Adding New Components to a Composite Quality Metric

How Good Is Good Enough?

Stephen Salerno, PhD,† Eileen Yang, MS,*‡ Claudia Dahlerus, PhD,*§ Richard A. Hirth, PhD,*‖ Peisong Han, PhD,* Tao Xu, PhD,* Ashley Eckard, MS,* Wilfred Agbenyikey, ScD, MPH,¶ Golden M. Horton, MS,¶ Stephanie Clark, MD,¶ Joseph M. Messana, MD,*§‖ and Yi Li, PhD*‡*

Objectives: This study illustrates how the statistical reliability of an individual measure relates to the overall reliability of a composite metric, as understanding this relationship provides additional information when evaluating measures for endorsement.

Background: National quality measure endorsement processes typically evaluate individual metrics on criteria such as importance and scientific acceptability (eg, reliability). In practice, quality measures may be used in composite rating systems, which aid in the interpretation of overall quality differences.

Methods: We define an individual measure's reliability by its intraclass correlation and analytically establish the relationship between a composite's reliability and the reliability of its components. We use real data to confirm this relationship under various scenarios. We are motivated by 8 quality measures, which comprise the Quality of Patient Care Star Ratings on Dialysis Facility Care Compare. These measure 4 primary outcomes (mortality, hospitalizations, readmissions, and blood transfusions), vascular access (2 measures), and facility processes (2 measures).

Results: Depending on the reliability of the individual measures, their respective weights in the composite, and their pairwise correlations, there are circumstances when adding a new measure, even if it is less reliable, increases the composite's reliability. For the dialysis facility Star Ratings, we find that the combined reliability of measures grouped within certain domains of care exceeded the reliability of the individual measures within those domains.

Conclusions: New quality measures may add utility to a composite rating system under certain circumstances—a consideration that should, in part, factor into quality measure endorsement processes.

Key Words: Dialysis, Public Reporting, Reliability

(*Med Care* 2025;63:293–299)

Measuring and reporting the quality of care delivered by health care providers has been a major focus of health policy for several decades.[1,2] Quality of care is multidimensional, encompassing various aspects that reflect patient outcomes and experiences. Health care facilities perform at different levels on various aspects of care, and few excel in all measured outcomes.

Whether to measure a particular aspect of quality depends on 2 criteria: (1) "importance" and (2) "scientific acceptability."[6,7] Importance reflects whether an outcome is meaningful to stakeholders and whether there is variability in performance between providers that can be improved upon.[8,9] Scientific acceptability relates to both a measure's clinical soundness—is it tied to provider-owned structures or processes and related to primary outcomes (eg, mortality), and measurement soundness—is it a reliable and valid metric?[10] A surrogate for reliability in the quality measure endorsement process is whether there is appreciable variation in quality measured between providers (variance of interest, or signal) when compared with the variation among patients within a given provider (random variation). The ratio of the variance of interest to the total variance (variance of interest plus random variation) is otherwise known as a measure's intraclass correlation (ICC).[11]

From the *Department of Biostatistics, Kidney Epidemiology and Cost Center, University of Michigan, Ann Arbor, MI; †Department of Biostatistics, Public Health Sciences Division, Biostatistics, Fred Hutchinson Cancer Center, Seattle, WA; ‡Department of Biostatistics, University of Michigan, Ann Arbor, MI; §Division of Nephrology, Department of Biostatistics, University of Michigan Health System, Ann Arbor, MI; ‖Department of Health Policy and Management, University of Michigan, Ann Arbor, MI; and ¶Department of Biostatistics, The Centers for Medicare and Medicaid Services, Baltimore, MD.

The analyses upon which this publication is based were performed under Contract Number HHSM-500-2013-13017I0 and Contract Number 75FCMC18D0041, Task Order Number 75FCMC18F0001 entitled, "Kidney Disease Quality Measure Development, Maintenance, and Support," sponsored by the Centers for Medicare and Medicaid Services, Department of Health and Human Services. Further, the authors attest that the content of this manuscript is solely the responsibility of the authors and does not reflect the official views of the Centers for Medicare and Medicaid Services.

The authors declare no conflict of interest.

