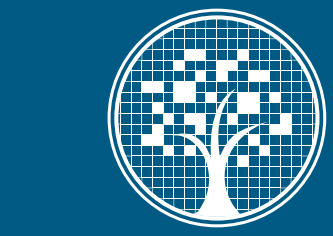

HCUP Methods Series

Calculating Nationwide Readmissions Database (NRD) Variances

Report # 2017-01

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 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: Yoon F, Sheng M, Jiang HJ, Steiner CA, Barrett ML. *Calculating Nationwide Readmissions Database (NRD) Variances.* HCUP Methods Series Report # 2017-01 ONLINE. January 24, 2017. U.S. Agency for Healthcare Research and Quality. Available: http://www.hcup-us.ahrq.gov/reports/methods/methods.jsp.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The Nationwide Readmissions Database (NRD) is part of the Healthcare Cost and Utilization Project (HCUP) that is sponsored by the Agency for Healthcare Research and Quality. The NRD addresses a large gap in health care data—the lack of nationally representative information on hospital readmissions for all types of payers and the uninsured. The NRD was created to enable analyses of national readmission rates and to support public health professionals, administrators, policymakers, and clinicians in their decisionmaking.

The NRD is drawn from HCUP State Inpatient Databases and poststratified to reflect the target universe of inpatient discharges treated at community hospitals in the United States that are not rehabilitation or long-term acute care facilities. The target universe is based on the American Hospital Association Annual Survey of Hospitals.

The NRD is designed to be flexible to various types of analyses of national readmissions for all types of payers and the uninsured. Outcomes of interest include national readmission rates, reasons for returning to the hospital for care, and the hospital costs for discharges with and without readmissions.

Importantly, analyses must incorporate statistical precision of national estimates based on the NRD design, specifically by appropriate calculation of variances. The purpose of this report is to guide analysts in calculating variances for estimates of readmission outcomes using the NRD.

- We illustrate variance calculations and describe their implementation in standard software routines, based on important concepts for discharge weights, poststratification and clustered sampling.
- We also recommend that analysts consider, based on the research question, whether a finite or superpopulation framework is suitable for producing estimates and their variances. These two models are defined and compared.
- Finally, we provide sample programs for analyzing NRD readmissions data from patients admitted to the hospital with a principal diagnosis of acute myocardial infarction.

This report is an overview of the primary issues and not intended to be a comprehensive account of all possible scenarios in which variances of estimated NRD outcomes should be calculated. We provide additional resources for understanding the detailed issues surrounding variance calculations for the NRD and other related topics.

INTRODUCTION

The Nationwide Readmissions Database (NRD) is an all-payer hospital inpatient database that is drawn from the Agency for Healthcare Research and Quality (AHRQ) Healthcare Cost and Utilization Project (HCUP) State Inpatient Databases (SID). The SID contain verified patient linkage numbers that can be used to track a person across hospitals within a State while adhering to strict privacy guidelines.

The NRD can be used to generate national estimates of readmissions. Outcomes of interest include national readmission rates, reasons for returning to the hospital for care, and hospital costs for discharges with and without readmissions. The NRD supports analysis on hospital readmissions for all types of payers and for the uninsured. It is designed to be flexible to various types of analyses of readmissions. However, the NRD is not designed to support regional, State-, or hospital-specific readmission analyses.

The NRD is a sample of convenience from the SID and not a sample of hospitals or discharges. To reflect the target universe, weights were calculated and applied through poststratification by hospital and discharge characteristics. Standard error calculations should take into account the poststratification weights and clustering of discharges within hospitals.

It is important for researchers to calculate a measure of precision for national estimates based on the NRD. The purpose of this report is to provide a review of the NRD design and basic guidelines for variance calculations using statistical software. The introduction and second section describes the design of the NRD, key readmission data elements, and implications for missing data. The third section discusses statistical issues for variance calculations, including theoretical illustrations based on the NRD design. The fourth section (and Appendix) provides examples of variance calculation in common statistical software packages such as SAS® (SAS Institute Inc.), Stata® (StataCorp LP), and R© (The R Foundation). The report concludes with recommendations and further considerations for variance calculation in common applications of the NRD, such as subpopulation analysis and risk adjustment. Detailed software code is provided in the Appendix.

Sample Design

The target universe was limited to inpatient discharges treated at community hospitals in the United States that were not rehabilitation or long-term acute care (LTAC) facilities. Information on the target universe was available from the American Hospital Association (AHA) Annual Survey of Hospitals. The AHA Survey yields data on the number of inpatient discharges and hospital characteristics such as ownership, number of beds, and location.

The SID contain inpatient discharges for all community hospitals provided by HCUP State Partners. The AHA defines *community* hospitals as “nonfederal, short-term general, and special hospitals, including special children’s hospitals, whose facilities and services are available to the public.”[1] Specialty hospitals included in the AHA definition of community hospitals are obstetrics-gynecology, ear-nose-throat, short-term rehabilitation, and orthopedic. Also included in the SID are pediatric institutions, LTAC facilities, cancer hospitals, psychiatric hospitals, public hospitals, and academic medical centers.

[1] American Hospital Association Data Viewer: Glossary. American Hospital Association Web site. https://www.ahadataviewer.com/glossary. Accessed December 5, 2016.

The sampling frame for the NRD was limited to SID discharges for patients treated at community hospitals that were not rehabilitation or LTAC facilities. All of the discharges in the sampling frame were included, making the NRD a sample of convenience from the SID. Sampling discharges or hospitals was not recommended because the sample needed to balance the database's ability to estimate readmissions for common conditions such as chronic illnesses with the ability to estimate readmissions for rare diseases such as sickle cell anemia. Developing the database using a 100 percent sample allows researchers to study both all-cause and condition-specific readmissions.

