HCUP

HEALTHCARE COST AND UTILIZATION PROJECT

Design of the HCUP Kids' Inpatient Database (KID), 2000

January 29, 2003

TABLE OF CONTENTS

INDEX OF TABLES

EXECUTIVE SUMMARY

Introduction

The Kids' Inpatient Database (KID) 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.

The KID is a unique and powerful database of hospital inpatient stays for children. The KID development team designed the database to permit researchers to study a broad range of conditions and procedures related to hospitalizations of children. Researchers and policymakers can use the KID to identify, track, and analyze national trends in hospital utilization, access, charges, quality, and outcomes for children.

The KID is a nationwide sample of pediatric discharges from HCUP State Inpatient Databases (SID) community, non-rehabilitation hospitals weighted to all pediatric discharges in the target universe. The target universe includes all pediatric discharges from community hospitals in the United States that were open during any part of the calendar year. Beginning with the 2000 KID, rehabilitation hospitals were excluded 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.

This report describes the 2000 KID sample design and summarizes the sample contents. Sample weights were developed to obtain national estimates of inpatient parameters. These weights are described in detail. The previous KID release contained data for calendar year 1997. Cumulative information is presented for both 1997 and 2000 to provide a longitudinal view of the database.

Sample Design

Design Considerations

The overall design objective was to select a sample of pediatric discharges that accurately represents the target universe, which includes discharges outside the frame (with zero probability of selection). Moreover, this sample was to be geographically dispersed, yet drawn only from data supplied by HCUP State 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 KID. Ideally, relationships among outcomes and their correlates calculated from the KID should hold across all U.S. hospitals. However, since the 2000 KID includes data from only 27 HCUP State Partners, some estimates may differ from the U.S. When possible, estimates based on the KID should be checked against national benchmarks, such as the National Hospital Discharge Survey, to determine the appropriateness of the KID for specific analyses. (Refer to the report *HCUP Kids' Inpatient Database Comparative Analysis, 1997,* which is available on the 1997 KID Documentation CD-ROM and on the HCUP Website at http://www.ahrq.gov/data/hcup/.)

Sampling Frame

The KID sampling frame included all pediatric discharges from community, non-rehabilitation hospitals in the HCUP State Inpatient Databases (SID) that could be matched to the corresponding AHA survey data (subject to state-specific restrictions). For the 2000 KID, pediatric discharges were defined as having an age at admission of 20 or less. This is a change from the 1997 KID which included discharges with an admission age of 18 or less. Discharges with missing, invalid, or inconsistent ages were excluded.

Sampling Procedure

Unlike the Nationwide Inpatient Sample (NIS), the KID development team did not execute a two-stage sampling procedure. Instead, the KID includes a sample of pediatric discharges from all hospitals in the sampling frame. For the sampling, we stratified the pediatric discharges by uncomplicated in-hospital birth, complicated in-hospital birth, and pediatric non-birth. To further ensure an accurate representation of each hospital's pediatric case-mix, we also sorted the discharges by state, hospital, DRG, and a random number within each DRG. We then used systematic random sampling to select 10 percent of uncomplicated in-hospital births and 80 percent of other pediatric cases from each frame hospital.

Discharge Weights

To obtain national estimates, we developed discharge weights using the AHA universe as the standard. For the weights, we post-stratified hospitals on six characteristics contained in the AHA hospital files. These were the same characteristics used to define the NIS sampling strata, with the addition of an additional stratum for freestanding children's hospitals. Some of the NIS strata definitions were revised for 1998 and subsequent data years, and the 2000 KID used these revised strata. Hospital stratification variables were defined as follows:

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

6. Hospital Type – children's or other hospital.

If there were fewer than two frame hospitals, 30 uncomplicated births, 30 complicated births, and 30 non-birth pediatric discharges sampled in a stratum, we merged that stratum with an "adjacent" stratum containing hospitals with similar characteristics. We created the discharge weights by stratum in proportion to the number of AHA newborns for newborns and in proportion to the total number of (non-newborn) AHA discharges for non-newborns.

Weight Data Elements

In addition to the regular discharge weight data element, DISCWT, the 2000 KID contains a new discharge weight data element named DISCWTCHARGE. Texas discharges were not included in the calculation of DISCWTCHARGE. This data element 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 hospitals in the South.