Correspondence to: Yi Li, PhD, Department of Biostatistics, 1415 Washington Heights, M2202, University of Michigan, Ann Arbor, MI 48109. E-mail: yili@umich.edu.

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.lww-medicalcare.com.

DOI: 10.1097/MLR.0000000000002116

National quality measure endorsement processes typically evaluate metrics on the above criteria,[12] and measures may be endorsed on importance but fail on scientific acceptability.[13,14] To date, these measure evaluations are performed individually. However, agencies often use quality measures to form composite metrics, which can be useful for differentiating providers based on metrics reflecting different aspects of primary and intermediate clinical outcomes or care processes.[5,15,16] Such composite metrics include those developed by the Centers for Medicare and Medicaid Services (CMS) public reporting and value-based purchasing programs.[17–21] This study illustrates how the statistical overall reliability of a composite metric relates to the reliabilities of the individual measures comprising the composite, as understanding these relationships provides additional information when evaluating measures for endorsement.

CMS's Quality of Patient Care Star Ratings for dialysis facilities is a composite measure representing overall care quality. It includes various aspects like preventing hospitalizations and deaths, avoiding unnecessary transfusions, effective bloodstream access, waste removal, and balanced bone minerals, summarized by 8 specific measures reported on Medicare's Dialysis Facility Care Compare.[22,26,27] A composite score for each facility based on the 8 quality measures is first calculated and then mapped onto ratings ranging from 1 to 5 stars (with 1, 3, and 5 representing much below average, average, and much above average care quality, respectively) to be interpretable to the general public. Stakeholders include ESRD care consumers (patients/caregivers), providers (dialysis facilities), and payers/regulators (Medicare/CMS, private insurers). While all stakeholders benefit from the Star Ratings, its primary goal is to offer patients and caregivers an easy-to-use, balanced summary of facility quality to inform their care decisions.[17,23–25] We justify the Star Ratings' reliability by demonstrating that the composite measure is more reliable than its individual measures. This approach is also relevant to other health care public reporting programs using composite summaries.

METHODS

We calculate the overall reliability of a composite measure and examine the relationship between overall and individual measure reliabilities in the Star Ratings as a real-world example.

Individual Measure Reliability

The individual quality metrics are defined at the facility level based on measurements taken at the patient level. The ICC definition of reliability is the proportion of the total variation in the facility-level measure that can be attributed to variation between facilities.

$$\frac{\text{Between-FacilityVariation}}{\text{Between-FacilityVariation} \pm \text{Within-FacilityVariation/FacilitySize}}$$

Larger values indicate that a larger portion of a measure's variation is due to between-facility differences, and the measure is interpreted as more reliably distinguishing facility differences. Smaller values indicate more of a measure's variation is random, and the measure is less reliable.

Theoretical Relationships Between Overall and Individual Measure Reliability

Given the ICC for a single measure, we can consider the combined reliability of 2 or more measures (ie, "overall reliability"). The reliability of a composite, r, comprised of K measures is

$$r = \frac{\sum_{k=1}^{K} w_k^2 r_k + \sum_{k=1}^{K} \sum_{k' \neq k} w_k w_{k'} \rho_{k,k'}}{\sum_{k=1}^{K} w_k^2 + \sum_{k=1}^{K} \sum_{k' \neq k} w_k w_{k'} \rho_{k,k'}}, \quad (1)$$

where r_k is the reliability of the kth measure ($k = 1, \ldots, K$), w_k is the kth measure's weight within the composite measure, and $\rho_{k,k'}$ is its pairwise correlation with another measure, k'.[28–31] See the Appendix (Supplemental Digital Content, http://links.lww.com/MLR/C934) for the derivation of (1) and numerical illustrations of various hypothetical scenarios.