All discharges from the SID are included in the NRD except the following:

- Discharges from patients with an age of 0 with patient linkage numbers inconsistently reported for this age[2]
- Discharges with missing patient linkage numbers
- Discharges with questionable patient linkage numbers, defined as 20 or more discharges in a year, hospitalized after discharged dead, and overlapping inpatient stays
- Discharges from hospitals with more than 50 percent of their total discharges excluded for any of the above causes, because patients treated at these hospitals may not be reliably tracked over time
- Discharges from short-term rehabilitation hospitals and LTAC facilities.

After exclusions, the NRD contains more than 80 percent of SID discharges from the participating States. Unweighted, the NRD contains approximately 15 million discharges each year. Weighted, it represents 35 million discharges in the United States.

Poststratification for Weighting

Poststratification for the purpose of weighting compensates for any over- or under-represented types of hospitals and discharges in the sampling frame (the NRD) with respect to the distribution in the target universe (AHA data). The NRD was poststratified by hospital and discharge characteristics. Hospital characteristics for poststratification to the target universe included Census region, urban/rural location, teaching status, hospital size, and hospital control. Discharge characteristics included patient sex and age category. For more information, see the report titled *Introduction to the HCUP Nationwide Readmissions Database (NRD) 2014*.[3]

The target-universe discharge counts within the strata were derived from all SID discharges from all HCUP Partners, unless there were missing hospitals. If there were hospitals missing from the stratum according to the AHA, then the target universe total included SID discharges for all available hospitals plus the AHA discharge counts for the missing hospitals. This approach took advantage of the fact that the SID included over 95 percent of discharges from community hospitals that are not rehabilitation or LTAC hospitals in the United States.

---

[2] The number of States in the NRD that have patients who are younger than 1 year varies by data year (e.g., 9 of the 21 SID in 2013 and 12 of 22 SID in 2014). The weights for pediatric discharges often are higher than those for adult discharges.

[3] Agency for Healthcare Research and Quality. Introduction to the HCUP Nationwide Readmissions Database (NRD) 2014. Rockville, MD: Agency for Healthcare Research and Quality; November 2016. https://www.hcup-us.ahrq.gov/db/nation/nrd/Introduction_NRD_2014.pdf

To determine discharge-level weights, we calculated the number of discharges for the target universe and for the sampling frame by strata defined by hospital characteristics (Census region, urban/rural location, hospital teaching status, size of the hospital defined by the number of beds, and hospital control) and by patient characteristics (sex and five age groups). Within each stratum s, each NRD discharge received the weight N_s/n_s which is the ratio of the number of discharges in the target universe, N_s, to the number in the NRD sample, n_s. Therefore, each discharge's weight is equal to the number of discharges it represents in the universe in stratum s during the given NRD year.

A potential limitation, which is not addressed in illustrating variance calculations in this report, is that the 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 HCUP. To the extent that poststratification does not fully account for differences between States in the frame and other States, this could lead to biased estimates for readmission analyses. Consequently, users should compare estimates from the NRD to other benchmarks whenever they are available.

Analytic Considerations

Defining Index Events and Readmission Rates

The NRD was designed to support many different types of readmission analyses. Analysts can use the information contained in the NRD to define the index event and readmission specific to their topic of interest. Analytic specifications include defining the index event, specifying the criteria for a readmission, selecting the appropriate time period to qualify the readmission, and reporting readmission rates.

When creating the NRD, no attempt was made to determine whether repeat visits were related or unrelated. This decision is left to the analyst using the NRD. Although the definition of *readmission rate* seems simple—number of readmissions divided by number of cases followed—our research into readmission rates showed no such standard definition. In some cases, the unit of observation was a patient; in others, the unit of observation was index events, and individual patients were counted more than once. Some studies focused on the first readmission following an index event, whereas others counted all readmissions. The definitions of the readmission rate were specific to the purpose of the analyses.

The example code in the Appendix shows how to specify 30-day readmissions following discharge from treatment of acute myocardial infarction (AMI). For further details about how to specify readmission outcomes and alternative specifications, we refer the reader to the *Introduction to the Nationwide Readmissions Database (NRD)* report, the Nationwide Readmissions Database Tutorial,[4] and other NRD-related HCUP Methods Series Reports.[5]

[4] Agency for Healthcare Research and Quality. Nationwide Readmissions Database Tutorial. Healthcare Cost and Utilization Project Web site. http://www.hcup-us.ahrq.gov/tech_assist/nrd/index.html. Accessed December 5, 2016.

[5] Agency for Healthcare Research and Quality. NRD Related Reports. Healthcare Cost and Utilization Project Web site. http://www.hcup-us.ahrq.gov/db/nation/nrd/nrdrelatedreports.jsp. Accessed December 5, 2016.

Combining Multiple Years of the NRD

The NRD are annual files containing inpatient records for patients discharged in a calendar year. The files include patients admitted in the prior year and discharged in the current year, while excluding patients admitted to a hospital in the current year but discharged in the next year. Therefore, readmissions for patients admitted in the latter part of the year may not be captured if the subsequent admission crossed into the next year. In addition, one year of discharge data probably is an insufficient length of time for examining readmissions that are more than 90 days apart.

The 2013 and 2014 NRD cannot be combined to create a 24-month database, because the patient linkage numbers (NRD_VISITLINK) do not track the same person from 2013 into 2014; in addition, the hospital identifiers (HOSP_NRD) do not track the same hospitals between 2013 into 2014. Each year of the NRD must be considered as a separate sample.

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 readmission outcomes for discharges with missing values is systematically different from those 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 titled *Missing Data Methods for the NIS and SID*.[6] Another strategy is to use sample weight adjustments to compensate for missing values.[7] 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.