To produce national estimates, use DISCWT or DISCWTCHARGE to weight sampled discharges in the Core file to the discharges from all U.S. community, non-rehabilitation hospitals. For the 2000 KID, 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. In the 1997 KID, DISCWTCHARGE is not available, and DISCWT should be used to create all national estimates.

The 2000 KID Sample

The Agency for Healthcare Research and Quality (AHRQ) obtained agreements with 27 HCUP State Partners to participate in the 2000 KID. Over 90% of the hospital universe is included in the sampling frame for all but six of these states. Five State Partners – Georgia, Hawaii, Missouri, South Carolina and Virginia – imposed sampling restrictions that limited the percentage of state hospitals included in the frame to between 50 and 82 percent. (Restrictions from other states did not have an appreciable effect on the percentage of hospitals in the sampling frame.) One State Partner, Texas, supplied data from only 70% of the state's hospitals. This is because certain Texas state-licensed hospitals, primarily the smaller hospitals, are exempt from statutory reporting requirements. As a result, small Texas 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.

Although pediatric discharges from hospitals from each region are selected for the KID, the comprehensiveness of the sampling frame varies by region. The percentage of hospitals included in the sampling frame is highest in the Northeast (90%) and in the West (77%), while figures are lower for the South (63%) and the Midwest (30%).

Because the KID sampling frame has a disproportionate representation of the more populous states and contains hospitals with more annual discharges, its comprehensiveness in terms of discharges is higher. The proportion of the regional population in the KID states ranges from 95% in the Northeast to 26% in the Midwest. The five Southern states added for 2000 have substantially increased the percentage of the Southern population represented, from 38.7% in the 1997 KID to 80.8% in the 2000 KID.

There were 2,788 hospitals in the 2000 sampling frame, a 10% increase from the 1997 KID. The final 2000 KID sample included 2,516,833 discharges of children from 2,784 hospitals drawn from 27 frame states representing each region of the United States. The 2000 KID is larger than the 1997 KID across several dimensions:

- The number of states included increased from 22 to 27.
- The number of hospitals included increased from 2,521 to 2,784.
- The number of discharges increased from 1.9 million to 2.5 million.

Variance Calculations

It may be important for researchers to calculate a measure of precision for some estimates based on the KID sample data. Variance estimates must take into account both the sampling design and the form of the statistic. 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. Discharges were randomly selected from within each hospital. Standard formulas for stratified, two-stage cluster sample without replacement may be used to calculate statistics and their variances in most applications.

The KID 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.

INTRODUCTION

The 2000 Kids' Inpatient Database (KID) of the Healthcare Cost and Utilization Project (HCUP) was developed to enable analyses of hospital utilization by children across the United States. The target universe includes all pediatric discharges from all community, non-rehabilitation hospitals in the United States in 2000. The 2000 KID is a nationwide sample of pediatric discharges from the HCUP State Inpatient Databases (SID) community, non-rehabilitation hospitals weighted to all pediatric discharges in the target universe.

Potential research issues focus on both discharge- and hospital-level outcomes. Discharge outcomes of interest include:

- frequency
- costs
- 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 2000 KID sample design, as well as a summary of the resultant sample. Sample weights were developed to obtain national estimates of inpatient parameters. These weights are described in detail. Some tables in this report include information for the previous KID release, the 1997 KID, to provide a longitudinal view of the database.

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. The AHA defines community hospitals as follows: "All nonfederal short-term general and other specialty hospitals, excluding hospital units of institutions." Included among community hospitals are academic medical centers and specialty hospitals such as obstetrics-gynecology, ear-nose-throat, short-term rehabilitation, orthopedic, and pediatric hospitals. Excluded are federal hospitals (Veterans Administration, Department of Defense, and Indian Health Service hospitals), long term hospitals, psychiatric hospitals, alcohol/chemical dependency treatment facilities and hospitals units within institutions such as prisons. Beginning with the 2000 KID, rehabilitation hospitals were excluded 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. (The 1997 KID includes rehabilitation hospitals.) Table 1 shows the number of universe hospitals for each year based on the corresponding AHA Annual Survey.

Table 1. Hospital Universe[1]

Year	Number of Hospitals
1997	5,113
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 for the 2000 KID. Therefore, when two or more hospitals merged to create a new hospital, the original hospitals and the newly formed hospital were all considered separate hospital entities in the universe during the year they merged. Likewise, if a hospital split, the original hospital and all newly created hospitals were separate entities in the universe during the year they split. Finally, hospitals that closed during a year were included as long as they were in operation during some part of the calendar year.