Reliability Comparisons in the Star Ratings

The Star Ratings consist of 8 quality metrics, scored either as standardized ratios or rates. The standardized mortality ratio (SMR), standardized hospitalization ratio (SHR), standardized readmission ratio (SRR), and standardized transfusion ratio are defined as the number of observed events (mortality, hospitalizations, readmissions, and blood transfusions, respectively) in a facility divided by the number of expected events based on that facility's patient mix, with higher values indicating worse performance. The 4 rate-based measures are defined as the proportion of patients in a facility meeting a defined threshold. These include 2 measures of hemodialysis vascular access—the standardized fistula rate, defined as the adjusted rate of adult patient-months using an arterial venous fistula as the sole means of vascular access (a positive outcome), and the long-term catheter rate, defined as the rate of prolonged use of a tunneled catheter (a negative outcome)—and 2 measures of facility processes. The process measures include total Kt/V, which measures dialysis adequacy as the proportion of patients meeting a prespecified threshold for "adequate" small solute clearance of urea nitrogen based on the patient's age (adult vs pediatric) and treatment modality (hemodialysis vs peritoneal dialysis), and hypercalcemia, the proportion of patients with hypercalcemia (blood serum or plasma calcium > 10.2 mg/dL) taken over a 3-month rolling average. Additional details regarding the definition of individual quality measures can be found in the ESRD measures manual.[23]

The quality measures are first normalized to standardized measure scores. Certain measures (eg, mortality) are realigned to follow the paradigm that higher scores mean better quality.[25] Correlated standardized measure scores are empirically grouped into 3 groups, or

"domains," of care through factor analysis. Factor analysis identifies which measures are more correlated. Those most correlated are grouped into the same domains. The 3 resulting domains each generally represent a different aspect of care quality. Specifically, SMR, SHR, SRR, and standardized transfusion ratio form one domain, reflecting primary outcomes, standardized fistula rate and long-term catheter form a second domain, reflecting vascular access management, and total Kt/V and Hypercalcemia form a third domain, reflecting waste removal/dialysis management. Each facility receives domain-specific scores by averaging the within-domain standardized measure scores, as well as a final score, which is aggregated from these domain scores. Final scores are translated into ratings (1–5 stars) by comparing facility scores to thresholds established using historical, baseline data (for detailed methodology, see the technical notes).[25] Supplemental Table S1 (Supplemental Digital Content, http://links.lww.com/MLR/C934) reports quality measures, their within-domain weights, and the weights of their respective domains in calculating a facility's final score and subsequent Star Rating. We also report correlations between standardized measure scores in Table 1. These correlations confirm that some measures are more closely related than others, namely measures within the same domain, though all have some degree of positive correlation.

For our analyses in this work, we utilize only the reliabilities and pairwise correlations of the 8 individual quality measures from the calendar year 2019, that is, the last data collected before the start of the COVID-19 pandemic, to mitigate confounding. All individual measure reliability values have been publicly reported in their respective National Quality Forum measure endorsement testing forms (downloadable from https://www.qualityforum.org/qps/). The exception is hypercalcemia and total Kt/V. For these 2 measures, we calculate the individual measure reliabilities using patient and facility-level data from dialysis facility claims data and EQRS data (not publicly reported). The individual measure reliabilities are provided in the first row of Table 2. Pairwise correlations between all measures were obtained from the Star Rating technical notes and are provided in Table 1.[25] Using reliabilities and pairwise correlations of the individual measures, we calculate the overall reliability of the Star Rating composite final score. We also calculate the overall reliability within each of the 3 domains. Finally, as an illustrative exercise, we calculate how the overall reliability would change if the Star Ratings were hypothetically initially composed of only one measure and then other measures were sequentially added to the measure set.

RESULTS

We apply the combined reliability calculation to the measures in the dialysis facility Star Ratings. We report the combined reliability estimates for domain-specific scores and final facility scores in Table 2. As the total Kt/V measure is a combination of 4 other measures, we calculate its reliability using an approximate ANOVA approach in the same spirit as the other measures. As shown, the combined reliability for each domain (0.71, 0.82, and 0.92, respectively) is estimated to be at least as high as their individual measures' reliabilities. In addition, the reliability of the final scores (0.85) is higher than the individual measures, apart from the dialysis process measures. This may be due to an overestimation of the variance explained for these highly skewed measures when using traditional (ie, model-based) estimates for the variance components. To remedy this, Table 2 also reports the variance explained for the domain scores and final scores under an alternate, rank-based calculation for skewed measures, though the variance explained by a facility's final score remained the same (0.85).

