H·CUP

HEALTHCARE COST AND UTILIZATION PROJECT

HCUP Methods Series

Calculating National Inpatient Sample (NIS) Variances for Data Years 2012 and Later

Report # 2015-09

AHRQ
Agency for Healthcare
Research and Quality

U.S. Department of Health and Human Services
Agency for Healthcare Research and Quality

Contact Information:
Healthcare Cost and Utilization Project (HCUP)
Agency for Healthcare Research and Quality
5600 Fishers Lane
Room 07W17B
Mail Stop Number 7W25B
Rockville, MD 20857
http://www.hcup-us.ahrq.gov

For Technical Assistance with HCUP Products:

Email: hcup@ahrq.gov

or

Phone: 1-866-290-HCUP

Recommended Citation: Houchens R, Ross D, Elixhauser A. *Final Report on Calculating National Inpatient Sample (NIS) Variances for Data Years 2012 and Later*. 2015. HCUP Methods Series Report # 2015-09 ONLINE. December 14, 2015. U.S. Agency for Healthcare Research and Quality. Available: http://www.hcup-us.ahrq.gov/reports/methods/methods.jsp.

TABLE OF CONTENTS

PREFACE **1**

EXECUTIVE SUMMARY **2**

INTRODUCTION **3**

NIS Sample Design 4

NIS Sample Weights 4

MISSING VALUES **5**

RATIONALE AND FORMULAS FOR NIS VARIANCE CALCULATIONS **6**

EXAMPLES OF NIS VARIANCE CALCULATIONS **8**

SAS Programming Statements 10

Stata Programming Statements 12

Comparison of Estimates 14

Finite Population Corrections 14

Combining Multiple Years of the NIS 14

Analyzing Subpopulations 15

DISCUSSION **15**

Alternative Concepts of Variance 15

Estimation Techniques 16

CONCLUSIONS **17**

APPENDIX A: CODE FOR ANALYZING SUBPOPULATIONS USING SAS AND STATA **18**

SAS Programming Statements 18

Stata Programming Statements 20

LIST OF TABLES

Table 1. SAS Output for DIABETES = 1, Domain Statistics 11

Table 2. Stata Output for Diabetes = 1 13

Table 3. Comparison of SAS and Stata Results for Complicated Diabetes, National Inpatient Sample (NIS), 2012 14

PREFACE

This version of the National Inpatient Sample (NIS) variance report applies to the latest NIS sample design, which is effective for data years 2012 and later. For data years 2011 and earlier, users should consult the previous version of the NIS variance report.[1]

[1] Houchens R, Elixhauser A. Final Report on Calculating Nationwide Inpatient Sample (NIS) Variances for Data Years 2011 and Earlier. HCUP Methods Series Report No. 2003-02. Rockville, MD: Agency for Healthcare Research and Quality; December 14, 2015 http://www.hcup-us.ahrq.gov/reports/methods/methods.jsp.

EXECUTIVE SUMMARY

The Healthcare Cost and Utilization Project (HCUP) is a Federal-State-Industry partnership to build a standardized, multistate health data system. HCUP databases contain encounter-level health care data submitted by participating States. One HCUP database, the National Inpatient Sample (NIS), is the largest all-payer inpatient care database in the United States.

Beginning with the 2012 data year, the NIS is a stratified sample of discharges from all hospitals in the sampling frame defined by States that make their data available to the HCUP project and that can be matched to data from the American Hospital Association (AHA) Annual Survey of Hospitals*.* Hospitals are stratified by region, location and teaching status, bed size category, and ownership. Prior to the 2012 NIS, the samples included all discharges from a sample of hospitals in the sampling frame. For data prior to 2012, see the document titled *Calculating National (Nationwide) Inpatient Sample Variances, Data Year 2011 and Earlier*.[2]

This document describes how to calculate simple statistics, including variances, from the NIS while taking into account the sampling design and sample discharge weights. Data from the 2012 NIS are used in all examples in this report, although the same methods can be applied to all subsequent data years. This report contains the program code required to calculate sample totals, means, rates, and their variances with two commonly used statistical programming languages that run on personal computers: SAS (SAS Institute Inc) and Stata (StataCorp LP). This report also provides results of example calculations from both statistical packages and demonstrates that the results are virtually the same for both statistical packages.

Two approaches to calculating variances for subpopulations are suggested. The first, described in the body of the report, uses the entire NIS sample. The second, described in Appendix A, uses only the subsample of the NIS corresponding to the subpopulation of interest. Finally, we discuss alternative concepts of variance and other methods that could be applied to calculating variances.

[2] Houchens R, Elixhauser A. Final Report on Calculating Nationwide Inpatient Sample (NIS) Variances, for Data Years 2011 and Earlier. HCUP Methods Series Report No. 2003-02. Rockville, MD: Agency for Healthcare Research and Quality; December 14, 2015 http://www.hcup-us.ahrq.gov/reports/methods/methods.jsp.