Stratification Variables

For the purpose of calculating discharge weights, hospitals were post-stratified on six characteristics contained in the AHA hospital files. These were the same characteristics used to define the HCUP Nationwide Inpatient Sample (NIS) sampling strata, with the addition of an additional stratum for stand-alone children's hospitals. The definitions of some of the NIS strata were revised for 1998 and subsequent data years, and the 2000 KID used the revised strata. A description of the strata used for the 1997 KID can be found in the report *Kids' Inpatient Database (KID) Design Report, 1997.* This report is available on the 1997 KID Documentation CD-ROM and on the HCUP Website at http://www.ahrq.gov/data/hcup/. For the 2000 KID, 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. Table 2 shows the states in each region.

Table 2. 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

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. The teaching status of children's hospitals identified by the National Association of Children's Hospitals and Related Institutions (NACHRI) is based on a teaching status indicator provided by NACHRI. Other hospitals are considered to be teaching hospitals if they have an AMA-approved residency program, are members of the Council of Teaching Hospitals (COTH), or have 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. 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 fall within 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 containing small numbers of hospitals.

   Rural hospitals were not split according to teaching status, because rural teaching hospitals were rare. For example, in 2000 rural teaching hospitals comprised less than one percent 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.

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+

6. *Hospital Type – freestanding children's or other hospital*. Children's hospitals restrict admissions to children, while other hospitals admit both adults and children. There may be significant differences in practice patterns, severity of illness, and available services between children's hospitals and other hospitals. Data from NACHRI were used to help verify and correct the AHA list of children's hospitals. Children's units in general hospitals were not stratified as children's hospitals.

SAMPLING FRAME

The target universe is all pediatric discharges from community, non-rehabilitation hospitals located in the U.S. The *universe* of hospitals for the 2000 KID was established as all AHA community hospitals located in the U.S. with the exception of rehabilitation hospitals. However, it was not feasible to obtain and process all-payer discharge data from the entire universe of hospitals for at least two reasons. First, all-payer discharge data were not available from all hospitals for research purposes. Second, based on the experience of prior hospital discharge data collections, it would have been too costly to obtain data from individual hospitals, and it would have been too burdensome to process each hospital's unique data structure.

Therefore, the KID *sampling frame* was constructed from the subset of universe hospitals that released their discharge data to HCUP for research use. The Agency for Healthcare Research and Quality (AHRQ) obtained agreements with 27 HCUP Partner organizations that maintain statewide, all-payer discharge data files to include their data in the 2000 KID. These HCUP State Partners were either state agencies or private data organizations, primarily state hospital associations. The 1997 KID included five fewer states, as shown in Table 4.

Table 4. Number of KID States, Hospitals, and Discharges, by Year

Calendar Year	States in the Frame	Number of States	Sample Hospitals	Sample Discharges (Millions)
1997	Arizona, California, Colorado, Connecticut, Florida, Georgia, Hawaii, Iowa, Illinois, Kansas, Massachusetts, Maryland, Missouri, New Jersey, New York, Oregon, Pennsylvania, South Carolina, Tennessee, Utah, Washington, and Wisconsin	22	2,521	1.9
2000	Add Kentucky, Maine, North Carolina, Texas, Virginia, and West Virginia; Drop Illinois	27	2,784	2.5

The list of the entire frame of hospitals was composed of all AHA community, non-rehabilitation hospitals in each of the frame states *that could be matched to the discharge data provided to HCUP*. If an AHA 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 Connecticut, Georgia, Hawaii, South Carolina, Missouri, Tennessee, and Virginia, as described below.

To help ensure the confidentiality of hospitals in Connecticut, the only stand-alone community children's hospital in Connecticut was dropped from the database; bed size for two unique hospitals was changed to mask their identity; and the NACHRI hospital type (NACHTYPE) was set to missing for four hospitals.

Georgia, Hawaii, South Carolina, and Tennessee stipulated that only hospitals appearing in sampling strata with two or more hospitals from the state were to be included in the KID. Because of this restriction, four Georgia hospitals, five Hawaii hospitals, seven South Carolina hospitals, and two Tennessee hospitals were excluded from the KID sampling frame. An additional 41 Georgia hospitals were randomly dropped from the KID sampling frame because Georgia allowed no more than 60% of the state's hospitals to be included in the KID. 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.