Table 2 also reports the changes to the combined reliability of the composite score that would occur if we were to hypothetically add the Star Rating measures sequentially to the measure set. In practice, the Star Rating measure set includes all measures described previously. We provide an illustrative example to demonstrate how the overall reliability would hypothetically change if the Star Ratings originally consisted of only one measure, SMR, and then more measures were added one by one. The results of this illustration can be connected to the numerical illustrations described in the Appendix and Supplemental Figures S1 and S2 (Supplemental Digital Content, http://links.lww.com/MLR/C934). Going across the second rows under the "moment-based estimators for all domains" and "rank-based estimates for the dialysis processes" domain sections of Table 2, we see that if SMR were the only measure in the Star Ratings, the reliability of the Star Rating final score would be 0.5 (the same as SMR's individual reliability). Then, if SHR was added to the measure set, the combined reliability of SMR and SHR would be 0.59 [which is higher than both SMR's individual reliability (0.5) and SHR's individual reliability (0.53)]—this is an example of a region (I) scenario as described in the numerical illustration represented by Supplemental Figure S1. If SRR was added next, we see that the overall reliability of SMR, SHR, and SRR is 0.64, despite SRR having a relatively lower individual reliability (0.39) than SMR and SHR. This situation is an example corresponding to the region (II) in Supplemental Figure S1. Continuing across the row in the table, we see that each measure addition would be an example of a region (I) or region (II) scenario; each time a measure is added, there are gains in the overall reliability. As shown, since the measures are all weakly correlated and have non-zero individual reliability, the combined reliability increases as more measures are added, a similar result to our illustrations.

DISCUSSION

Composite metrics are often created to be interpretable and comprehensive summaries of overall care quality that can be used by stakeholders to differentiate providers. By drawing attention to the differences in care quality between providers, composite measures can pro-

TABLE 1. Pairwise Correlations for the Standardized Clinical Quality Measure Scores of the Star Ratings

Measure	SMR*	SHR†	SRR‡	STrR§	SFR‖	LTC¶	Kt/V#	Hyp.**
SMR*	1.00	—	—	—	—	—	—	—
SHR†	0.20	1.00	—	—	—	—	—	—
SRR‡	0.10	0.41	1.00	—	—	—	—	—
STrR§	0.17	0.35	0.17	1.00	—	—	—	—
SFR‖	0.08	0.14	0.08	0.09	1.00	—	—	—
LTC¶	0.09	0.18	0.08	0.11	0.46	1.00	—	—
Kt/V#	0.13	0.19	0.09	0.15	0.18	0.26	1.00	—
Hyp.**	0.08	0.09	0.03	0.05	0.16	0.22	0.28	1.00

*Standardized mortality ratio for dialysis facilities.
†Standardized hospitalization ratio for dialysis facilities.
‡Standardized readmission ratio for dialysis facilities.
§Standardized transfusion ratio for dialysis facilities.
‖Hemodialysis vascular access: standardized fistula rate.
¶Hemodialysis vascular access: long-term catheter rate.
#Total Kt/V measure.
**Hyp: Proportion of patients with hypercalcemia.
LTC indicates long-term catheter; SFR, standardized fistula rate; SHR, standardized hospitalization ratio; SMR, standardized mortality ratio; SRR, standardized readmission ratio; STrR, standardized transfusion ratio.

mote improvements in care quality. For example, in the Star Rating program, to facilitate improved performance, facilities are provided with not only their categorical star rating but also a detailed breakdown of their composite score so they can compare their performance on the individual quality measures to other facilities' performance and identify which specific outcomes to focus on improving. In a more general sense, regulators of care quality (such as CMS/Medicare) can implement incentives based on composite metrics to motivate providers to improve their care quality. Evaluating the reliability of a composite metric is important to ensure that the composite metric is statistically sound. We derived the overall reliability of a composite metric based on the reliabilities and pairwise correlations of the individual measures and explored how adding more individual measures to a composite could impact the overall reliability in different situations.

Quality measure evaluation and endorsement processes often act on stand-alone measures. Despite their potential, quality metrics that are important to measure, demonstrate a high level of variability among providers and are clinically sound may fail the endorsement process based on low measurement reliability or validity.[32] However, such measures may add utility to a composite rating system under certain circumstances—a consideration which should, in part, factor into their assessment. For example, an individual measure that is deemed to have low reliability may still increase the reliability of the composite, adding value to its overall assessment of care quality. This manuscript described some of the conditions under which a measure with lower perceived reliability would nonetheless enhance a composite measure by increasing its ability to differentiate provider performance.