INTRODUCTION

The National Inpatient Sample[3] (NIS), a database of United States hospital discharge data, is designed to inform policy decisions regarding health and health care at the national and regional levels. Through NIS data, researchers can make inferences about national trends in health care utilization, access, cost, quality, and outcomes. 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 all-payer inpatient care database in the United States. The NIS has been made publicly available since the 1988 data year.

This document describes how to calculate statistics, including variances, from the NIS while taking into account the sampling design and sample discharge weights. Data from the 2012 NIS are used in the examples. This report also contains program code required to calculate sample totals, means, rates, and their variances with two commonly used statistical programming languages that run on personal computers: SAS and Stata. Although the examples in this report used SAS Version 9.4 and Stata Version 12, the example code is likely to run on earlier and later versions of these software packages, perhaps with some minor alterations.

Both programming languages have procedures for calculating sample statistics and appropriate variances based on data from complex sampling designs.[4,5] This is important, because unweighted statistics or analyses that otherwise fail to account for the NIS sample design could yield biased estimates. Although this report does not cover multivariate statistical procedures such as regression analysis, some concepts introduced in this report also carry over to those types of analysis.

Several other commercial and noncommercial statistical programming packages allow weighted analyses. If the user prefers to use a statistical package other than SAS or Stata, it is likely that the options and statements for other packages will be similar to those for SAS and Stata.

This report also gives the results of example calculations from both statistical packages. Therefore, users who have a copy of the 2012 NIS can run the program code in this document and check the results obtained against the results reported here.

This introduction continues below with a brief overview of the NIS sample design and the discharge weights that accompany the NIS database. Users desiring a more comprehensive account should refer to the final report on the NIS sampling and weighting strategies.[6] This introduction ends with a brief discussion on the treatment of missing values in the data.

[3] Beginning with the 2012 data year, the Agency for Healthcare Research and Quality redesigned the NIS and changed its name from *Nationwide Inpatient Sample* to *National Inpatient Sample*.

[4] SAS/STAT® 14.1 User's Guide. Cary, NC: SAS Institute Inc; 2015. https://support.sas.com/documentation/onlinedoc/stat/141/whatsnew.pdf

[5] Stata Survey Data Reference Manual, Release 14. College Station, TX: StataCorp; 2015. http://www.stata.com/bookstore/survey-data-reference-manual/

[6] Houchens RL, Ross DN, Elixhauser A, Jiang J. Nationwide Inpatient Sample Redesign Final Report. Rockville, MD: Agency for Healthcare Research and Quality; April 4, 2014. https://www.hcup-us.ahrq.gov/db/nation/nis/reports/NISRedesignFinalReport040914.pdf

NIS Sample Design

For the NIS data years 2012 and later, the final sample design is as follows. The hospital universe is defined by all hospitals that were open during any part of the calendar year and were designated as community hospitals in the American Hospital Association (AHA) Annual Survey of Hospitals, excluding rehabilitation and long-term acute care (LTAC) hospitals. The NIS uses the AHA's definition of a community hospital: "all nonfederal short-term general and other specialty hospitals, excluding hospital units of institutions." Consequently, U.S. Department of Veterans Affairs hospitals and other federal hospitals also are excluded.

The NIS is a self-weighted, stratified, systematic, random sample of *discharges* from *all* hospitals in the sampling frame, after sorting discharges by diagnosis-related group (DRG), hospital, and admission month. The sampling frame is the subset of hospitals in States that make their data available to the HCUP project and *that can be matched to the AHA survey data.*

For 2012, there are 196 strata. Hospitals are stratified by census division, location and teaching status (within region), bed size category (within region and within location and teaching status), and ownership (within region, location and teaching, and bed size categories).

There are nine census divisions. We categorized hospitals with a Core-Based Statistical Area (CBSA) type of metropolitan or division as *urban*, whereas we designated hospitals with a CBSA type of micropolitan or rural as *rural*. Teaching hospitals are those with membership in the Council of Teaching Hospitals (COTH), with a residency program approved by the American Medical Association, or with an intern-to-bed ratio of 25 percent or higher. Bed size categories are small, medium, and large, with separate size cut points defined for each combination of hospital region, teaching status, and urban or rural designation. Ownership breakdowns are based on the degree of observed ownership variation within each region across bed size categories.

Details about the present NIS sample design, including the rationale for changes from the previous NIS design and the effects of those changes on estimates, are available in the document titled *Nationwide Inpatient Sample Redesign Final Report.*[7]

NIS Sample Weights

In the NIS files from 1998 and later years, the discharge weight data element is named DISCWT. To produce national estimates, we use DISCWT to weight sampled discharges in the NIS to the discharges from all nonrehabilitation, non-LTAC community hospitals located in the United States.