Thirty-three Missouri hospitals that did not agree to allow public release of their data were excluded from the sampling frame, leaving 72 hospitals from the state in the frame.

Because Virginia allowed only 50% or less of the state's hospitals be included in the KID, 35 hospitals from the state were randomly dropped from the KID sampling frame.

Table 5 shows the number of AHA, HCUP SID, and KID hospitals by state. A total of 130 hospitals were restricted from the KID sampling frame, leaving 2,788 hospitals with pediatric discharges in the frame. The columns in Table 5 are defined are as follows:

- "AHA Universe Hospitals" lists all community, non-rehabilitation hospitals in the AHA Survey data.
- "All SID Hospitals" lists all hospitals available in the SID.
- "SID Community, Non-Rehabilitation Hospitals" lists potential KID sampling-frame hospitals before applying restrictions to the frame and before dropping hospitals without any pediatric discharges.
- "SID Community, Non-Rehabilitation Hospitals with Pediatric Discharges" lists potential KID sampling-frame hospitals with pediatric discharges before applying restrictions to the frame.
- "KID Sampling-Frame Hospitals" lists hospitals with pediatric discharges in the sampling frame after applying restrictions to the frame.
- "KID Sample Hospitals" lists the hospitals selected for the KID. Four hospitals in the frame were not included in the KID because they had so few pediatric discharges that none were randomly sampled.

The KID sampling frame includes all pediatric discharges from community, non-rehabilitation hospitals in HCUP State Inpatient Databases (SID) matched to the AHA Survey data for the corresponding year (subject to the state-specific restrictions discussed above). For the 2000 KID, pediatric discharges were defined as having an age at admission of 20 or less. This is a change from the 1997 KID which included discharges with an admission age of 18 or less. Discharges with missing, invalid, or inconsistent ages were excluded.

Table 5. Number of AHA, HCUP SID, and KID Hospitals, by State

State	AHA Universe Hospitals	All SID Hospitals	SID Community, Non-Rehabilitation Hospitals	SID Community, Non-Rehabilitation Hospitals with Pediatric Discharges	KID Sampling-Frame Hospitals	KID Sample Hospitals
Non-Frame	1,685	362	332	329	0	0
Arizona	59	59	55	55	55	55
California	384	436	379	374	374	374
Colorado	67	70	66	65	65	65
Connecticut	34	31	31	31	30	30
Florida	193	216	190	189	189	189
Georgia	150	173	150	148	103	103
Hawaii	21	21	18	17	12	12
Iowa	115	117	115	115	115	115
Kansas	133	122	121	120	120	119
Kentucky	100	103	95	95	95	95
Massachuse	73	72	68	68	68	68
Maryland	48	49	47	47	47	47
Maine	36	39	36	36	36	36
Missouri	119	110	105	105	72	72
North	112	122	109	108	108	108
New Jersey	76	76	75	75	75	75
New York	213	222	213	211	211	211
Oregon	59	59	58	58	58	58
Pennsylvani	192	230	187	187	187	187
South	62	62	60	60	51	51
Tennessee	116	111	109	108	106	106
Texas	408	384	287	269	269	267
Utah	41	47	40	40	40	40
Virginia	86	98	84	84	49	49
Washington	83	92	82	81	81	81
Wisconsin	120	139	119	119	119	119
West	54	55	54	53	53	52
Total	4,839	3,677	3,285	3,247	2,788	2,784

SAMPLE DESIGN

Design Considerations

The overall design objective was to select a sample of pediatric discharges that accurately represents the target universe, which includes discharges outside the frame (with zero probability of selection). Moreover, this sample was to be geographically dispersed, yet drawn only from data supplied by HCUP State 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 KID. Ideally, relationships among outcomes and their correlates calculated from the KID should hold across all U.S. hospitals. However, since the 2000 KID includes data from only 27 HCUP State Partners, some estimates may differ from the U.S. When possible, estimates based on the KID should be checked against national benchmarks, such as the National Hospital Discharge Survey, to determine the appropriateness of the KID for specific analyses. (Refer to the report *HCUP Kids' Inpatient Database Comparative Analysis, 1997,* which is available on the 1997 KID Documentation CD-ROM and on the HCUP Website at http://www.ahrq.gov/data/hcup/.)

