Split-Half Reliability Method Examples

Example 1

We tested the reliability of the facility measure score by calculating the intra-class correlation coefficient (ICC) of the measure score. To calculate the ICC, we used the Medicare FFS FY 2012-2015 Dataset. For ASCs with two or more urology procedures, these procedures were then randomly split into the two samples (2 years of combined data for each sample). The ICC evaluates the agreement between the risk-standardized hospital visit rates (RSHVRs) calculated in the two randomly selected samples.

The ICC [2,1] score of 0.45, calculated for two years of data, indicates moderate measure score reliability.

Example 2

We tested the reliability of the facility measure score by calculating the intra-class correlation coefficient (ICC) of the measure score. To calculate the ICC, we used the Medicare FFS CYs 2012-2015 Dataset. For ASCs with two or more general surgery procedures, these procedures were randomly split into the two samples within each facility. The ASCs with one procedure were randomly split into the two samples. The ICC evaluated the agreement between the risk-standardized hospital visit ratios (RSHVRs) calculated in the two randomly selected samples [1].

The ICC [2,1] score of 0.530, calculated for four years of data, indicates moderate measure score reliability.

Example 3

We defined reliability as described by Lord and Novick using split-sample methodology. (Lord FM, Novick MR. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley; 1968)

Using split-sample methodology, FTR had a split half sample correlation estimate of 0.32, with the upper bound on validity (provided by the square root of the Spearman-Brown reliability correction) being 0.56.

Example 4

The reliability of a measurement is the degree to which repeated measurements of the same entity agree with each other. For measures of hospital performance, the measured entity is naturally the hospital, and reliability is the extent to which repeated measurements of the same hospital give similar results. In line with this thinking, our approach to assessing reliability is to consider the extent to which assessments of a hospital using different but randomly selected subsets of patients produces similar measures of hospital performance. That is, we take a "test-retest" approach in which hospital performance is measured once using a random subset of patients, then measured again using a second

random subset exclusive of the first, and finally comparing the agreement between the two resulting performance measures across hospitals (Rousson et al., 2002). For test-retest reliability, we combined index admissions from successive measurement periods into one dataset, randomly sampled half of patients within each hospital, calculated the measure for each hospital, and repeated the calculation using the second half. Thus, each hospital is measured twice, but each measurement is made using an entirely distinct set of patients. To the extent that the calculated measures of these two subsets agree, we have evidence that the measure is assessing an attribute of the hospital, not of the patients. As a metric of agreement we calculated the intra-class correlation coefficient (ICC) (Shrout and Fleiss, 1979), and assessed the values according to conventional standards (Landis and Koch, 1977). Specifically, we used dataset 1 split sample and calculated the RSRR for each hospital for each sample. The agreement of the two RSRRs was quantified for hospitals using the intra-class correlation as defined by ICC (2,1) by Shrout and Fleiss (1979).

Using two independent samples provides a stringent estimate of the measure's reliability, compared with using two random but potentially overlapping samples which would exaggerate the agreement. Moreover, because our final measure is derived using hierarchical logistic regression, and a known property of hierarchical logistic regression models is that smaller volume hospitals contribute less ´signal´, a split sample using a single measurement period would introduce extra noise. This leads to an underestimate in the actual test-retest reliability that would be achieved if the measure were reported using the full measurement period, as evidenced by the Spearman Brown prophecy formula (Spearman 1910, Brown 1910), which estimates the reliability of the measure if the whole cohort were used, based on an estimate from half the cohort.

There were 991,007 admissions in the combined 3-year sample, with 494,297 in one sample and 496,710 in the other randomly selected sample. The agreement between the two RSMRs for each hospital was 0.55, which according to the conventional interpretation is "moderate" (Landis & Koch, 1977). Note that this analysis was limited to hospitals with 12 or more cases in each split sample. The intra-class correlation coefficient is based on a split sample of three years of data, resulting in a volume of patients in each sample equivalent to only 1.5 years of data, whereas the measure is reported with the full three years of data. The correlation coefficient is expected to be higher using the full three-year sample since it would include more patients.

Example 5

To test the reliability of facility-level risk-standardized readmission rates (RSRRs), we calculated the intra-class correlation coefficient (ICC) using a test-retest approach that examines the agreement between repeated measures of the same IPF for the same time period. The randomly sampled sets of admissions from a given hospital are assumed to reflect an independent set of re-measurement of readmission rates for the hospital. Good reliability is assumed if the risk-standardized measure rates calculated from the random datasets for the same IPF are similar. Higher ICC values indicate stronger agreement, and hence, better measure reliability.

We used two test-retest approaches to generate independent samples of patients within the same IPF: a split-half sampling design and bootstrapping. For split-half sampling, we randomly sampled half of all

eligible index admissions in each facility over the two-year period, resulting in two samples that cover the same two-year period but with case volume the size of a measure that would be calculated with one year of data. The ICC in the split-half sampling design was estimated using the RSRRs of the two split-half samples.