We highlight several specific results regarding the relationship between the reliability of individual dialysis facility quality measures and their combined reliability in the Star Ratings. First, the proportion of variance explained for each domain of care surpassed those of the individual measures within the domain. For example, the mortality, hospitalization, readmission, and blood transfusion measures had reliability estimates ranging from 0.39 to 0.65, while the combined variance explained for their domain was 0.71. This result reinforces the rationale for constructing domains in the development of the Star Rating. As the combined reliability is a function of the individual measures and their pairwise correlations, we expect the variance explained for each domain to exceed the individual measures' if they are indeed correlated. Though some measures have moderate individual reliability, we show that combining them increases the overall reliability. Thus, in addition to importance and scientific acceptability at the individual measure level, the evaluation of a quality measure for inclusion in a composite score should consider its impact on the reliability of the composite. The composite proportion of variance explained for the dialysis facility Star Ratings was 0.85, further exceeding that of the individual measures for all but the dialysis process measures (total Kt/V and hypercalcemia).

Austin and colleagues (2019) yielded similar conclusions in a different context. In the simulation, they pooled 3 binary indicators with low ICC into a composite indicator. They found that creating a composite indicator with high reliability was possible when the individual components were moderately correlated.[33] We draw similar conclusions both analytically and through illustrations of simulated and real data.[34] Rudner (2001) provides 2 key results when considering just 2 measures: (1) the lowest possible reliability value is that of the less reliable measure and (2) if the measures are equally reliable, composite reliability is maximized with equal measure weighting.[35] These results extend to multiple measures, as we can always add a new measure to an existing composite.

If a new measure is not strongly correlated with existing measures, such as one developed for a previously uncaptured aspect of care, adding that measure could reduce the composite reliability. This has substantive

TABLE 2. Combined Reliability (Proportion of Variance Explained) Calculations for the Domain-Specific Scores and Final Score for the Star Ratings

	Primary Outcomes				Vascular Access Management		Waste Removal/Dialysis Management	
	SMR*	SHR†	SRR‡	STrR§	SFR‖	LTC¶	Kt/V#	Hyp.**
Reliability using moment-based estimators for all domains								
Individual measure	0.50	0.53	0.39	0.65	0.74	0.76	0.92	0.87
Cumulative, sequentially adding measures	0.50	0.59	0.64	0.71	0.76	0.81	0.84	0.85
Domain-specific	0.71	0.71	0.71	0.71	0.82	0.82	0.92	0.92
Final score	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
Reliability using rank-based estimates for the dialysis processes domain								
Individual measure	0.50	0.53	0.39	0.65	0.74	0.76	0.86	0.83
Cumulative, sequentially adding measures	0.50	0.59	0.64	0.71	0.76	0.81	0.84	0.85
Domain-specific	0.71	0.71	0.71	0.71	0.82	0.82	0.88	0.88
Final score	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85

*Standardized mortality ratio for dialysis facilities.
†Standardized hospitalization ratio for dialysis facilities.
‡Standardized readmission ratio for dialysis facilities.
§Standardized transfusion ratio for dialysis facilities.
‖Hemodialysis vascular access: standardized fistula rate.
¶Hemodialysis vascular access: long-term catheter rate.
#Total Kt/V measure.
**Proportion of patients with hypercalcemia.
LTC indicates long-term catheter; SFR, standardized fistula rate; SHR, standardized hospitalization ratio; SMR, standardized mortality ratio; SRR, standardized readmission ratio; STrR, standardized transfusion ratio.

implications but may not preclude adding the measure to the composite score (Supplemental Table S2, Supplemental Digital Content, http://links.lww.com/MLR/C934). If quality is multidimensional, not including such a measure may create incentives for the facility to skimp on that aspect of care. This leads to an important question to address: if a new measure does not substantially harm the overall reliability, but adds diversity in capturing a different aspect of care, should the measure be included in the overall measure set? It is our hope that this and related work provide a catalyst for such discussions when considering new measures for endorsement and public reporting.