The discharge sample weights are calculated within each sampling stratum as the ratio of discharges in the universe to discharges in the sample. Starting with the 2012 NIS, the discharge samples are self-weighted, meaning that the discharge weight is constant throughout the sample. Therefore, estimates of sample means are the same whether the discharges are weighted or unweighted. The discharge sampling rate for each stratum is dictated by the proportion of universe discharges contained in the sampling frame. Similar to the previous NIS

[7] Houchens RL, Ross DN, Elixhauser A, Jiang J. Nationwide Inpatient Sample Redesign Final Report. Rockville, MD: Agency for Healthcare Research and Quality; April 4, 2014. https://www.hcup-us.ahrq.gov/db/nation/nis/reports/NISRedesignFinalReport040914.pdf

weight calculations, this requires an estimate of discharges in the universe. However, the basis of this estimate differs from the method used prior to the 2012 NIS, which estimated the number of discharges in the universe exclusively on the basis of estimates taken from the AHA annual hospital survey.

Starting with the 2012 NIS, the number of discharges in the universe is estimated on the basis of the AHA data solely for hospitals outside the sampling frame, representing only about five percent of the universe. For hospitals contained inside the sampling frame, the number of discharges in the universe is estimated from the count of discharges observed in the HCUP data. This has implications for trends estimated from multiple years of the NIS. These implications are discussed in the NIS sample redesign final report cited previously.

MISSING VALUES

The procedures presented in this report omit cases with missing values from all calculations. Missing values for any reason 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. There are several techniques available to help overcome this bias. One strategy is to use imputation to replace missing values with acceptable values. For more information see the HCUP report *Missing Data Methods for the NIS and SID*.[8] Another strategy is to use sample weight adjustments to compensate for missing values.[9] These types of data preparation and adjustment are outside the scope of this report. However, if these adjustments are necessary, they should be completed before analyzing data with the statistical procedures presented here.

It should be noted that if the cases with and without missing values are assumed to be similar with respect to their outcomes, then no adjustment may be necessary for estimates of means and rates. This assumes that the means and rates based on nonmissing cases would be representative of the means and rates of missing cases. However, some adjustment still may be necessary for the estimates of totals. Totals (of non-negative variables) would tend to be underestimated in the presence of missing values of the variable for which the total is estimated, because the cases with missing values would be omitted from the calculations.

The next section establishes some sampling concepts in a short discussion of a formula that could be used to calculate the variance of a total from the NIS sample. The following sections contain the program code required to estimate some sample statistics and their variances using SAS and Stata. We demonstrate that the results are identical or very similar for both programs. Finally, we discuss the finite population correction, alternative concepts of variance, and other methods that could be applied to calculate variances.

[8] Houchens R. Missing Data Methods for the NIS and the SID. HCUP Methods Series Report No. 2015-01. Rockville, MD: Agency for Healthcare Research and Quality; January 22, 2015. http://www.hcup-us.ahrq.gov/reports/methods/2015_01.pdf

[9] See, for example, Foreman EK. Survey Sampling Principles. New York: Dekker; 1991, Chapter 10.

RATIONALE AND FORMULAS FOR NIS VARIANCE CALCULATIONS

For a simple random sample of discharges, the usual variance calculations are appropriate. For example, the unbiased estimate for the variance of hospital length of stay (LOS) based on a sample of n discharges would be calculated as the following:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

where x_i is the LOS for discharge i, and $\bar{x}$ is the mean LOS over the sample of *n* discharges. Consequently, the estimated standard error of the mean would be calculated as $\hat{\sigma} \big/ \sqrt{n}$.

However, the sample of NIS discharges is not a simple random sample. As stated previously, the NIS is a self-weighted, stratified, systematic, random sample of *discharges* from *all* hospitals in the sampling frame, after sorting discharges by DRG, hospital, and admission month. Although the NIS sample design is *not* a cluster sample per-se, the sample is stratified on hospital characteristics, and the variance calculations still should account for the clustering of discharges within hospitals. This clustering tends to induce dependence among discharges within hospitals, because the patients discharged from a hospital share a set of treatment resources (e.g., staff and facilities) that differs from the treatment resources available to patients discharged from another hospital.

Discharges were sampled within each stratum at the rate necessary to achieve a sample size equal to 20 percent of the total discharges in the hospital universe in that stratum. Consequently, the NIS sample resembles a stratified two-stage cluster sample. Within each stratum, hospitals (clusters) were selected from the sampling frame at the rate of 100 percent and discharges within hospitals were sampled at a rate of 20 percent or more.

A complication, which we ignore in calculating variances below, is that the hospital sampling frame did not contain the entire universe of U.S. hospitals. The frame contained only hospitals in the States for which all-payer discharge data were made available to the HCUP project. To the extent that States in the frame differ from other States on outcomes within each stratum, this could lead to biased estimates. Consequently, users should compare estimates from the NIS to other benchmarks whenever they are available. For many data years, comparisons between the NIS and other data sources are available on the HCUP User Support Web Site (www.hcup-us.ahrq.gov).