For bootstrapping, we sampled 1,000 pairs of samples from the original measure cohort with replacement (stratified sampling by IPF), resulting in 1,000 pairs of new samples within each IPF with the identical sample size as in the original measure cohort, thus maintaining the sample size of a two-year measure. The ICC in the bootstrap sampling was estimated for each pair of the bootstrap samples. With the 1,000 ICC estimates from the 1,000 pairs of bootstrap samples, we determined the distribution of estimated ICC coefficients and thus could calculate the mean and 95% CI of the ICC.

RSRR distributions across IPFs obtained for the two randomly split-half samples that we established for test-retest reliability testing are displayed below. We estimated RSRR for each sample using a hierarchical logistic regression model and RSRR calculations described in 2b5. The average RSRR in the two split-half samples is very similar with means of 21.03 and 20.93 percent (Table 3). The corresponding intra-class correlation coefficient is 0.60.

Table 1. RSRR distributions for IPFs in split-half samples (January 2012–December 2013)

	# Index Admissions	# of IPFs (n≥25)	Mean	SD	Min	10th Percentile	Lower Quartile	Median	Upper Quartile	90th percentile	Max
Sample 1	358,087	1,594	21.03	2.71	12.62	17.73	19.20	20.89	22.72	24.50	31.02
Sample 2	358,087	1,593	20.93	2.56	13.29	17.85	19.14	20.73	22.41	24.36	30.89

The ICC obtained from the bootstrapping approach, comparing 1,000 pairs of samples of the original measurement cohort, which were sampled with replacement yielding an identical sample size as the original measurement cohort, is 0.78 (95% CI 0.77-0.80).

Example 6

The reliability of a measurement refers to the degree to which repeated measurements of the same entity agree with each other. Specifically, for hospital-level performance measures, reliability characterizes to what extent repeated measurements of the same hospital generate similar results. In line with this thinking, our approach to assessing reliability is to consider the extent to which assessments of a hospital using different but randomly selected subsets of patients produces similar measures of hospital performance. That is, we assessed measure reliability for the measure using split-half correlations with claims data for all hospitals. To calculate split-half reliability, we randomly divided the hospital-level data into two equal samples; thus, each hospital is measured twice, but each measurement is made using an entirely distinct set of patients. We calculated the measure performance in both samples for each hospital; to the extent that the calculated measures of these two

subsets agree, we have evidence that the measure is assessing an attribute of the hospital, not of the patients. As a metric of agreement, we calculated the Pearson correlation between the performance rate estimates: the higher the correlation, the higher the reliability of the measure. In order to produce estimates that are as stable as possible, we repeated this approach 1,000 times, a technique known as bootstrapping [1].

Because we expect hospitals with relatively few cases to have less reliable estimates, we only included scores for hospitals with at least 60 patients in the reliability calculation (i.e., with 30 patients in each of the split samples). This approach is consistent with a reporting strategy that includes smaller hospitals in the measure calculation, but does not publicly release the measure score for smaller hospitals (i.e., labels them in public reporting as having "too few cases" to support a reliable estimate). We note that the minimum sample size for public reporting is a policy choice that balances competing considerations such as the reliability of the measure score and transparency for consumers, and that the cutoff used for this analysis is one of many that might be reasonably used.

In addition, we conducted a second analysis of measure reliability using the intraclass correlation coefficient (ICC) signal-to-noise method to determine a recommended minimum case count to maintain a moderate level of reliability. The ICC is estimated from the random effects model that produces the risk-standardized hospital visit rates, as ICC = V / (V + σ), where V is the between variance and σ is the sampling variance of the estimated provider level results. Because π 2/3 is the sampling variance of the logit distribution, the ICC of the measure, which is based on a logit model, is ICC = V / (V + π 2/3).

We used the intercept variance from the hierarchical logit models used to estimate the measure (0.0909 for inpatient admission, and 0.1108 for ED visits) as the estimate of the between variance. The ICC can be used to calculate the reliability (R) of individual hospitals using the formula: R= N/(N + (1 - ICC)/ICC).[1] The case size required for a given R is: N = R*(1 - ICC)/(ICC*(1 - R)). We looked for the N required to maintain a reliability level of 0.4 or higher.

Citations

1. Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test–retest reliability of continuous measurements. Statistics in Medicine 2002; 21:3431-3446.

There were 942 hospitals with ≥60 patients in their cohorts in the full sample. This sample was randomly split 1,000 times and the Pearson correlation was calculated each time. For the inpatient admission measure, on average, the agreement between the two hospital visit rates for each hospital was 0.413 (95% confidence interval (CI) = 0.37-0.45), which according to conventional interpretation is "moderate." For the ED visit measure, on average, the agreement between the two hospital visit rates for each hospital was 0.270 (95% confidence interval (CI) = 0.22-0.33), which according to conventional interpretation is "moderate."

In addition, we found to achieve a reliability (ICC) of 0.4, we only require 25 patients for the inpatient admissions rate and 20 patients for the ED visit rate per performance period.