DEPARTMENT OF HEALTH & HUMAN SERVICES
Centers for Medicare & Medicaid Services
7500 Security Boulevard, Mail Stop S3-02-01
Baltimore, Maryland 21244-1850

Office of Clinical Standards & Quality

MEMORANDUM

DATE: February 13, 2012

TO: Users of CMS' Hospital Value-Based Purchasing Program Website at http://www.cms.gov/hospital-value-based-purchasing

FROM: Quality Improvement Group
Office of Clinical Standards & Quality
Centers for Medicare & Medicaid Services

SUBJECT: Results of Reliability Analysis from Mathematica Policy Research

The Centers for Medicare & Medicaid Services (CMS) is committed to make public the analysis supporting the reliability and minimum case thresholds for measures in the outcome domain originally proposed for the FY2014 Hospital Value-Based Purchasing Program.

Mathematica, one of CMS's vendors, provided CMS the analysis that follows as one input informing measure selection and minimum case thresholds for measures in the outcome domain of the FY2014 Hospital Value-Based Purchasing Program. In determining which measures and minimum case thresholds to propose and finalize, CMS considered many factors, including this analysis and the statutory requirement to include measures that cover healthcare-associated infections in Hospital Value-Based Purchasing.

Mathematica's analysis follows as a memorandum to CMS' policy analyst, Sophia Chan, on the proceeding pages. Questions about this analysis, as well as CMS' use of it, may be directed to HospitalVBP@cms.hhs.gov.

MATHEMATICA Inc.

MEMORANDUM

TO: Sophia Chan

FROM: Eric Schone, Mai Hubbard and David Jones **DATE:** 11/18/2011

SUBJECT: Reporting Period and Reliability of AHRQ, CMS 30-day and HAC Quality Measures - Revised

---

Reliability of an outcome measure is the extent to which variation in the measure is due to variation in quality of care rather than random variation due to the sample of cases observed. Reliability should be distinguished from validity, which is the accuracy with which the measure reflects the quality of care on average. Validity depends on the accuracy of calculations: whether complete records are available and whether the information they contain is correct; and on the relation of the measure to quality: whether better quality care results in a better outcome. Reliability depends on whether variation in real performance is large compared to random variation due to sampling. The statistical concept of reliability (R) used to determine minimum case size for a particular measure is whether a hospital's ranking on that measure, compared to its performance in other periods or compared to other hospitals, is likely to be the same if we take repeated samples of the hospital's cases.[1] R depends on the rate's variance between hospitals, the variance of the rate within a hospital's own cases, and the number of discharges from a given hospital. For its evaluation, Yale proposed a standard of R=0.4, which is considered to be the lower limit of "moderate" reliability.

This memo revises an earlier memo that described analysis of reliability for three sets of measures: (1) 30-day risk-standardized mortality and readmission measures reported by CMS on *Hospital Compare*; (2) AHRQ's Inpatient Safety Indicators (IQI) and Patient Safety Indicators (PSI); and (3) Hospital Acquired Condition (HAC) measures. Since the time that earlier memo was prepared, we have made several revisions to our calculation of the AHRQ measures. This memo contains results for all three measures, including updated results for the AHRQ measures.

The memo presents the number of cases required for each VBP measure to meet a moderate standard for reliability as well as the percentage of hospitals that would meet this standard when data periods of differing lengths are used. These results may be relevant to CMS decisions concerning both the length of the data period and minimum N for the VBP program.

[1] Though the statistical basis of this measure is a hypothetical sampling distribution at one point in time, one consequence of an unreliable measure is likely to be implausible fluctuations in a hospital's classification when measured at different points in time.

METHODS

Results were calculated using data submitted for hospital value-based purchasing baseline performance measurement. The results include prospective payment hospitals, and exclude critical access hospitals. Additionally, for 30-day measures only, results include hospitals from the state of Maryland.

For the IQI and PSI measures, values were projected from the 7 months of input data used to estimate rates, by assuming that the number of discharges per hospital would change in proportion to the change in the number of months in the time period. The data covers prospective payment hospitals over the time period of March 2010 to September 2010.