In order to sample and weight births up to the number of births reported by the AHA, which reports in-hospital births, the KID development team wanted to identify all in-hospital births in the KID data. We also wanted to further separate the in-hospital births into uncomplicated "normal" births and complicated births. We sampled uncomplicated births at a lower rate because they have little variation in their outcomes.

To determine the best way to identify in-hospital births, during the development of the 1997 KID, we ran cross-tabulations of different combinations of variables on all cases that had any of the following possible birth indicators: age of zero days (AGEDAY=0), neonatal diagnosis (NEOMAT>=2), neonatal MDC (MDC 15), or admission type of birth (ATYPE=4). Based on reviews of the cross-tabulations, the MDC 15 DRG definitions, and ICD-9-CM birth codes, the following screen was selected for births: an in-hospital birth diagnosis code (any DX code in the range V3000 - V3901 with a fourth digit of zero and a fifth digit of zero or one), without an admission source of another hospital or health facility (ASOURCE not equal to 2 or 3).

We classified neonates transferred from other facilities as pediatric non-births because they are not included in births reported by the AHA. An age of zero days was not a reliable in-hospital birth indicator since neonates transferred from another hospital or born before admission to the hospital could also have an age of zero days. There were also some cases with birth diagnoses, but with ages of a few days. Since the HCUP data are already edited for neonatal diagnoses inconsistent with age, we did not include any age criteria in the in-hospital birth screen.

"Normal" uncomplicated in-hospital births are identified as cases that meet the above screen and are in DRG 391, "Normal Newborn." Less than one percent of the cases in DRG 391 do not meet the in-hospital birth screen. These cases have diagnoses that imply a newborn, but do not specifically indicate an in-hospital birth. It is possible that some of these may have actually been born in the hospital but lacked the proper V3nnn code. Others, however, may be readmissions or may have been born before admission to the hospital and did not receive a V3nnn code. Less than 0.2% of cases in DRG 391 have an admission type of newborn (ATYPE = 4) but do not meet the in-hospital birth screen.

Using the above in-hospital birth screen, we identified 2,672,707 in-hospital births in community, non-rehabilitation hospitals in the 2000 KID compared to 2,705,106 births reported by the AHA in these hospitals. There were 32,399 more births reported by the AHA, a difference of about 1.2%.

Sampling Procedure

Unlike the NIS, we did not execute a two-stage sampling procedure. Instead, the KID includes a sample of pediatric discharges from all hospitals in the sampling frame. For the sampling, we stratified the pediatric discharges by uncomplicated in-hospital birth, complicated in-hospital birth, and pediatric non-birth. To further ensure an accurate representation of each hospital's pediatric case-mix, we also sorted the discharges by state, hospital, DRG, and a random number within each DRG. We then used systematic random sampling to select 10 percent of uncomplicated in-hospital births and 80 percent of other pediatric cases from each frame hospital.

It should be observed that, for the NIS, states wanted to make it difficult or impossible to identify individual hospitals in part because the NIS included 100% of the discharges from hospitals in the NIS sample. Consequently, outcomes could have been estimated without sampling error for individual hospitals that could be identified in the sample. However, the KID includes fewer than 100% of the pediatric discharges for each hospital in the database. Therefore, researchers will not be able to calculate hospital-specific outcomes with certainty.

SAMPLE WEIGHTS

To obtain national estimates, we developed discharge weights using the AHA universe as the standard. For the weights, we post-stratified hospitals on six characteristics contained in the AHA hospital files. These were the same characteristics used to define the NIS sampling strata, with the addition of an additional stratum for freestanding children's hospitals. We also stratified the KID discharges by whether the discharge was an uncomplicated in-hospital birth, a complicated in-hospital birth, or a non-newborn pediatric discharge. If there were fewer than two frame hospitals, 30 uncomplicated births, 30 complicated births, and 30 non-birth pediatric discharges sampled in a stratum, we merged that stratum with an "adjacent" stratum containing hospitals with similar characteristics.

The discharge weights were created by stratum in proportion to the number of AHA newborns for newborns and in proportion to the number of non-newborn AHA discharges for non-newborns. Refer to the report *Design of the HCUP Kids' Inpatient Database (KID), 1997*, for a discussion of the analysis and development of the KID weighting scheme. This report is available on the 1997 KID Documentation CD-ROM and on the HCUP Website at http://www.ahrq.gov/data/hcup/.