The variance formula for a stratified, two-stage cluster sample employs weights and components for two stages of sampling. This is necessary to account for the possibility that sample discharges within hospitals may be more homogeneous in their outcomes than sample discharges between hospitals. If the analyst wants finite population estimates, then finite population correction (fpc) factors also are needed to correct for the proportion of the universe included in the sample at each level. However, we generally recommend against using the fpc for typical uses of the NIS, because inferences usually are not specific to the exact population of discharges defined by the finite universe of discharges for that particular set of patients treated in that particular year.

The following example is meant to illustrate the “behind the scenes” calculations that statistical programs make for variances based on sample designs. The reader may safely go to the next section of this report without understanding the technical details of this example.

For this example, consider the estimated total of a variable Y, calculated as the weighted sum:

$$T = \sum_{s}\sum_{h}\sum_{d} w_{shd} y_{shd}$$

where:

y_{shd} = the observed value of variable Y for sample discharge *d* within sample hospital *h* within stratum *s*.

w_{shd} = a set of discharge weights or any other constants over the set of sample discharges, hospitals, and strata. The NIS sample weights are constant. However, we retain the subscript *d* to account for the possibility that an analyst would want to adjust the weights to account for missing values in a way that creates unequal weights across patients. For example, the weights might be made to vary according to some patient-level characteristic, such as the patient’s DRG, if the rate of missing values varies by that characteristic.

In any case, an estimate of the variance of T from the sample follows:

$$\hat{\sigma}_T^2 = \sum_{s}(1 - f_s)n_s V_s + \sum_{s} f_s \sum_{h}(1 - f_{sh})n_{sh} V_{sh} \qquad (1)$$

where:

f_s = the proportion of the universe hospitals sampled in stratum *s*. In the NIS, this usually is close to 100 percent, but it varies because the number of frame hospitals is sometimes less than the number of universe hospitals within the strata. If we wish to generalize results to a broader set of hospitals and patients outside that year’s hospital population (recommended), then we would set $f_s = 0$. This might be desirable, for example, if the analyst wishes to draw inferences about a future year or wishes to use the results to set policy going forward.

n_s = the number of hospitals within stratum s.

f_{sh} = the proportion of the discharges in the sample from sample hospital *h* within stratum *s*. For the NIS, f_{sh} is ≥ 0.20. Again, the analyst may wish to consider the NIS a sample from an infinite population (of possible patients), in which case we would assign $f_{sh} = 0$ rather than a finite population.

n_{sh} = the number of discharges in hospital h within stratum s.

V_s = the component of variance due to the first stage of sampling (variation among hospitals within stratum s):

$$V_s = \frac{\sum_h \left(\sum_d w_{shd} y_{shd} - \frac{\sum_h \sum_d w_{shd} y_{shd}}{n_s} \right)^2}{(n_s - 1)}$$

Notice that the numerator is the sum of squared deviations of the individual hospital totals from the mean hospital total, and the sum is over all hospitals in stratum s; this is similar to the familiar calculation for the variance of any sample statistic. Also notice in equation (1) that this term is multiplied by zero if $f_s = 1$. In that case, all hospitals within stratum s are sampled, and the estimated total for that stratum has no sampling error associated with it.

V_{sh} = the component of variance due to the second stage of sampling (variation among discharges within hospital h in stratum s):

$$V_{sh} = \frac{\sum_d \left(w_{shd} y_{shd} - \frac{\sum_d w_{shd} y_{shd}}{n_{sh}} \right)^2}{n_{sh} - 1}$$

Again, this calculation of a variance is familiar. The numerator is the sum of squared deviations of the individual weighted discharge totals from the mean weighted discharge total for each hospital h in stratum s. If the sampling rate is $f_{sh} = 1$ for hospital h in stratum s, then this term is multiplied by zero in equation (1) because the hospital total is estimated without error. However, $f_{sh} < 1$ for the NIS.

Many statistical packages use similar formulas to estimate variances for simple statistics such as means and totals.

It is important to recognize that these variance calculations assume that the analyst is interested in making inferences to the finite population of hospital discharges in the year of the data. As the sampling fraction *f* approaches 1, the sampling variance approaches zero. If the analyst is interested in making inferences to another population rather than to the specific discharges represented in the discharge population for the year of the data, then the sampling fraction *f* should be set to zero. Our examples will not use the fpc. However, we will indicate the effect of the fpc and how the fpc could be incorporated.

