY */
     ELSE ST_TEACH = 0 ;                     /* NONTEACH STATUS*/
END ;
```

Beginning with the 1998 NIS, we used the following SAS code to assign the teaching hospital status indicator, HOSP_TEACH:

```
/*******************************************************/
/* FIRST ESTABLISH SHORT-TERM BEDS DEFINITION          */
/*******************************************************/
IF BDH NE . THEN BEDTEMP = BDH ;      /* SHORT TERM BEDS  */
ELSE IF BDH =. THEN BEDTEMP = BDTOT ; /* TOTAL BEDS PROXY */
/*******************************************************/
/* ESTABLISH IRB NEEDED FOR TEACHING STATUS            */
/* BASED ON F-T P-T RESIDENT INTERN STATUS             */
/*******************************************************/
IRB = (FTRES + .5*PTRES) / BEDTEMP ;
/*******************************************************/
/* CREATE TEACHING STATUS VARIABLE */
/*******************************************************/
IF (MAPP8 EQ 1) OR (MAPP3 EQ 1) THEN HOSP_TEACH = 1 ;
ELSE IF (IRB GE 0.25) THEN HOSP_TEACH = 1 ;
ELSE HOSP_TEACH = 0 ;
```

[3] U.S. Census Bureau, Population Division. “Table NST-EST2005-01 – Annual Estimates of the Population for the United States and States, and for Puerto Rico: April 1, 2000 to July 1, 2005.” Internet Release Date: December 22, 2005.

[4] Refer to Chapter 10 in Foreman, EK, *Survey Sampling Principles*. New York: Dekker, 1991.

[5] Carlson BL, Johnson AE, Cohen SB. “An Evaluation of the Use of Personal Computers for Variance Estimation with Complex Survey Data.” *Journal of Official Statistics*, vol. 9, no. 4, 1993: 795-814.