We used NACHRI data to help verify and correct the AHA list of children's hospitals in the target universe. Many of these children's hospitals are units of larger institutions (AHA hospital type 10). Consequently, we do not have separate reporting for them either in the AHA survey or in the HCUP SID. However, data analysts may find it useful to identify hospitals that contain children's units within them.

Discharge Weights

The discharge weights usually are constant for all discharges of the same type (uncomplicated in-hospital birth, complicated in-hospital birth, other pediatric discharge) 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 their full year of data to the KID. For those hospitals, we *adjusted* the number of observed discharges by a factor 4) Q, where Q was the number of calendar quarters that the hospital contributed discharges to the KID. For example, when a sample hospital contributed only two quarters of discharge data to the KID, the *adjusted* number of discharges was double the observed number.

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 numbers of total discharges and births were available for every hospital in the universe from the AHA files.

Discharge weights to the universe were calculated by post-stratification. Hospitals were stratified on geographic region, urban/rural location, teaching status, bed size, control, and hospital type. In some instances, strata were collapsed for sample weight calculations. Within stratum k, for hospital i, each KID sample discharge's universe weight was calculated as:

$$W_{ik} = [T_k / (R_k * A_k)] * (4\ )\ Q_i)$$

In the birth strata (both complicated and uncomplicated):

- T_k is the total number of births reported in the AHA survey.
- A_k is the total number of adjusted births in the restricted sampling frame.
- In the uncomplicated birth strata, R_k is the frame sampling rate for uncomplicated in-hospital births calculated as (sum of adjusted number sampled)/(sum of adjusted number in the restricted frame).
- In the complicated birth strata, R_k is the frame sampling rate for complicated in-hospital births.

In the non-newborn strata:

- T_k is the total number of non-newborns reported in the AHA survey.
- A_k is the total number of adjusted non-newborn discharges in the sampling frame.
- R_k is the frame sampling rate for non-newborns from all non-newborn discharges in the sampling frame.

Q_i is the number of quarters of discharge data contributed by hospital i to the KID (usually Q_i = 4).

T_k / A_k estimates the number of discharges in the population that is represented by each discharge in the sampling frame. R_k adjusts for the fact that we are taking a sample of the frame in each stratum.

Uncomplicated in-hospital births were sampled at a lower rate than other discharges because the variation in hospital outcomes for uncomplicated births is considerably less than that of other pediatric cases and because we expect research to focus much more on other pediatric patients. We sampled uncomplicated births at the nominal rate of 10 percent and sampled other pediatric discharges at the nominal rate of 80 percent from the discharges available in the (restricted) frame. To avoid rounding errors in the weights calculation, the actual sampling rate for a discharge type (uncomplicated in-hospital birth, complicated in-hospital birth, or non-birth pediatric discharge) in stratum k, R_k, was calculated as follows:

$R_k, = S_k / H_k$

- S_k is the number of adjusted discharges sampled for the discharge type in stratum k.
- H_k is the number of adjusted discharges in the sampling frame for the discharge type in stratum k.

The AHA birth counts include both uncomplicated and complicated births. Therefore, the weights in the uncomplicated birth strata implicitly assume that the proportion of births that are uncomplicated in the frame is representative of the proportion of births that are uncomplicated in the population for each stratum. A similar assumption is made for complicated newborns.

Similarly, the non-birth AHA discharge counts include <u>all</u> non-birth discharges, not just non-birth pediatric discharges. Consequently, the weights in the non-birth strata implicitly assume that the proportion of discharges that are non-birth pediatric across the HCUP SID hospitals is the same as the proportion of discharges that are non-birth pediatric across the universe of AHA hospitals, in the aggregate within each stratum.

Weight Data Elements

In addition to the regular discharge weight data element, DISCWT, the 2000 KID contains a new discharge weight data element named DISCWTCHARGE. Texas discharges were not included in the calculation of DISCWTCHARGE. This data element 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 hospitals in the South.

To produce national estimates, use DISCWT or DISCWTCHARGE to weight sampled discharges in the Core file to the discharges from all U.S. community, non-rehabilitation hospitals. For the 2000 KID, 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. In the 1997 KID, DISCWTCHARGE is not available, and DISCWT should be used to create all national estimates.

THE 2000 KID SAMPLE

There were 2,788 hospitals in the 2000 sampling frame, a 10% increase from the 1997 KID. The final 2000 KID sample included 2,516,833 discharges of children from 2,784 hospitals drawn from 27 frame states representing each region of the United States. The 2000 KID is larger than the 1997 KID across several dimensions:

- The number of states included increased from 22 to 27.