EXAMPLES OF NIS VARIANCE CALCULATIONS

The example analysis is for a subpopulation of the 2012 NIS defined by a Clinical Classifications Software (CCS) diagnosis category code equal to 50: diabetes mellitus with complications. CCS is a tool developed by AHRQ for clustering patient diagnoses and procedures into a manageable number of clinically meaningful categories.[10]

To obtain estimates, we created an indicator variable using SAS to identify the subset of discharges with complicated diabetes for the analysis (DXCCS1 = 50). Using Stata, we also generated a data file in American Standard Code for Information Interchange (ASCII) format.

The SAS and Stata program codes for the analysis of the diabetes subpopulation are shown below, along with examples of the output produced by each program.

In these examples, the following conventions apply:

- UPPERCASE WORDS denote NIS variable names.
- Lowercase words denote keywords and options that are part of the programming language.
- *Italicized words* denote information to be supplied by the researcher.
- **Bold words** denote comments.

Note that the example programs shown here use the entire NIS. This approach generally is recommended for calculating standard errors because it will yield correct standard errors. If computing constraints force the use of a subset of the NIS (such as a specific subpopulation defined by condition or by patient characteristics), then refer to Appendix A for alternative methods.

[10] Healthcare Cost and Utilization Project. Clinical Classifications Software for ICD-9-CM Fact Sheet. Rockville, MD: Agency for Healthcare Research and Quality; January 2012. http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccsfactsheet.jsp

SAS Programming Statements

The following are the SAS programming statements for the example described above:

```
/* Create analysis file. */
libname IN “location of NIS file” ;
data DIABETES ;
   set IN.NIS_2012_CORE ;
   if DXCCS1 = 50 then DIABETES = 1; else DIABETES = 0 ;
   DISCHGS = 1 ;
run ;

/* Obtain estimates: The following SAS code produces estimates of
the sums, means, and standard errors for the number of discharges,
length of stay, and total hospital charges. */

/* Note: If finite population estimates of standard errors are    */
/* desired, then the PROC SURVEYMEANS statement could include the */
/* option "RATE = .20", which gives the approximate hospital      */
/* sampling rate in each stratum.                                 */

proc surveymeans data= DIABETES sum std mean stderr /*rate = .20*/ ;
   weight DISCWT ;
   cluster HOSP_NIS ;
   strata NIS_STRATUM ;
   var DISCHGS LOS DIED TOTCHG ;
   domain DIABETES ;
run ;
```

Note the following:

- The proc surveymeans statement invokes the SAS procedure.

- The data= option requests that the analysis be performed on the file specified. If this statement is omitted, SAS uses the most recently created dataset.

- The sum option requests the sum for variables listed in the var statement. For example, the variable dischgs is set to equal 1 for every record, so its sum estimates the total number of discharges.

- The std option requests the standard error of the sum.

- The mean and stderr options request that the mean and its standard error be printed. The default statistics are the mean, its standard error, and 95% confidence limits.

- If finite population estimates of standard errors are desired, then the PROC SURVEYMEANS statement could include the option "RATE = .20", which gives the approximate hospital sampling rate in each stratum.

- The weight statement weights each record by the value of the variable DISCWT.

- The stratum statement specifies NIS_STRATUM as the stratum identifier.

- The cluster statement specifies HOSP_NIS as the cluster identifier.

- The var statement requests the statistics for the variables DISCHGS, LOS, TOTCHG, and DIED. If the var statement is omitted, statistics will be calculated for all of the variables in the dataset except for those listed in the weight, stratum, or cluster statement.

- The domain statement requests statistics for the subpopulation of diabetics (and nondiabetics).

These commands produced the output shown in Table 1.

Table 1. SAS Output for DIABETES = 1, Domain Statistics

Variable	Label	Mean	Std Error of Mean	Sum	Std Dev
DISCHGS		1.000000	0	528,030	5,451.714493
LOS	Length of stay (cleaned)	4.601903	0.025091	2,429,116	29,480
DIED	Died during hospitalization	0.005531	0.000231	2919.999322	124.540093
TOTCHG	Total charges (cleaned)	33,914	422.923358	17,532,625,874	290,339,323

HCUP National Inpatient Sample, 2012. Principal diagnosis only.

Stata Programming Statements

The following are SAS programming statements to prepare the data for Stata for the example described above:

```
/* Using SAS, create an ASII file for use by STATA. */
libname IN “location of NIS file” ;
data _null_ ;
   set IN.NIS_2012_CORE ;
   retain dischgs 1 ;
   if DXCCS1 = 50 then DIABETES = 1 ; else DIABETES = 0 ;
   if DIED < 0 then DIED = . ;
   if TOTCHG < 0 then TOTCHG = . ;
   if LOS < 0 then LOS = . ;
   file FILEREF ;
   put NIS_STRATUM “,” Hosp_NIS “,” DIED “,” LOS “,” DISCHGS “,”
       TOTCHG “,” DIABETES “,” DISCWT ;
run ;
```