[6] 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

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

VARIANCE CALCULATIONS FOR POSTSTRATIFIED, CLUSTERED DATA

To accurately calculate variances from the NRD, appropriate statistical software and techniques must be used. A multitude of outcomes can be analyzed using computer programs such as SAS, Stata, and R, which calculate statistics and their variances from sample survey data. These programs use general methods of variance calculations (e.g., Taylor-series expansion, or jackknife and balanced half-sample replications) that take into account the sampling design.

In this section, we discuss and illustrate how to derive some basic calculations that incorporate the NRD design elements for poststratification and clustering, as well as considerations for finite versus superpopulation models. The reader can safely forgo the detailed illustrations and formulas that appear later, but all researchers are encouraged to review the following section about finite population and superpopulation models prior to embarking on an analysis using NRD data.

Finite Population and Superpopulation Models

For cross-sectional, nationwide estimates of NRD readmission outcomes specific to a given year and discharge population, variance calculations are based on finite population theory. According to this theory, the intent of the estimation process is to obtain precise representations of the nationwide population at a specific point in time. In the context of the NRD, any estimates that attempt to describe readmission outcomes during a specific year should be governed by finite population theory. Examples would be cross-sectional estimates of readmission rates that are specific to the discharge population in that year.

Similarly, any estimates that attempt to accurately describe characteristics (such as readmission rates) and inter-relationships among characteristics of hospitals and discharges during a specific year should be governed by finite population theory. 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.

On the other hand, analysts usually are interested in the long-run results for hospitals. For example, policy interest may be focused on the true, long-run readmission rates in the United States, rather than on the rates actually observed in 2014. 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 NRD finite population (hospitals and discharges) and time period from which the *sample* was drawn than they are in hypothetical characteristics of a conceptual "superpopulation" (infinite in size) from which any particular NRD (finite) population may 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 superpopulation model, procedures have been developed to draw inferences using weights from complex samples.[8] In this context, the discharge weights are not used to weight the NRD discharges to the superpopulation because it is conceptually infinite in size. Instead, these weights are used to produce unbiased estimates of parameters that govern the superpopulation.

In the following illustrations, we assume the superpopulation framework for NRD analyses (e.g., long-run readmission rates); this is the default setting for most software routines. The analyst is reminded that the alternative decision to adopt the finite population framework will depend on whether analysis focuses on the specific population of discharges in the target universe in a given NRD year.

For a more in-depth discussion of the difference between finite and superpopulation models, users may want to consult the document titled *Inferences With HCUP State Databases Final Report*.[9]

Variance of Poststratified Sample Mean

For illustration, we show basic "behind the scenes" calculations that statistical programs make for variances based on poststratified sample designs in a superpopulation framework. The reader may safely go to the example code in the next section of this report without understanding the technical details of this example.

The NRD is a poststratified convenience sample of all discharges from all hospitals in the sampling frame. For the present illustration, we ignore the clustering of discharges within hospitals. However, in most settings, the analyst should account for clustering, which tends to induce dependence among discharges within hospitals and thereby reduce the effective sample size in calculating standard errors for certain readmission outcomes that are not population totals. We further describe the clustering issue in the next section.

In the NRD, the stratum-specific counts of hospitals and discharges are known in the target universe. Using these counts, we calculate and apply poststratification weights so that the stratum-specific counts in the NRD sample equal those in the target universe. Those weights, as previously described, are the relative number of discharges in the target universe that are represented by each discharge in the NRD sample. With discharge weights within stratum s defined as N_s/n_s, the poststratification estimator for the population mean is given by the following formula:

$$\overline{y}_{post} = \sum_{s=1}^{S} \frac{N_s}{N} \overline{y}_s = \frac{1}{N} \sum_{s=1}^{S} \sum_{d \in A_s} \frac{N_s}{n_s} y_{sd}$$

where N is the size of the target universe; $s = 1, \ldots, S$ indexes the strata; N_s is the stratum size in the target universe; n_s is the stratum size in the NRD sample; y_{sd} is the observation for

[8] 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.

[9] 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

discharge d in stratum s; A_s is the collection of discharges in stratum s; and $\overline{y}_s$ is the estimate of the mean in stratum s.

The standard approach treats stratum sample sizes n_s as random variables. Poststratification estimators customarily are defined on the sample of respondents, whose propensity to respond is modeled stochastically. Given the stratum sizes in the sample (and known sizes in the target universe), a conditional variance can be estimated. This variance is based on a superpopulation model for (1) a common mean within each stratum, (2) correlation in outcomes within clusters, and (3) independence in outcomes between clusters.[10] In contrast, the unconditional variance essentially integrates the conditional expression over all possible configurations of the stratum sample; the unconditional variance is "slightly larger."[11] For a simple random sample (for illustration only), the conditional variance is given by the following:

$$\mathrm{Var}(\overline{y}_{post}|n_s, s = 1, \dots, S) = \sum_{s=1}^{S} \left(\frac{N_s}{N}\right)^2 \left(1 - \frac{n_s}{N_s}\right)\left(\frac{V_s}{n_s}\right)$$

where V_s is the within-stratum population variance estimated by $\frac{1}{n_s - 1}\sum_{d \in A_s}\left(y_{sd} - \overline{y}_s\right)^2$.

Accounting for Clustered Data

In the NRD, discharges are naturally clustered within hospitals. This clustering tends to induce dependence in readmission outcomes, such as underlying rates 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. As such, variance calculations should account for the dependencies; importantly, the effective sample size will be smaller, because there is inherently less information in correlated (as opposed to uncorrelated) outcomes.

