

Benchmarking Physician Performance: Reliability of Individual and Composite Measures

eAppendix Reliability Formula

Calculation of Reliability Estimates

An important question in physician quality measurement is “how big a sample size is needed to make fair judgments about quality?” In physician profiling, it is common to assess sample size adequacy by estimating the reliability of a measure at a particular sample size. When used as a term-of-art, rather than colloquially, reliability is the ratio of the physician-to-physician variance to the sum of physician-to-physician variance and the measurement variance within physicians. The measurement variance within physicians (error variance) is dependent on the sample size as well as the performance rate for the measure. High reliability estimates suggest that it is easy to tell physicians apart. Low reliability estimates suggest that it is difficult to be confident that physicians are different from one another given the available sample sizes. The reliability estimate varies between 0 and 1. A reliability of 0.70 or higher is typically considered acceptable for psychometric purposes.¹

The average level of performance can affect the reliability estimate. Low reliability estimates for measures with high average scores may be less important since they are a consequence of small variation in a compressed scale. Low reliability for measures with large variation near the middle of a scale is much more directly a sample size problem. The reliability metric does help make measures with different performance rates and different physician-to-physician variation comparable. For profiling problems it is more sensible to ask: “What is the minimum reliability?” than to ask: “What is the minimum sample size?” Authors often advocate minimum reliabilities for various purposes.^{2,3}

As noted above, reliability estimates require an estimate of physician-to-physician variation. Generally this cannot be obtained by simply summarizing the physician level scores. Variation in sample sizes from physician to physician as well as small sample sizes in general can bias simple variance calculations and inflate the between-physician variances. To estimate the physician-to-physician variance for this study, we fit beta-binomial models to each measure using a SAS macro developed for this purpose.⁴ Conceptually, this is estimating what physician-to-physician variation would be if we had very large sample sizes for each physician.

Once we have an estimate of physician-to-physician variance, we can estimate what the reliability would be for a particular measure at a particular sample size using the Spearman-Brown formula:

$$reliability = \sigma_{physician-to-physician}^2 / (\sigma_{physician-to-physician}^2 + \sigma_{error}^2 / n)$$

where the error variance is the binomial variation and n is the sample size.⁵

The error variance depends on the performance rate for the physician. In this paper, we use an average performance rate for each category.

We estimated reliability based on 2 different sample size assumptions, based on the average sample size per measure/volume category and for the minimum sample size per volume category. The reliability estimate at the average number of observations for the measure/volume category reflects the typical experience of physicians in this population. The reliability at the minimum sample size per volume category shows the “worst-case” reliability for the volume category. The reliability estimate at the minimum sample size also facilitates comparison of reliability estimates across measures since the average number of observations varies by measure. To promote comparability across volume categories for a measure, the physician-to-physician variance for the group of physicians with 10 or more observations was used for computing reliability for all groupings.

Rather than creating one reliability estimate for the entire national database, reliability estimates were calculated within each plan. We wanted the physician-to-physician variance used in calculating the reliability estimate to reflect the natural variation *within a plan* and not to be unduly influenced by patients or other factors that are likely to vary *across plans*. We selected 2 plans for inclusion in our presented findings to illustrate potential differences in reliability results between different performance measures and regions, plans C and I. These plans were selected based on relative size, with plan C being smaller and plan I larger. These plans were representative of the remaining plans in terms of market share and other characteristics.

The reliability for the composite score is conceptually similar but is based on a normal hierarchical model fit with Proc Mixed in SAS version 9.2.⁶

Calculation of Composite Scores

To calculate the composite score, we first grouped the performance measures into 3 composite scores for chronic, acute, and preventive care, as well as an overall composite comprising all 27 measures.

The simplest way to calculate a composite score would be to determine the number of quality events met for each physician and to divide it by the number of quality events attributed to the physician. This approach presents several concerns for making comparisons: the distribution of quality events may vary across physicians, the “degree of difficulty” of performing well on a measure varies, and the degree of variability in performance results also varies. To address these issues, we constructed a composite score for each physician based on the *standardized average rate* for the relevant measures for that physician’s patient population. By standardizing the measures prior to averaging we took into account differences in measure means and standard deviations that might otherwise distort comparisons across physicians. This method allows for the comparison of physicians’ performance on the same scale even if their patient mixes and applicable measures were different. For each physician, we calculated an average of the standardized measure rates weighted by the number of observations that the physician has for the measure. In computing a composite score for a physician, only individual indicators with 4 or more observations were used. With less than 4 observations, a standardized rate is constrained to be 0, 1, or 0.5, and thus had unacceptably strong

influence on the composite results. This approach scores a physician more heavily on what that physician does by focusing on the individual physician's mix of patient characteristics and attributed quality events.

To create the reliability estimate for the composite scores using the method described above, we computed the physician-to-physician variation using a hierarchical modeling approach. For reasons described above, the reliability estimates were computed at the plan level, and within each composite score at different volume thresholds (10 or more observations, 20 or more, 30 or more, and 50 or more). Using the physician-to-physician variance for each plan and volume group (eg, 10+) for each composite (eg, overall, preventive, etc), a reliability was computed for each physician. A sample of up to 900 physicians was used for each composite for each plan. The median of the reliability estimates across the physicians is reported.

References

- 1. Nunnally J, Bernstein I.** *Psychometric Theory*. 3rd. New York: McGraw-Hill; 1994.
- 2. Safran DG, Karp M, Coltin K, et al.** Measuring patients' experiences with individual primary care physicians. Results of a statewide demonstration project. *J Gen Inter Med*. 2007;21 (1):13-21.
- 3. Hofer TP, Hayward RA, Greenfield S, Wagner EH, Kaplan SH, Manning WG.** The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. *JAMA*. 1999;281(22):2098-2105.
- 4. Wakeling I.** BETABIN macro. 2004.
<http://www.sensory.org/library/files/SAS/betabin-v22.sas>.
- 5. Shrout PE, Fleiss JL.** Intraclass correlations: uses in assessing rater reliability. *Psych Bull*. 1979;2:420-428.
- 6. SAS version 9.2** Cary, NC: SAS Institute; 2005.