Memorandum

DATE: July 1, 2004
TO: Users of the HCUP Nationwide Inpatient Sample (NIS) and Kids' Inpatient Database (KID)
SUBJECT: Corrections to instructions on how to calculate variances (standard errors)

We have recently uncovered a problem in the method we describe for calculating variances using the HCUP Nationwide Inpatient Sample (NIS). The same methods can be applied to Kids' Inpatient Database (KID), so this information is relevant for KID users as well. The method for calculating variances is outlined in the special report entitled "Calculating Nationwide Inpatient Sample Variances". The report is available online at the HCUP User Support Web site http://www.hcup-us.ahrq.gov. This memo describes the nature of the problem.

For the NIS and the KID, interest is sometimes limited to a subset of the sampled population. For example, interest might center on patients with a given medical condition like diabetes or cystic fibrosis or on patients with certain characteristics like males under the age of 18. We have recently determined that eliminating individuals outside the subpopulation from the NIS or KID before variance estimation will yield correct means and totals, but it can possibly yield incorrect standard errors.

Incorrect standard errors will be produced if a hospital gets eliminated from the sample in the process of excluding patients outside the subpopulation of interest. In other words, the standard errors from a subset will be correct if every sample hospital has at least one observation in the subset. For example, if every hospital treated at least one cystic fibrosis patient, then the standard errors produced by subsetting the data would be correct. However, the standard errors will be incorrect if the NIS was subset to only those patients treated for cystic fibrosis and some NIS hospitals had no cystic fibrosis patients in the sample.

Standard errors will always be calculated appropriately if all of the data are retained in the analysis and the subsets are defined by variables in the DOMAIN statement in SAS, by the SUBGROUP statement in SUDAAN, and by the SUBPOP statement in Stata. For example, an indicator variable could be created equal to 1 for cystic fibrosis patients and equal to 0 for all other patients, which could then be used in the DOMAIN, SUBGROUP, or SUBPOP statements.

The original NIS Variance report showed an example analysis of the subpopulation of diabetic patients from the NIS. The example program code first subsets all diabetic patients from the NIS, then uses SAS, SUDAAN, and Stata to analyze the subsets. In this particular case, every hospital contained at least one diabetic patient; consequently, the standard errors shown in the original report were correct for that subpopulation. However, they would not have been correct for a subpopulation defined by a rarer condition. Therefore, the report has been revised to use the entire NIS in the examples.

However, because of the large number of records in the NIS and the KID, most analyses on subsets of patients will not retain all discharges. Therefore, Appendix B in the revised report provides an approach that will allow use of subsets of cases while still calculating variances correctly. The programming code Appendix B basically "tricks" the software into believing that all NIS hospitals are in the analysis, even though not all hospitals may contribute discharges to the analysis.

You may well wonder how erroneous the standard errors were that you may have produced by subsetting in previous studies. It depends partly on the number of hospitals that got excluded from

the sample. The degree of error will be zero for subpopulations defined by the stratification variables (e.g., region). The degree of error was probably zero or very slight for high-frequency subpopulations (like diabetes). The degree of error was probably greatest for rare subpopulations. The difference between the "correct" standard error (using all the data) and the "incorrect" standard error (using only a subset of the data) could have been large or small and it could have been positive or negative. The only way to know for sure is to recalculate the standard error using the full data approach.

We apologize for any inconvenience this oversight may have caused. If you have any questions, please contact us at hcup@arhq.gov.