The following are the Stata programming statements for the example described above:

```
/* Obtain STATA estimates. */
/* Note: Stata commands should be entered in lowercase text. */

infile int    nis_stratum  ///
       long   hosp_nis     ///
       int    died         ///
       long   los          ///
       int    dischgs      ///
       double totchg       ///
       int    diabetes     ///
       double discwt       ///
       using “dataset name”
svyset hosp_nis [pw=discwt], strata(nis_stratum)
svy: total dischgs, subpop(diabetes)
svy: mean los, subpop(diabetes)
svy: mean totchg, subpop(diabetes)
svy: ratio died dischgs, subpop(diabetes)
```

Note the following:

- The infile command lists the variables to read from the dataset created for this analysis.
- The svyset command identifies the weight variable, the stratification variable, and the primary sampling unit.
- The syv: TOTAL command requests the estimate of the total and standard error for the variable listed.

- The svy: MEAN command requests the estimate of the mean and its standard error for the variable listed. Separate MEAN commands should be used for each variable. Otherwise Stata will exclude observations with missing values for any of the listed variables from the estimate.

- The svy: RATIO command requests the ratio of the two variables listed; in this case, the ratio of those who died to total discharges.

- The subpop option requests statistics for the subpopulation of diabetics.

These commands produce the output shown in Table 2.

Table 2. Stata Output for Diabetes = 1

Output Measure	Survey Total Estimation, Discharges	Survey Mean Estimation, Length of Stay	Survey Mean Estimation, Total Charges	Survey Ratio Estimation, Died/Dischgs
No. of strata[†]	195	195	195	195
No. of PSUs	4,375	4,375	4,375	4,375
No. of observations[*]	7,296,566	7,296,530	7,294,355	7,296,544
Population size*	36,482,837	36,482,657	36,471,782	36,482,727
No. of observations, subpopulation*	105,606	105,570	103,395	105,584
Subpopulation size*	528,030.28	527,850.28	516,975.24	527,920.28
Design degrees of freedom	4,180	4,180	4,180	4,180
Linearized				
Total or Mean	528,030.3	4.601903	33,913.86	.0055311
Standard error	5,451.715	.0250909	422.9234	.0002311
95% confidence interval	517,342–538,718.5	4.552712–4.651095	33,084.71–34,743.01	.0050781–.0059842

HCUP National Inpatient Sample, 2012. Principal diagnosis only.

[†] One stratum was omitted because it contains no subpopulation members.

[*] Numbers of observations and population sizes exclude records with missing values.

Comparison of Estimates

Table 3 displays the estimates from the two statistical programming packages using the program code described previously. All the estimates are identical. For Stata, separate MEAN commands should be used for each variable. Otherwise Stata will exclude observations with missing values for any of the listed variables from the estimate. In this example, that would cause the estimates to differ slightly for mean length of stay and mean total charges because a few records have one but not the other of these variable missing.

Table 3. Comparison of SAS and Stata Results for Complicated Diabetes, National Inpatient Sample (NIS), 2012

	SAS		Stata	
Variable	No.	Standard Error	No.	Standard Error
Total discharges	528,030	5,452	528,030	5,452
In-hospital mortality, %	.553	.0231	.553	.0231
Mean length of stay, days	4.60	.025	4.60	.025
Mean total charges, $	33,914	423	33,914	423

Finite Population Corrections

The NIS sample contains about 20 percent of discharges from nearly 100 percent of hospitals nationwide. Analysts therefore may want to “correct” variance estimates to account for the fact that sampling error is attributable only to the remaining 80 percent of the discharge population. Hence, the finite population correction factor—the multiple of the infinite population variance—is equal to about 80 percent. This means that the standard errors reported in the above table would be multiplied by 89.4 percent (the square root of 80 percent). Although this decreases the estimated standard error by a little over 10 percent, the fpc should be applied only when inferences are being made to the specific population of patients actually hospitalized during the data year (in this case, 2012).

Analysts usually prefer not to use the fpc because they are interested in the long-run results for hospitals. For example, interest centers on the true, long-run mortality rate for a hospital rather than on the mortality rate actually observed in 2012. For a more in-depth discussion of the difference between finite populations and infinite populations, users may consult the document titled *Inferences With HCUP State Databases Final Report*.[11]

Combining Multiple Years of the NIS

Users may want to combine multiple NIS datasets for their analyses. The data element “year” then should be added as a stratification variable. Otherwise, the statistical routines will treat the data as though they were sampled from a single year. AHRQ also provides “trend weights” that

[11] Houchens R. Inferences With HCUP State Databases Final Report. HCUP Methods Series Report No. 2010-05. Rockville, MD: Agency for Healthcare Research and Quality; October 12, 2010. http://www.hcup-us.ahrq.gov/reports/methods/2010_05.pdf

reduce the impact of sample design changes on trend analyses. These NIS trend weights are documented and can be downloaded from the HCUP Web site.[12]

Analyzing Subpopulations

For the NIS, interest is sometimes limited to a subpopulation of the sampled population (e.g., domain, subset, or subgroup). For example, interest might center on patients with a given medical condition such as diabetes or heart disease or on patients with certain characteristics such as males who are younger than 18 years.