We illustrate variance calculations for clustered NRD outcomes using analogous calculations from the two-stage sampling framework. This framework can be used to calculate the variance of a NRD estimate, as it separates the hospital and discharge components and thereby closely resembles the NRD sample design. Our illustrations here easily generalize to both the finite and superpopulation frameworks, as they are based on the finite population correction (fpc). In the finite population framework, the fpc is a number between 0 and 1 that represents how much information we have from the target universe; specifically, it is the ratio of NRD sample-stratum sizes to population-stratum sizes. For superpopulation models, the fpc is zero, thereby indicating that the target universe is theoretically infinite in size.

Suppose the analyst is conducting inference on the total of a readmission outcome y_{shd}, such as total cost of care, of discharge d in hospital h in stratum s. To our previous formulation, we include indexing by hospitals in order to illustrate the clustering of outcomes within them (index by $h = 1, \dots, H_s$) and we denote by B_s the collection of hospitals in stratum s. We also slightly revise the notation A_h to refer to the collection of discharges within hospital h. Following the example of cost of care for NRD admissions, the total cost of care in the NRD sample is given by this formula:

[10] Valliant R. Poststratification and conditional variance estimation. Journal of the American Statistical Association. 1993;88(421): 89-96.

[11] Lohr SL. Sampling: Design and Analysis. Pacific Grove, CA: Brooks/Cole Publishing Company; 1999.

$$T_{post} = \sum_{s=1}^{S} \sum_{h \in B_s} \sum_{d \in A_h} \frac{N_s}{n_s} y_{shd}$$

From the NRD sample, we can estimate the variance of T_{post} as follows:

$$\mathrm{Var}\left(T_{post} | n_s, s = 1, \ldots, S\right) = \sum_{s=1}^{S} \left\{ (1 - f_s) \cdot m_s U_s + f_s \cdot \sum_{h \in B_s} (1 - f_{sh}) \cdot n_{sh} V_{sh} \right\}$$

where f_s is the hospital-level fpc indicating the proportion of hospitals sampled in stratum s, m_s is the number of NRD hospitals in stratum s, f_{sh} is the discharge-level fpc, and n_{sh} is the number of NRD discharges from the target universe in hospital h in stratum s. The first component of variance U_s is generated by the selection of hospitals (in stratum s) and is given by the following:

$$U_s = \frac{1}{m_s - 1} \sum_{h \in B_s} \left(\sum_{d \in A_h} \frac{N_s}{n_s} y_{shd} - \frac{1}{m_s} \sum_{h \in B_s} \sum_{d \in A_h} \frac{N_s}{n_s} y_{shd} \right)^2$$

In this formulation, the first summation over $h \in B_s$ yields the sum of squared deviations of individual hospital totals from the mean hospital total; heuristically, this is similar to the calculation of the variance of any sample statistic. This first component of variance in $\mathrm{Var}\left(T_{post} \middle| \cdot\right)$ is governed by f_s, which is the proportion of hospitals within population-stratum s that were included in the NRD sample-stratum s. In an infinite population context, the analyst would set $f_s = 0$, thereby assuming the NRD hospitals were sampled at an indefinite point in time (i.e., irrespective of the year of data collection).

In similar fashion, the second component of variance V_{sh} arises from the selection of discharges within hospitals. It can be calculated by the following:

$$V_{sh} = \frac{1}{n_{sh} - 1} \sum_{d \in A_h} \left(\frac{N_s}{n_s} y_{shd} - \frac{1}{n_{sh}} \sum_{d \in A_h} \frac{N_s}{n_s} y_{shd} \right)^2$$

Likewise, for the variance that comes from discharge selection, the summation over discharges in hospital h (indexed by $d \in A_h$) yields the sum of squared deviations of weighted-discharge totals from the mean-weighted discharge total in hospital h in stratum s. For a NRD analysis based on the finite population model, the analyst may set the sampling fraction $f_{sh} = 1$, because all discharges in hospital h in stratum s were selected; then, the second term of $\mathrm{Var}\left(T_{post} \middle| \cdot\right)$ disappears, because there is no uncertainty surrounding the hospital totals in the NRD sample (as all discharges within each hospital were sampled). On the other hand, for inference using a superpopulation approach, the analyst would set $f_{sh} = 0$, so that the second term contributes the uncertainty arising from selecting a random sample of discharges from the superpopulation.

EXAMPLES OF VARIANCE CALCULATIONS

The NRD discharge weights are needed to calculate national estimates of readmission counts and rates. In most cases, computer programs are readily available to perform these calculations in weighted analyses.[12] They incorporate weights and can specifically calculate statistics and their standard errors from survey data. For example, SAS Version 8 or later contains procedures (PROC SURVEYMEANS and PROC SURVEYREG) for calculating statistics based on specific sampling designs. The Stata statistical software package calculates numerous statistics arising from stratified, cluster sampling designs. Finally, survey analyses can be conducted in the R programming language, which presents several libraries for complex survey data (the primary one being the *survey* package).

The example analysis is for a subpopulation of the 2014 NRD that includes patients admitted to hospitals with principal diagnosis of AMI, defined by diagnosis codes from the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). Detailed code in the Appendix shows how index and readmission events were identified.

Note that the example programs shown here use the entire NRD—that is, every hospital in the NRD is included in the analysis, even those that do not contain discharges for the given subpopulation. 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 NRD (such as a specific subpopulation defined by condition or by patient characteristics), then refer to the Appendix, which provides example code using a smaller subset of the data.[13]

Importantly, for a readmissions analysis, the entire NRD should be used to first identify index and candidate readmission events; that is, subsetting must be done *after* this step, otherwise index and readmission events may be inadvertently excluded and thereby counted incorrectly.