The calculations for the 30-day measures included discharges from July 2009 through June 2010.

For HAC measures, we used the data over which VBP baselines were estimated. However, in order to create stable hospital-level estimates for these extremely rare events, for calculations in which we estimated the relationship between reliability and case size, we restricted the data set to those hospitals that were largest in terms of measured case size. We performed estimates over the largest 200 hospitals. The data covers the 7 month period from March 2010 to September 2010.

For the AHRQ measures, the output from the measure calculation software contains a reliability weight, which is the ratio of the between variance to the sum of the between variance and the hospital's estimated within variance divided by the number of observations. The reliability is recalculated for 6 month through 24 month time periods by assuming that the within and between variances would remain the same, and only the number of observations would change. Then the median reliability and the proportion of hospitals with reliability exceeding the threshold can be calculated. Because the within variance is permitted to vary by hospital, each hospital has a unique threshold for which R=0.4. We report the median of these thresholds as the N needed for moderate reliability. The reliability estimates used for the IQI and PSI measures were calculated using AHRQ's Version 4.2 software. We also approximated reliability for the composite measures using as estimates of signal and noise variances weighted averages of individual component elements.

For CMS's 30 day measures we approximated the signal to noise ratio used in AHRQ's smoothing of risk adjusted rates.[2] We approximated the signal to noise ratio by using hospitals'

[2] An accompanying memo describes our reasons for selecting this method, as well as results based on alternate methods, such as using the intraclass correlation (ICC), which is the ratio of between variance to total variance, estimated along with other parameters of the regression used to calculate risk adjusted mortality.

individual risk adjusted rates to calculate reliability for each hospital. For each hospital, we calculated the value of N for which R=0.4 and used the median of these values as the threshold value for N. The median reliabilities and the proportion exceeding the reporting threshold are calculated by assuming that the number of observations is a constant function of the number of months in the time period.

For HAC measures, we performed a similar approximation of the signal to noise ratio used in AHRQ's smoothing of risk adjusted rates, but adapted for rare events.

FINDINGS

Table 1 presents results for the IQI measures. They indicate that 4 of the measures, AMI, CHF, stroke, and pneumonia, achieve the moderate reliability threshold with a sample of about 100. Three others, hip fracture, gastrointestinal hemorrhage, and AAA repair, achieve moderate reliability with a sample of about twice that size. As before, hip fracture, gastrointestinal hemorrhage, and AAA repair are rarely sufficiently reliable for reporting or payment, while other measures, and the IQI composite achieve moderate reliability for a majority of hospitals in the 12 to 24 month time period.

As shown in Table 2, revised results for PSI measures indicate that several measures are not reliable even with a long time period. Post-operative hip fracture requires a sample of over 15,000 to achieve even moderate reliability, and rarely does. Post-operative sepsis and wound dehiscence have lower thresholds but also rarely exceed them because measure denominators are generally small. Two measures, decubitus ulcers and pulmonary embolism, are reliable for most hospitals with a time period of 6 months or longer, because they have lower reliability thresholds.

Two other measures, puncture/laceration and post-operative infection, have a substantially higher threshold, about 800, but also achieve moderate reliability for about half of hospitals at 6 months. That is because the denominators for these measures are large. The remaining measures: post-operative respiratory failure, pneumothorax and death from treatable conditions achieve moderate reliability for a half of hospitals or more in the 12 to 24 month range. Because it weights the most reliable measures heavily, the PSI composite achieves moderate reliability at a majority of hospitals for reporting periods of 6 months or longer.

Table 3 presents findings from CMS's 30-day measures. Most hospitals do not achieve reliability with 12 months of data. A little less than half achieve moderate reliability using 24 months of data. We did not analyze the reliability of a composite measure.[3]

[3] Using the ICC estimated from the regression parameters, higher values for reliability are obtained. When that method is used, over half of hospitals attain moderate reliability for heart failure and pneumonia with 12 months of data, and for AMI with 24 months.