Eliminating individuals outside the subpopulation from the NIS before variance estimation will yield correct means and totals; however, it can yield incorrect standard errors. In particular, incorrect standard errors could be produced if a hospital is eliminated from the sample in the process of excluding patients outside the subpopulation of interest. For example, the standard errors could be incorrect if the NIS was subset to the subpopulation of patients treated for cystic fibrosis and some hospitals had no patients with cystic fibrosis in the sample. The standard errors only will be correct if every sample hospital has at least one observation in the subset; that is, every hospital treated at least one patient with this condition.

Standard errors will be calculated appropriately if all of the NIS observations are retained in the analysis and the subpopulations are defined by variables in the DOMAIN statement in SAS or by the SUBPOP option in Stata. For example, an indicator variable could be created equal to 1 for patients with cystic fibrosis and equal to 0 for all other patients. This variable then could be used with the DOMAIN statement or with the SUBPOP option. This was the method used to illustrate the diabetes analysis in this report.

One of the difficulties analysts will face with this approach is the requirement to perform analyses on the entire sample. The NIS contains over 7 million observations. Therefore, compared with the subsetting approach, this approach will require more disk space and more central processing unit (CPU) time for analyses of subpopulations. To address this difficulty, we suggest an alternative approach of using a subset of the NIS for the subpopulation of interest and then augmenting that subset with an extra “dummy” observation for each hospital. For the 2012 NIS, this adds 4,378 observations—one for each hospital in the NIS. These additional observations induce the programs to use the correct formula for calculating standard errors. The program code for SAS and Stata is contained in Appendix A.

DISCUSSION

Alternative Concepts of Variance

Sometimes analysts require variance calculations based on finite sample theory. According to this theory, the intent of the estimation process is to obtain estimates that are precise representations of the nationwide population of patients treated during a specific year. In the context of the NIS, any estimates that attempt to accurately describe characteristics (such as expenditure and utilization patterns or hospital market factors) and interrelationships among characteristics of hospitals and discharges during a specific year should be governed by finite sample theory.

[12] Healthcare Cost and Utilization Project (HCUP). HCUP NIS Trend Weights. Rockville, MD: Agency for Healthcare Research and Quality; May 2015. http://hcup-us.ahrq.gov/db/nation/nis/trendwghts.jsp

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

Under the finite population model, the variances of estimates approach zero as the sampling fraction approaches one, because (1) the population is defined at that point in time and (2) 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.

Estimation Techniques

Different methods are used for calculating variances under the two sample theories. Under the superpopulation (stochastic) model, procedures have been developed to draw inferences using weights from complex samples.[13] In this context, the survey weights are not used to weight the sampled cases to the universe, because the universe is conceptually infinite in size. Instead, these weights are used to produce unbiased estimates of parameters that govern the superpopulation.

In addition to the methods shown in this report, 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 then can be calculated from the validation data. For example, it is well known that the percentage of variance explained by a regression, R^2, generally is overestimated by the data used to fit a model. The regression model could be estimated from a training subsample and then applied to the validation subsample. The squared correlation between the actual and predicted value in the validation subsample is an unbiased estimate of the model's true explanatory power when applied to new 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 10 equal-sized subsets. The estimation would take place in 10 iterations. At each iteration, the outcome of interest is predicted for one-tenth of the observations by an estimate based on a model fit to the other nine-tenths of the observations. Unbiased estimates of error variance then are obtained by comparing the actual values to the predicted values calculated in this manner.

Longitudinal analyses of hospital-level outcomes may be biased if they are based on any subset of NIS hospitals limited to continuous NIS membership. 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.

[13] Potthoff RF, Woodbury MA, Manton KG. "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. Journal of the American Statistical Association. 1992;87(418):383-396.

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 while also incorporating data from all hospitals in the sample during the study period.

CONCLUSIONS

We found that the two statistical packages, SAS and Stata, produced identical or very similar values for weighted sample statistics, including sample variances. Other mature statistical packages could be used, such as IBM SPSS Statistics or the R language, and we are confident that they would be equally effective.

APPENDIX A:
CODE FOR ANALYZING SUBPOPULATIONS USING SAS AND STATA

The program code in this appendix yields correct estimates of standard errors based on subsets of the NIS. These estimates can be used when computing constraints prevent use of the entire NIS.

The generally recommended approach for calculating standard errors—using the entire NIS—was illustrated for the diabetes subpopulation analyses shown in the main body of this report. The example in this appendix first subsets the NIS to patients with diabetes, then augments this subset with a “dummy” observation for each NIS hospital to ensure that the proper formula is used to calculate standard errors. This approach “tricks” the software into believing that all NIS hospitals are in the analysis, even though not all hospitals may contribute discharges.