- The number of hospitals included increased from 2,521 to 2,784.
- The number of discharges increased from 1.9 million to 2.5 million.

Table 6 summarizes the numbers of hospitals and discharges for children's hospitals and other hospitals. For each hospital type, the table shows the number of:

- AHA universe hospitals and total discharges, including births
- Pediatric discharges from non-rehabilitation community hospitals in the SID
- KID nationwide hospitals and discharges.

Table 6. Numbers of Hospitals and Discharges in AHA Universe, HCUP SID and KID, by Hospital Type, 2000

Hospital Type	AHA Universe Hospitals	AHA Discharges (Including Births)	SID Community, Non-Rehabilitation Hospitals with Pediatric Discharges	SID Pediatric Discharges	KID Hospitals	KID Pediatric Discharges
Not a Children's Hospital	4,765	35,937,297	3,210	5,302,275	2,754	2,314,466
Children's Hospital	74	480,268	37	320,109	30	202,367
Total	4,839	36,417,565	3,247	5,622,384	2,784	2,516,833

Table 7 summarizes the 2000 KID hospital sample by geographic region. For each region, the table shows the number of:

- AHA universe hospitals and total discharges, including births
- Pediatric discharges from non-rehabilitation community hospitals in the HCUP SID
- KID nationwide sample hospitals and discharges.

Table 7. Number of Hospitals and Discharges in AHA Universe, HCUP SID and KID, by Region, 2000

Region	AHA Universe Hospitals	AHA Discharges (Including Births)	SID Community, Non-Rehabilitation Hospitals with Pediatric Discharges	SID Pediatric Discharges	KID Hospitals	Percent of AHA Hospitals in KID	KID Pediatric Discharges
Northeast	675	7,349,649	608	1,178,096	607	89.93%	617,486
Midwest	1,398	8,426,138	788	952,421	425	30.40%	176,569
South	1,860	13,714,943	1,161	2,064,275	1,067	57.37%	1,015,791
West	906	6,926,835	690	1,427,592	685	75.61%	706,987
Total	4,839	36,417,565	3,247	5,622,384	2,784	57.53%	2,516,833

For example, in 2000, the Northeast region contained 675 hospitals in the AHA universe. Of these, 608 with pediatric discharges are available in the SID, and 607 appeared in the KID.

Table 8 summarizes the estimated U.S. population on July 1, 2000[2], by geographic region. For each region, the table shows:

- The estimated U.S. population
- The estimated population of states in the 2000 KID
- The percentage of estimated U.S. population included in KID states.

Table 8. Percentage of U.S. Population in 2000 KID States, by Region

Region	Estimated U.S. Population	Population of KID States	Percent of U.S. Population in KID States
Northeast	53,644,868	50,745,042	94.6
Midwest	64,473,034	16,595,055	25.7
South	100,562,076	81,302,933	80.8
West	63,444,653	56,280,631	88.7
Total	282,124,631	204,923,661	72.6

For example, the estimated population of the Northeast on July 1, 2000, was 53,644,868. The estimated population on July 1, 2000, of states in the Northeast that were included in the 2000 KID was 50,745,042. This represents 94.6% of the total Northeast population. The percentage of estimated U.S. population included in states in the 2000 KID was almost as high in the West (88.7%), but was much lower in the Midwest (25.7%). The five Southern states added for 2000 have substantially increased the percentage of the Southern population represented, from 38.7% in the 1997 KID to 80.8% in the 2000 KID.

Table 9 shows the number of hospitals and discharges in the AHA universe, in SID community, non-rehabilitation hospitals with pediatric discharges, and in the KID for each state in the sampling frame for 2000. Some states have fewer hospitals in the KID than the number of HCUP SID hospitals with pediatric discharges in the state for one or both of the following reasons: 1) four hospitals have so few pediatric discharges that none were selected for the nationwide sample; and 2) as previously described, Connecticut, Georgia, Hawaii, South Carolina, Tennessee, Missouri, and Virginia restricted which hospitals could be included in the KID.