To obtain estimates, we created an indicator variable to identify the subset of discharges with AMI for the analysis and thereby create the analytic dataset. The programming statements below are based on this dataset that contains index and readmission events (*IndexEvent* and *Readmit*), hospital identifiers (HOSP_NRD), stratum (NRD_STRATUM), and discharge weights (DISCWT). They show the basic steps for specifying the sampling design and then calculating variances of NRD estimates. For more details in the Appendix, the SAS, Stata, and R program codes include code to flag index and readmission events, as well as code to conduct subpopulation analysis. In these examples, the following conventions apply:

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

---

[12] Carlson BL, Johnson AE, Cohen SB. An evaluation of the use of personal computers for variance estimation with complex survey data. Journal of Official Statistics. 1993;9(4):795-814.

[13] Houchens RL and Ross DN. Calculating National Inpatient Sample (NIS) Variances for Data Years 2012 and Later. HCUP Methods Series Report # 2015-09. Rockville, MD: Agency for Healthcare Research and Quality; December 10, 2015. https://www.hcup-us.ahrq.gov/reports/methods/2015-09.pdf

Example Programming Statements

These programs illustrate only the sampling design and variance calculations for the AMI analysis, based on a user-defined analytic dataset; for detailed code to identify AMI index and readmission events and to subset the data to analyze subpopulations, please see the Appendix.

SAS

```
/* Specify the sampling design with sampling weights DISCWT,   */
/*   hospital clusters HOSP_NRD and stratification NRD_STRATUM */
/* Calculate national estimates */
proc surveymeans data = sasdata.readmissions_sql sumwgt sum mean ;
    cluster HOSP_NRD ;
    strata NRD_STRATUM ;
    weight DISCWT ;
    var Readmit ;
    /* We can subset the survey design object so that the target of */
    /* inference is the population of AMI index events only         */
    domain IndexEvent;
    ods output domain = readmissionRates ;
    format IndexEvent indexEvent. ;
run;
```

Stata

```
/* Specify the sampling design with sampling weights DISCWT,   */
/*   hospital clusters HOSP_NRD and stratification NRD_STRATUM */
svyset HOSP_NRD [pw = DISCWT], strata(NRD_STRATUM) ;

/* Calculate national estimates */
svy: total readmit ;
svy: mean readmit ;

/* We can subset the survey design object so that the target of */
/* inference is the population of AMI index events only         */
svy: total readmit, subpop(indexevent) ;
svy: mean readmit, subpop(indexevent) ;
```

R

```
install.packages("survey"); library(survey)
# Specify the sampling design with sampling weights DISCWT,
#   hospital clusters HOSP_NRD, and stratification NRD_STRATUM
strnrddsgn <- svydesign(ids = ~HOSP_NRD, weights = ~DISCWT,
      strata = ~NRD_STRATUM, data = core)

# Calculate national estimates
svytotal(~readmit, strnrddsgn)
svymean(~readmit, strnrddsgn)

# We can subset the survey design object so that the target of
#   inference is the population of AMI index events only
strnrddsgn.sub <- subset(strnrddsgn, IndexEvent)
svytotal(~readmit, strnrddsgn.sub)
svymean(~readmit, strnrddsgn.sub)
```

Comparison of Estimates From Different Software Routines

Table 1 contains readmission outcomes from analyses using the three software packages for all NRD discharges and for discharges with AMI index events. The estimates are nearly identical, within rounding for Taylor series-based approximation of standard errors.

Table 1. Output for 30-day All-Cause AMI Readmissions

Analytic Sample	Software	Mean	Standard Error	Sum	Standard Error
All NRD discharges	**SAS**	0.001918	4.1×10^{-5}	67,717	1,624.8
	Stata	0.001918	4.1×10^{-5}	67,717	1,624.8
	R	0.001918	0	67,717	1,624.8
Discharges with AMI index events	**SAS**	0.1417	0.0013	67,717	1,624.8
	Stata	0.1417	0.0013	67,717	1,624.8
	R	0.1417	0.0013	67,717	1,624.8

Abbreviations: AMI, acute myocardial infarction; NRD, Nationwide Readmissions Database.

Source: Agency for Healthcare Research and Quality, Healthcare Cost and Utilization Project Nationwide Readmissions Database, 2014.

FURTHER CONSIDERATIONS

The previous discussion focused on variance calculations in a basic NRD analysis—specifically how to incorporate the sample design through discharge weights, poststratification, and hospital clustering. We provide some additional considerations for analyses with specifications that generally are not covered by the previous discussion.

Analyzing Subpopulations

The analyst may wish to study readmissions in a subpopulation defined by clinical domains, demographics, or any other discharge or hospital features. For example, interest might center on readmission rates for patients admitted for treatment of heart disease or on patients with certain characteristics such as males aged 65 years and older.

Eliminating individuals outside the subpopulation from the NRD before variance estimation will yield correct means and totals; however, it also 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 NRD was subset to the subpopulation of patients treated for AMI and some hospitals in the NRD had no patients with AMI. As an example, the 2014 NRD has 325 hospitals for which there are no patients with a principal diagnosis of AMI. The standard errors only will be correct if every hospital in the NRD 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 NRD observations are retained in the analysis and the subpopulations are defined by (1) variables in the DOMAIN statement in SAS, (2) the SUBPOP option in Stata, or (3) the *subset* function in the R *survey* package. For example, an indicator variable could be created equal to 1 for patients with AMI 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 AMI analysis in this report.

One of the difficulties analysts will face with this approach is the requirement to perform analyses on the entire sample. For example, the 2014 NRD contains almost 15 million observations. Therefore, compared with the subsetting approach, this approach will require more disk space and more processing time for analyses of subpopulations. To address this difficulty, we suggest an alternative approach of using a subset of the NRD for the subpopulation of interest and then augmenting that subset with an extra “dummy” observation for each hospital. For the 2014 NRD, this adds 2,048 observations—one for each hospital in the NRD. These additional observations induce the programs to use the correct formula for calculating standard errors; for example code, please refer to the Appendix.