SAS Programming Statements

The following are the SAS programming statements for the example described above:

```
/* Create analysis file. */
libname IN “LOCATION OF nis FILE” ;

/* Subset diabetes cases. */
data DIABETES ;
   set IN.NIS_2012_CORE ;
   if DXCCS1 = 50 ;
   DISCHGS = 1 ;
run ;

/* Augment the diabetic subset with hospital-level observations. */
data COMBINED ;
   set DIABETES
      IN.NIS_2012_HOSPITAL(in=INHOSP KEEP=HOSP_NIS NIS_STRATUM) ;
   INSUBSET = 1 ;
   if INHOSP then do ;
      INSUBSET = 0 ;    * ASSIGN A VALUE OUTSIDE THE SUBSET ;
      DISCWT = 1 ;      * ASSIGN A VALID WEIGHT ;
      DIED = 0 ;        * SET ANALYSIS VARIABLES TO ZERO ;
      DISCHGS = 0 ;
      LOS = 0 ;
      TOTCHG = 0 ;
      FEMALE = 0 ;
   end ;
run ;

/* Obtain subpopulation estimates. */

/* Note: If finite population estimates of standard errors are    */
/* desired, then the PROC SURVEYMEANS statement could include the */
/* option "RATE = .20", which gives the approximate hospital      */
/* sampling rate in each stratum.                                 */

proc surveymeans data= COMBINED sum std mean stderr /*rate = .20*/ ;
   weight DISCWT ;
   cluster HOSP_NIS ;
   strata NIS_STRATUM ;
   var DISCHGS LOS DIED TOTCHG ;
   domain INSUBSET ;
run ;
```

The desired statistics correspond to the domain with INSUBSET = 1. Subgroups within the diabetes subpopulation (e.g., males and females) also can be analyzed using the domain statement. For example, if estimates are desired separately for male and female discharges with diabetes, female = 0 can be assigned to the hospital-level observations, and the following statement can be added to PROC SURVEYMEANS:

```
domain INSUBSET * FEMALE ;
```

Again, the desired statistics correspond to the domains with INSUBSET = 1.

Stata Programming Statements

In Stata, the hospital-level observations can be assigned zero weights to generate appropriate standard errors. Unlike SAS, it is not necessary to create a special domain variable such as INSUBSET for Stata.

The following are SAS programming statements to prepare the data for Stata for the example described above:

```
/* Using SAS, create an ASCII file for use by STATA. */
libname IN “location of NIS file” ;

/* Subset diabetes cases. */
data DIABETES ;
   set IN.NIS_2012_CORE ;
   if DXCCS1 = 50 ;
   DISCHGS = 1 ;
run ;

/* Augment the diabetic subset with hospital-level observations. */
data _null_ ;
   set DIABETES
      IN.NIS_2012_HOSPITAL(IN=INHOSP KEEP=HOSP_NIS NIS_STRATUM) ;
   if INHOSP then do ;
      DISCWT = 0 ; * Assign zero weights ;
      DIED = 0 ;   * Set analysis variables to zero ;
      DISCHGS = 0 ;
      LOS = 0 ;
      TOTCHG = 0 ;
      FEMALE = 0 ;
   end ;
   if LOS < 0 then LOS = . ;
   if DIED < 0 then DIED = . ;
   if TOTCHG < 0 then TOTCHG = . ;
   if FEMALE < 0 then FEMALE = . ;
   file FILEREF ;
   PUT nis_stratum “,” hosp_nis “,” died “,” los “,” dischgs “,”
       totchg “,” female “,” discwt ;
run ;
```

The following are the Stata programming statements for the example described above:

```
/* Obtain STATA estimates. */
/* Note: Stata commands should be entered in lowercase text. */

infile int    nis_stratum  ///
       long   hosp_nis     ///
       int    died         ///
       long   los          ///
       int    dischgs      ///
       double totchg       ///
       int    female       ///
       double discwt       ///
       using "dataset name"
svyset hosp_nis [pw=discwt], strata(nis_stratum)
svy: total dischgs
svy: mean los
svy: mean totchg
svy: ratio died dischgs
```

Subpopulations of the diabetic subset also can be analyzed using the SUBPOP option, which is similar to the Stata code shown for the diabetes example in the main body of this report. For example, if separate estimates are desired for female discharges with diabetes, the following statements could be substituted:

```
svy: total DISCHGS, subpop(FEMALE)
svy: mean LOS, subpop(FEMALE)
svy: mean TOTCHG, subpop(FEMALE)
svy: ratio DIED DISCHGS, subpop(FEMALE)
```

These commands will produce estimates for female discharges with diabetes (observations with female = 1). If estimates are desired for male discharges with diabetes, then an indicator variable for males should be used.