Comment on E&M Guidebook Version 3.0

Submitted by the University of Michigan Kidney and Epidemiology Cost Center
June 25, 2025

For measures based on non-Binomial data, PQM recommends that entity-level reliability is evaluated based on the Intraclass Correlation Coefficient (ICC) with a Spearman-Brown prophecy adjustment. It also specifically recommends against using the Inter-Unit Reliability (IUR) for reliability assessments, citing previous literature that highlighted undesirable properties of the IUR (e.g., that the IUR often decreases with improved risk-adjustment).[1] However, we note that the population parameters for the ICC with Spearman-Brown prophecy adjustment and the IUR are mathematically equivalent, sharing identical theoretical formulas. [1,2] Therefore, these metrics should share the same general properties, and many of the limitations of the IUR also apply to the recommended ICC metric. While PQM also recommends a method for estimating these parameters (i.e., permutation sampling) that is different from the typical bootstrap implementation of the IUR, empirical evidence has shown that these differences alone do not result in large changes to the properties of the reliability statistics.[3]

The literature that PQM cites[1] on the IUR statistic is more of a criticism of how reliability statistics are typically used and interpreted (i.e., by comparing them to very strict thresholds such as 0.6 or 0.7 for measure endorsement criteria) than a specific criticism of the IUR formula relative to other similar reliability metrics. In many cases, important steps to improve measure validity such as thorough risk-adjustment and broader inclusion of facilities can actually reduce the estimated reliability metric.[4] Therefore, rejecting quality measures based solely on the numeric value of these reliability statistics may unnecessarily exclude many useful measures and create a negative incentive against key aspects of the measure development process. Instead, we argue that a more nuanced approach should be taken that considers the tradeoffs involved in measure development and the possibility for very useful measures to have entity-level reliability statistics below fixed thresholds.

Furthermore, it is well-known that reliability metrics depend on entity size, where smaller entities tend to have lower reliability. For many types of healthcare entities such as dialysis facilities, the entity size needed to achieve a reliability threshold of 0.4 or 0.6 is much larger than the number of patients that these facilities typically treat, even in cases where the between-entity variation in the outcome of interest is large. We show two examples of riskadjusted dialysis facility measures in the table below that would require a highly unrealistic distribution of facility sizes to achieve these reliability thresholds across most facilities, despite clinically meaningful variation in the measures that has been shown to be useful in quality incentive programs. Therefore, strict thresholds of reliability that are universally applied to all entity types may be too rigid to fully capture meaningful performance.

	Facility Size Needed		**Observed Distribution of Facility Sizes from Real Data**		
	IUR=0.4	IUR=0.6	25th perc.	median	75th perc.
Standardized Readmission Ratio	64 index discharges	143 index discharges	34	54	82
Standardized Mortality Ratio	171 patients	384 patients	49	75	109

1. Kalbfleisch, J.D., He, K., Xia, L. Li, Y. (2018). Does the inter-unit reliability (IUR) measure reliability? Health Services and Outcomes Research Methodology. 18:215-225.
2. Warrens, M.J., (2017) Transforming intraclass correlation coefficients with the Spearman–Brown formula. Journal of Clinical Epidemiology. (85): 14-16.
3. Nieser, K.J. and Harris, H.S. (2024) Comparing methods for assessing the reliability of health care quality measures. Statistics in Medicine: 43(23).
4. Hartman, N., Shahinian, V. B., Ashby, V. B., Price, K. J., & He, K. (2024). Limitations of the Inter-Unit Reliability: A Set of Practical Examples. Health services & outcomes research methodology, 24(2), 156–169.