It is important to note that subsetting to identify a target subpopulation must be done *after* the identification of index and candidate readmission events in the entire NRD; otherwise, readmission events (such as all-cause readmission events) may be inadvertently excluded from the analysis, thereby yielding incorrect estimates.

Risk Adjustment

Severity or risk adjustment also may be beneficial when comparing readmission rates across geographical regions, hospital types, or patient populations. Specifically, comparison of readmission outcomes can be facilitated by careful use of the NRD stratification and available

clinical covariates (e.g., major diagnostic category and Clinical Classifications Software category). A simple risk adjustment would include the age and sex of the patient. A more complex adjustment might also include comorbidities, severity classified by the 3M All-Patient Refined DRG severity score, patient income quartile, or any other factor that could considerably increase or decrease the risk of subsequent hospital care.

Although much of the current policy debate concerns the comparison of readmission rates between hospitals in order to ascertain their underlying quality (e.g., in public reporting and payment programs), we emphasize that the NRD is not designed to support comparisons of specific hospitals. Hospital identities are masked in the NRD and cannot be linked to hospital-specific external data sources such as the AHA Annual Survey Database or the Centers for Medicare & Medicaid Services public-reporting website Hospital Compare. In addition, a comparative analysis of readmission rates at the hospital level would be restricted to the finite *sample* of hospitals represented in the NRD; it would be difficult to generalize to a broader superpopulation (i.e., one that also represents hospitals that were not included in the NRD) without imposing strong sampling assumptions. For example, the analyst would need to assume that the NRD hospitals would accurately reflect the performance of other non-NRD hospitals on readmission outcomes.

Alternative Concepts of Variance and Estimation Techniques

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. For NRD analysis, the analyst first would specify and produce readmission outcomes in a separate analytic file (with one observation or outcome per row) and then split the analytic file into training and validation subsamples. 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.

APPENDIX: SAS, STATA, AND R SOFTWARE CODE FOR READMISSIONS ANALYSIS

The program code in this appendix yields correct estimates of standard errors of AMI readmission estimates based on both the full population of the NRD. A smaller subset of the data (e.g., discharges with an AMI index event) can be used when computing constraints prevent use of the entire NRD.

SAS

```
/* Data load and prep   */
/* User-defined         */
%let Path_='[project path]' ;
libname sasdata "&path_\data" ;
libname nrd "[data directory]" access=readonly ;
%let Obs_ = Max ;

/* Format index and readmission event outcomes */
options FormChar='|----|+|---+=|#/\<>*' ;
ods noptitle;
proc format;
    value indexEvent
        0 = '0: No AMI index'
        1 = '1: AMI index' ;
    value readmit
        0 = '0: No AMI readmit'
        1 = '1: AMI readmit' ;
run;

/* Identify AMI index events */
Title "1 : Index Admissions" ;
data nrd_2014_indexEvents
ReadmCandidates ( drop=DISCWT LOS NRD_STRATUM IndexEvent ) ;
    set nrd.nrd_2014_core ( obs= &obs_
                            keep= HOSP_NRD KEY_NRD DX1 DISCWT NRD_STRATUM
                                  AGE DMONTH DIED LOS NRD_VISITLINK
                                  NRD_DAYSTOEVENT DXCCS1 PRCCS: ) ;
    attrib IndexEvent length=3 label='AMI index event' ;
    IndexEvent = 0 ;

    if '41000' le DX1 le '41091' and substr( DX1, 5, 1 ) ne '2'
      and AGE ge 18
      and 1 le DMONTH le 11
      and DIED eq 0
      and not missing(LOS) then IndexEvent = 1;
    drop DX1 AGE DMONTH DIED;
    * Retain index events only ;
    if IndexEvent = 1 then output nrd_2014_indexEvents;
    output ReadmCandidates ;
run;

/* Tabulate (unweighted) index events */
proc freq data= nrd_2014_indexEvents;
    tables IndexEvent / list missing;
    format indexEvent indexEvent. ;
run ;
```

```
Title "2 : 30-day All-Cause Readmission Events" ;
/* Select all readmissions within 30 days */
proc sql ;
    create table readmissionsAll as
    select i.HOSP_NRD as HOSP_NRD_Index
        , i.KEY_NRD as KEY_NRD_Index
        , r.*
    from nrd_2014_indexEvents i                          /* Index Events    */
        inner join ReadmCandidates r                     /* Readmissions    */
            on i.NRD_VISITLINK = r.NRD_VISITLINK         /* Link patients   */
                and i.KEY_NRD ne r.KEY_NRD               /* Not a self join */
                and r.NRD_DAYSTOEVENT - ( i.NRD_DAYSTOEVENT + i.LOS )
                  between 0 and 30
                and i.indexEvent = 1
    order by i.HOSP_NRD, i.KEY_NRD, r.NRD_DAYSTOEVENT;  /* Sort by date    */
quit ;

/* Identify closest readmission if there are multiple readmission events */
data readmissionsClosest;
    set readmissionsAll ( rename=(HOSP_NRD=HOSP_NRD_Readmit
                                  HOSP_NRD_Index=HOSP_NRD
                                  KEY_NRD = KEY_NRD_Readmit
                                  KEY_NRD_Index=KEY_NRD)) ;
    by HOSP_NRD KEY_NRD ;
    if first.KEY_NRD ;
run;

/* Merge readmissions and index events */
data readmissions_sql ;
    merge nrd_2014_indexEvents ( drop=DXCCS1 PRCCS: )
          readmissionsClosest ( in=inR
                                rename=( NRD_DAYSTOEVENT=DaysToReadmission )
                                drop=NRD_VisitLink KEY_NRD_Readmit ) ;
    by HOSP_NRD KEY_NRD ;
    attrib Readmit length = 3 label='Readmission within 30 days (0/1)' ;
    Readmit = inR ;
    label DaysToReadmission = 'Readmission date';
run ;

/* Augment the index subset with hospital dummy records */
data combined ;
    set readmissions_sql
        NRD.NRD_2014_HOSPITAL (in = inhosp KEEP = HOSP_NRD NRD_STRATUM ) ;
    attrib InSubset length = 3 label = 'In Subset' ;
    InSubset = 1 ;
    if inhosp then do;
        * Assign a value outside the subset ;
        InSubset = 0 ;
        * Assign a valid weight ;
        DISCWT = 1 ;
        * Set analysis variables to zero ;
        IndexEvent = 0 ;
        Readmit = 0 ;
    end ;
run;
```

```
/* Tabulate (unweighted) index and readmission events */
proc freq data = combined ;
    tables InSubset*IndexEvent * Readmit / list missing ;
    format indexEvent indexEvent. readmit readmit. ;
run ;

Title "3 : National Readmission Rates" ;
/* National estimates based on NRD design                          */
/* Specify the sampling design with sampling weights DISCWT,       */
/*   hospital clusters HOSP_NRD, and stratification NRD_STRATUM    */
proc surveymeans data= combined sumwgt sum mean ;
    cluster HOSP_NRD ;
    strata NRD_STRATUM ;
    weight DISCWT ;
    var Readmit ;
    /* Subset on index events */
    domain InSubset*IndexEvent ;
    ods output domain = readmissionRates ;
    format IndexEvent indexEvent. ;
run ;
```