HACs, as shown in Table 4, fall into three groups. In the first group, foreign object retained after surgery, air embolism, and blood incompatibility have very low reliability on the basis of their extreme rarity in reported data. In fact, valid results could not be obtained for blood incompatibility or air embolism, because the estimated signal variance is negative or missing. Though it is labeled undefined in the table, its reliability can be assumed to be 0 and the N required for R=0.4 to be infinite. In the second group, foreign object retained after surgery, falls and trauma and poor glycemic control exhibit low reliability over any of the time periods presented.

In the third group, catheter associated urinary tract infection (UTI), pressure ulcers, and vascular catheter associated infections have moderate reliability thresholds of 1,000 to 3,500 cases. Because denominators are large, including all medical and surgical cases, they exhibit moderate reliability for half or more of hospitals, when the reporting period is lengthened to 21 months. In fact, UTIs exhibit moderate reliability in over half of hospitals when the time period is 6 months. We also calculated a composite measure, which is the combined occurrence rate across all types of HAC. With a moderate reliability threshold of about 2,000, this composite is also moderately reliable for the majority of hospitals when 12 months of data are used.

CONCLUSIONS

The most recent plan for calculation of AHRQ measure scores for value-based purchasing postulated scoring the IQI and PSI composites using a measurement period of 7 months. The results of this analysis suggests that though the PSI scores resulting would be moderately reliable for a majority of hospitals, IQI scores would not meet this standard of reliability without accumulating 24 months of data. The results suggest that a similar length of time (24 months) is needed for CMS 30-day measures to achieve the same standard. Like that of PSIs, the reliability of HAC measures varies widely. However, the more reliable measures and the composite measure that sums the total number of HACs meet the moderate reliability standard for a majority of hospitals after accumulating 12 months of experience.

The results presented for IQIs and PSIs are based on the version 4.2 AHRQ measures now being publicly reported. Specifications for several measures are changing for version 4.3 or have recently changed. These changes to specifications will affect results in the future.

The impact of using unreliable measures for value-based purchasing program will vary depending on the array of measures, the scoring method and design of the value-based modifier. Investigation of all three factors using results of these measures should be undertaken as part of design of the program.

cc: Marian Wrobel

Table 1. Reliability and Time Period of Calculation
AHRQ Inpatient Quality Indicators

		6 Months		12 Months		18 Months		24 Months	
Measure	N at which R=0.4	Median Reliability*	R≥0.4** (%)	Median Reliability*	R≥0.4** (%)	Median Reliability*	R≥0.4** (%)	Median Reliability*	R≥0.4** (%)
IQI-15 AMI[C]	84	0.12	10	0.21	27	0.28	38	0.35	45
IQI-16 CHF[C]	110	0.25	29	0.40	50	0.50	60	0.57	67
IQI-17 Stroke[C]	97	0.11	18	0.21	31	0.28	39	0.34	45
IQI-18 Gastro Hemorrhage[C]	242	0.07	0	0.12	7	0.17	17	0.22	25
IQI-19 Hip Fracture[PR,C]	228	0.05	0	0.10	5	0.14	12	0.18	18
IQI-20 PN[C]	79	0.29	32	0.45	56	0.55	69	0.62	75
IQI-11 AAA Repair[PR]	216	0.00	0	0.00	1	0.00	3	0.00	5
IQI Composite	***	0.16	15	0.28	34	0.37	44	0.44	50

Note: Estimated over hospitals required to submit POA data, March 2010 to September 2010. N=3401.

* Reliability of measure of hospital of median case size.
** Proportion of hospitals with case size large enough that R≥0.4.
*** Composite does not have a single N at which R=0.4 because it is a combination of measures.

PR – Publicly reported indicators.

C – These indicators are not publicly reported but are part of a publicly reported AHRQ composite.