Table 9. Number of Hospitals and Discharges in AHA Universe, SID, and KID, by State, 2000

State	AHA Universe Hospitals	AHA Discharges (including births)	SID Community, Non-Rehabilitation Hospitals with Pediatric Discharges	SID Pediatric Discharges	KID Hospitals	KID Pediatric Discharges
Non-Frame	1,685	10,346,143	329	554,244	0	0
Arizona	59	606,450	55	142,931	55	73,253
California	384	3,800,576	374	875,615	374	434,913
Colorado	67	453,135	65	105,382	65	53,062
Connecticut	34	388,905	31	72,641	30	31,654
Florida	193	2,281,770	189	379,562	189	217,268
Georgia	150	974,507	148	220,801	103	70,399
Hawaii	21	109,807	17	24,407	12	10,887
Iowa	115	385,380	115	69,470	115	35,519
Kansas	133	340,185	120	62,247	119	31,397
Kentucky	100	617,652	95	89,429	95	48,867
Massachusetts	73	792,642	68	139,252	68	71,392
Maryland	48	637,909	47	115,211	47	63,933
Maine	36	157,478	36	25,141	36	14,002
Missouri	119	822,011	105	145,044	72	48,091
North Carolina	112	1,077,957	108	199,136	108	103,830
New Jersey	76	1,167,820	75	196,735	75	102,282
New York	213	2,666,558	211	462,725	211	242,679
Oregon	59	372,518	58	76,002	58	35,802
Pennsylvania	192	1,865,167	187	281,602	187	155,477
South Carolina	62	538,833	60	103,617	51	48,001
Tennessee	116	792,386	108	147,402	106	76,778
Texas	408	2,674,629	269	608,782	267	309,879
Utah	41	238,193	40	75,724	40	36,004
Virginia	86	810,965	84	159,201	49	53,608
Washington	83	573,673	81	127,531	81	63,066
Wisconsin	120	625,842	119	121,416	119	61,562
West Virginia	54	298,474	53	41,134	52	23,228
Total	4,839	36,417,565	3,247	5,622,384	2,784	2,516,833

Only 70% of Texas hospitals were supplied to HCUP, as can be seen in Table 9. This is because 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. As can be seen from Table 10 below, small hospitals are substantially less likely to be included in the sampling frame, while larger hospitals are more likely to be included.

Table 10. Texas Hospitals Included and Excluded from Sampling Frame by Bed Size Category

Bed Size Category	Non-Rehabilitation AHA Community Hospitals	AHA Discharges	Hospitals in Sampling Frame	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	287	121	70%	94%

While the number of hospitals omitted appears sizable, Table 11 below shows that the hospitals available to the KID include 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 KID sampling frame.

Table 11. Texas Hospitals and Discharges Included and Excluded from Sampling Frame By Control

Included in Sampling Frame	Control	Non-Rehabilitation AHA Community Hospitals	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%

Table 12 shows the non-weighted and weighted number of uncomplicated births, complicated births, and pediatric non-births by hospital type in the 2000 KID.

Table 12. KID Discharges

Hospital Type	Uncomplicated Births	Complicated Births	Pediatric Non-Births	Total Pediatric Discharges
Non-Weighted:				
Not a Children's Hospital	193,018	587,650	1,533,798	2,314,466
Children's Hospital	406	3,117	198,844	202,367
Total	193,424	590,767	1,732,642	2,516,833
Weighted:				
Not a Children's Hospital	2,822,093	1,060,993	2,951,852	6,834,938
Children's Hospital	2,452	2,351	451,291	456,094
Total	2,824,545	1,063,344	3,403,143	7,291,032

DATA ANALYSIS

Variance Calculations

It may be important for researchers to calculate a measure of precision for some estimates based on the KID sample data. Variance estimates must take into account both the sampling design and the form of the statistic. 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. Discharges were randomly selected from within each hospital. Standard formulas for stratified, two-stage cluster sample without replacement may be used to calculate statistics and their variances in most applications.

A multitude of statistics can be estimated from the KID 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 KID, 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 (1997 or 2000) should be governed by finite-sample theory.

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, 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, while 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.[3] 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 Website. For an excellent review of programs to calculate statistics from survey data, visit the following Website: http://www.fas.harvard.edu/~stats/survey-soft/.

The KID 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.

ENDNOTES

[1] Most AHA surveys do not cover a January-to-December calendar year for every hospital. The numbers of hospitals for 1997 and 2000 are based on the AHA Annual Survey files.

[2] State Population Estimates, July 1, 2000. Source: Population Division, U.S. Census Bureau. Internet Release Date: December 29, 2000.

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