Stata

```
/* Data load and prep   */
/* User-defined         */
#delimit ;
cd "mydirectory";

/* Read data elements from hospital file  */
infile
      byte   HOSP_BEDSIZE
      byte   H_CONTRL
      long   HOSP_NRD
      byte   HOSP_URCAT4
      byte   HOSP_UR_TEACH
      long   NRD_STRATUM
      long   N_DISC_U
      int    N_HOSP_U
      long   S_DISC_U
      long   S_HOSP_U
      long   TOTAL_DISC
      int    YEAR
using "NRD_2014_Hospital.csv" ;
keep HOSP_NRD NRD_STRATUM ;

/* Add hospital dummy records to account for subsetting by index events */
gen DISCWT = 1 ;
gen byte indexevent = 0 ;
gen byte readmit = 0 ;
save "NRD_2014_Hospital.dta", replace ;

/* Read data elements from core file */
infile
      int    AGE
      byte   AWEEKEND
      byte   DIED
      double DISCWT
      byte   DISPUNIFORM
      byte   DMONTH
      _skip(4)
      str5   DX1
      _skip(70)
      long   HOSP_NRD
      double KEY_NRD
      long   LOS
      _skip(6)
      double NRD_DAYSTOEVENT
      long   NRD_STRATUM
      str7   NRD_VISITLINK
      _skip(55)
using "NRD_2014_Core.csv", clear ;
```

```
/* Keep minimally required variables (to reduce memory used) */
keep HOSP_NRD KEY_NRD DX1 DISCWT NRD_STRATUM AGE DMONTH DIED LOS
     NRD_VISITLINK NRD_DAYSTOEVENT ;
label var AGE             "Age in years at admission" ;
label var DIED            "Died during hospitalization" ;
label var DISCWT          "Weight to discharges in AHA universe" ;
label var DMONTH          "Discharge month" ;
label var DX1             "Diagnosis 1" ;
label var HOSP_NRD        "NRD hospital identifier" ;
label var KEY_NRD         "NRD record identifier" ;
label var LOS             "Length of stay (cleaned)" ;
label var NRD_DAYSTOEVENT "Timing variable used to identify days
                           between admissions" ;
label var NRD_STRATUM     "NRD stratum used for weighting" ;
label var NRD_VISITLINK   "NRD_VisitLink" ;

/* Convert special values to missing values */
recode AGE                      ( -99 -88 -66 = . ) ;
recode DIED                     ( -9 -8 -6 -5 = . ) ;
recode DMONTH                   ( -9 -8 -6 -5 = . ) ;
recode LOS                      ( -9999 -8888 -6666 = . ) ;
recode NRD_DAYSTOEVENT          ( -999999999 -888888888 -666666666 = . ) ;
describe ;

/* Identify AMI index events */
gen byte indexevent = 0 ;
replace indexevent = 1 if
      DX1 >= "41000" & DX1 <= "41091" & substr(DX1, 5, 1) != "2" &
      AGE >= 18 & inrange(DMONTH, 1, 11) & DIED == 0 & LOS < .;

/* Tabulate (unweighted) index events */
table indexevent ;

/* Save datasets for merge */
save "NRD_2014_Core.dta", replace ;
drop DISCWT LOS NRD_STRATUM indexevent ;
save "NRD_2014_Core_r.dta", replace ;

/* Load index events and calculate discharge dates */
use "NRD_2014_Core.dta", clear ;
keep if indexevent == 1 ;
gen dischargedate = NRD_DAYSTOEVENT + LOS ;

/* Merge index and readmission events */
keep HOSP_NRD KEY_NRD NRD_VISITLINK dischargedate ;
rename HOSP_NRD HOSP_NRD_INDEX ;
rename KEY_NRD KEY_NRD_INDEX ;
joinby NRD_VISITLINK using "NRD_2014_Core_r.dta" ;

/* Select all readmissions within 30 days */
/* Keep matches only - not a self merge   */
keep if KEY_NRD != KEY_NRD_INDEX & NRD_DAYSTOEVENT >= dischargedate &
        NRD_DAYSTOEVENT <= ( dischargedate + 30 ) ;
```

```
/* Identify closest readmission if there are multiple readmission events */
sort HOSP_NRD_INDEX KEY_NRD_INDEX NRD_DAYSTOEVENT ;
by HOSP_NRD_INDEX KEY_NRD_INDEX: gen f_num = 1 if _n == 1 ;
keep if f_num == 1 ;

/* Save readmission events */
drop HOSP_NRD KEY_NRD ;
rename HOSP_NRD_INDEX HOSP_NRD ;
rename KEY_NRD_INDEX KEY_NRD ;
save "NRD_2014_Core_readmit.dta", replace ;

/* Load core file and subset by AMI index events */
use "NRD_2014_Core.dta", replace ;
sort HOSP_NRD KEY_NRD ;
keep if indexevent == 1 ;

/* Merge and flag readmission events */
merge 1:1 HOSP_NRD KEY_NRD using "NRD_2014_Core_readmit.dta" ;
gen readmit = ( _merge == 3 ) ;

/* Add hospital dummy records */
append using NRD_2014_Hospital ;

/* Tabulate (unweighted) readmissions */
table indexevent readmit;

/* National estimates based on NRD design                          */
/* Specify the sampling design with sampling weights DISCWT,       */
/*   hospital clusters HOSP_NRD, and stratification NRD_STRATUM    */
svyset HOSP_NRD [ pw=DISCWT ], strata( NRD_STRATUM ) ;
svy: total readmit ;
svy: mean readmit ;

/* Subset on index events */
svy: total readmit, subpop( indexevent );
svy: mean readmit, subpop( indexevent );

#delimit cr
```