Table 2. Reliability and Time Period of Calculation
AHRQ Patient Safety Indicators

		6 Months		12 Months		18 Months		24 Months	
Measure	N at which R=0.4	Median Reliability*	R≥0.4** (%)	Median Reliability*	R≥0.4** (%)	Median Reliability*	R≥0.4** (%)	Median Reliability*	R≥0.4** (%)
PSI-03 Decubitus Ulcer[C]	44	0.82	84	0.90	89	0.93	91	0.95	92
PSI-04 Death from Treatable[PR]	52	0.10	9	0.18	24	0.24	34	0.30	40
PSI-06 Iatrogenic Pneumo[PR,C]	3,029	0.18	16	0.30	37	0.39	49	0.47	57
PSI-07 Infection[C]	750	0.38	48	0.56	66	0.65	76	0.71	81
PSI-08 Post-Op Hip Fracture[C]	15,011	0.00	0	0.01	0	0.01	0	0.02	0
PSI-11 Post-Op Resp Failure	167	0.25	33	0.40	50	0.50	59	0.57	65
PSI-12 Pulmonary Embol/DVT [PR,C]	129	0.52	61	0.67	75	0.77	80	0.81	83
PSI-13 Post-Op Sepsis[C]	225	0.05	5	0.09	15	0.13	22	0.16	28
PSI-14 Post-Op Wound Dehiscence[PR,C]	814	0.02	0	0.04	0	0.07	2	0.09	4
PSI-15 Puncture/Laceration[PR,C]	862	0.43	53	0.60	69	0.70	77	0.75	82
PSI Composite	***	0.67	73	0.81	83	0.86	88	0.89	90

Note: Estimated over hospitals required to submit POA data, March 2010 to September 2010.. N = 3401

* Reliability of measure of hospital of median case size.
** Proportion of hospitals with case size large enough that R ≥0.4.
*** Composite does not have a single N at which R=0.4 because it is a combination of measures

PR – Publicly reported indicators
C – These indicators are not publicly reported but are part of a publicly reported AHRQ composite.

Table 3. Reliability and Time Period of Calculation
30-Day Risk-Standardized Mortality Measures

		6 Months		12 Months		18 Months		24 Months	
Measure	N at which R=0.4	Median Reliability*	R≥0.4** (%)	Median Reliability*	R≥0.4** (%)	Median Reliability*	R≥0.4** (%)	Median Reliability*	R≥0.4** (%)
MI Mortality	107	0.09	2	0.17	12	0.23	24	0.29	33
HF Mortality	195	0.11	2	0.20	14	0.27	28	0.33	40
PN Mortality	211	0.11	1	0.19	8	0.27	22	0.32	35

Note: Estimated over prospective payment and Maryland hospitals, July, 2009 to June, 2010. Only hospitals with N>0 are included for each measure. Hospitals with N<25 are included. N=3142, 3237, 3257, respectively.

* Reliability of measure of hospital of median case size.
** Proportion of hospitals with case size large enough that R≥0.4.

Table 4. Reliability and Time Period of Calculation
HAC Measures

		6 Months		12 Months		18 Months		24 Months	
Measure	N at which R=0.4	Median Reliability*	R≥0.4** (%)	Median Reliability*	R≥0.4** (%)	Median Reliability*	R≥0.4** (%)	Median Reliability*	R≥0.4** (%)
Foreign object retained after surgery	15,417	0.05	0	0.09	1	0.12	4	0.14	7
Air embolism	UNDEF	UNDEF	UNDEF	UNDEF	UNDEF	UNDEF	UNDEF	UNDEF	UNDEF
Blood incompatibility	UNDEF	UNDEF	UNDEF	UNDEF	UNDEF	UNDEF	UNDEF	UNDEF	UNDEF
Pressure ulcer stages III & IV	2,195	0.28	31	0.40	50	0.50	62	0.53	66
Falls and trauma	16,119	0.05	0	0.08	1	0.12	4	0.14	6
Vascular catheter-associated infection	3,498	0.19	15	0.29	34	0.38	48	0.42	52
Catheter-associated UTI	1,070	0.44	55	0.57	70	0.67	80	0.70	83
Manifestations of poor glycemic control	27,888	0.03	0	0.05	0	0.07	0	0.08	1
All HAC	1,950	0.30	35	0.42	53	0.53	65	0.56	69

Note: Estimated over hospitals required to submit POA data, March 2010 to September 2010. Reliability estimate based on 200 largest hospitals. N=3401.

* Reliability of measure of hospital of median case size.
** Proportion of hospitals with case size large enough that R≥0.4.

UNDEF – Undefined because estimated signal variance negative or undefined