R

```
# Data load and prep #
setwd("mydirectory")
core <- read.csv("\\\\mydirectory\\NRD_2014_Core.CSV", header = F)

# Variable/column names
# - Scan the file "FileSpecifications_NRD_2014_Core.TXT" included with the CD
  # Columns    Description
  # ========    ===========
  # 1 -   4     Database name
  # 6 -   9     Discharge year of data
  # 11 -  19    File name
  # 21 -  23    Data element number
  # 25 -  43    Data element name
  # 45          Length of data element
  # 48          Non-zero number of digits after decimal point for numeric
  #               data element
  # 50 -  53    Data element type (Num=numeric; Char=character)
  # 55 - 154    Data element label
varnames <- read.fwf( "FileSpecifications_NRD_2014_Core.TXT",
                     widths=c( 5, 5, 10, 4, 20, 5, 5, 100 ),
                     header = F, skip = 18, strip.white = T )
colnames(core) <- as.character( varnames$V5 )

# Keep variables of interest for AMI analysis
core <- core[ c( "HOSP_NRD", "KEY_NRD", "AGE", "LOS", "DMONTH", "DX1",
                 "DIED", "DISCWT", "NRD_STRATUM", "NRD_VisitLink",
                 "NRD_DaysToEvent" ) ]

# Identify index events
DXrange <- factor(41000:41091, levels = levels(core$DX1))
core$IndexEvent <- core$DMONTH >= 1 & core$DMONTH <= 11 &
  core$DIED == 0 & core$LOS >= 0 & core$AGE >= 18 &
  core$DX1 %in% DXrange & substr(core$DX1, 5, 5) != 2

cat("Frequency of IndexEvent : \n"); table(core$IndexEvent)

# Calculate discharge dates for index events
core$PseudoDDate[core$IndexEvent] <- core$NRD_DaysToEvent[core$IndexEvent] +
                                     ifelse(core$LOS >= 0,
                                     core$LOS, NA)[core$IndexEvent]

# Identify readmission events and subset AMI index events
indexE <- subset( core, IndexEvent == 1, select = c( HOSP_NRD,
                  KEY_NRD,NRD_VisitLink, PseudoDDate ))
readmits <- subset( core, select = c( KEY_NRD, NRD_VisitLink,
                    NRD_DaysToEvent ))
colnames(readmits)[colnames(readmits) == "KEY_NRD"] <- "KEY_NRD_R"

# Inner join index events with candidate readmissions
readmits <- merge(indexE, readmits, by = "NRD_VisitLink")
```

```
# Subset to 30-day readmissions
readmits <- subset( readmits, KEY_NRD != KEY_NRD_R &
                      NRD_DaysToEvent >= PseudoDDate &
                      NRD_DaysToEvent <= ( PseudoDDate + 30 ),
                    select = c( HOSP_NRD, KEY_NRD ))

# De-duplicate readmissions and flag 30-day readmission events
readmits <- unique(readmits)
readmits$readmit <- 1

# Merge readmission events with core file
core <- core[ with( core, order( HOSP_NRD, KEY_NRD )), ]
readmits <- readmits[ with( readmits, order( HOSP_NRD, KEY_NRD )), ]
core <- merge( core, readmits, by = c( "HOSP_NRD", "KEY_NRD" ), all.x = TRUE)
core$readmit[is.na(core$readmit)] <- 0

# Tabulate (unweighted) readmissions
cat("Frequency of readmit : \n"); table(core$readmit)

# National estimates based on NRD design
# Install and load the survey package
install.packages("survey")
library(survey)

# Specify the sampling design with sampling weights DISCWT,
#   hospital clusters HOSP_NRD, and stratification NRD_STRATUM
strnrddsgn <- svydesign( ids = ~HOSP_NRD, weights = ~DISCWT,
                         strata= ~NRD_STRATUM, data = core )
cat( "Total readmissions : \n" ); svytotal( ~readmit, strnrddsgn )
cat( "Readmissions rate : \n" ); svymean( ~readmit, strnrddsgn )

# Subset on index events
strnrddsgn.sub <- subset( strnrddsgn, IndexEvent )
cat( "Total readmissions (over AMI index events) : \n" )
svytotal( ~readmit, strnrddsgn.sub )
cat( "Readmissions rate (over AMI index events) : \n")
svymean( ~readmit, strnrddsgn.sub )
```