Our work has several strengths. First, this is a national study based on over 7000 Medicare-certified dialysis facilities providing care to over 500,000 patients with end-stage renal disease. We apply the proportion of variance explained methods to a composite measure that is based on classic test theory. Lastly, this method can be generalized to other national composites, and though it is structured differently, it would be interesting to see if the results presented here hold for other dialysis facility-related composites such as the total performance scores reported for the ESRD Quality Improvement Program or for other provider types such as the overall hospital star rating.

There are also some limitations to our work. First, as a general comment, using composites to measure care quality is sometimes questioned, as there is an inherent tradeoff between interpretability and information loss. Composite metrics like the Star Ratings are often designed and intended to be a simple and balanced summary of care quality across multiple care quality domains. Interpretability of the Star Ratings is further enhanced by categorizing the composite score into ratings ranging from 1 to 5 stars. This can provide patients, caregivers, and other stakeholders with a simple way to determine a facility's overall performance relative to others. Composites are often more easily interpreted by the general public than are individual quality measures. However, discerning the source of differences in quality from the composite measure may be difficult.[36] For example, if there are differences in 2 facilities' Star Ratings, for example, 2 versus 3 stars, it is difficult for the user to know what measure(s) in the composite is driving the difference between the two facilities. Further, the overall reliability we calculated is based on the continuous final score underlying a facility's Star Rating. Extending this work to a facility's categorical rating would better align with the goals of the public reporting program. Third, we base our work on the ICC coefficient, a commonly accepted reliability metric used in the endorsement of health care quality metrics at the national level. This assumes that the health care quality metrics are normally distributed and can be biased when this assumption does not hold. Other nonparametric approaches may be warranted for future consideration; however, the concepts presented here can easily be extended to other measurements of reliability as well. Fourth, we recognize that between-facility variation, which is often used in the signal-to-noise ratio definition of reliability, often includes unmeasured/unaccounted variation in patient case mix, provider actions, and external factors that may influence outcomes. However, in the case of race/ethnicity, differences in outcomes may not be due to race, but rather other (underlying) sources that could include provider treatment and historical disparities related to race. It is therefore difficult to sustain the assumption that all sources of variation in patient and

provider mix can be measured and accounted for in an individual or composite measure.[37] Sorting out this important issue is beyond the scope of this paper. Lastly, we limit our analysis to pre-COVID data to avoid introducing additional factors that may systematically affect the quality of care provided by facilities during the pandemic. Further study and an updated analysis with post-COVID data, when available, are warranted to validate our initial findings.

We note that validity is an important consideration when determining the scientific soundness of any measure, including composite measures. However, our focus is solely on the reliability aspect of composite measures. This study assumes some level of empirical validity has been achieved for all individual measures used in a composite measure. In our Star Ratings example, each of the 8 individual measures was reviewed and received initial endorsement by the National Quality Forum and thus went through rigorous validity testing during its development. It remains unclear whether adding more measures would enhance composite validity, another important but distinct criterion of scientific soundness. In addition, while a measure may achieve validity and may measure what it is intended to measure, a valid measure can still be unreliable and therefore not provide a stable assessment of performance variation between providers. Exploring the validity of composite measures in more detail is an important direction for future work.

CONCLUSIONS

Individual quality measures may be useful in composite reporting, even if certain criteria for consensus endorsement are not met. This depends on the measure's relationship with other measures in the composite, and their ability to inform the public about aspects of care that may be important to measure. This work highlights the utility of combined reliability for composite measures of care quality. Based on our results, it may be useful to add measures with lower ICC values to composites, if they add diversity and robustness to the overall rating system. Further, from this perspective, we advocate that the scientific acceptability of new measures or subsequent inclusion in public reporting programs should be considered with respect to the proportion of variance explained for the individual measure, as well as the additional aspect of care quality the measure brings to a composite set. As developing measures encompassing several aspects of care is necessary for creating a robust composite, we hope this work will further justify incorporating new measures into the Star Ratings and provide a stimulus for needed discussion about how individual measures are used in the larger scope of public reporting.

REFERENCES

1. Brook RH, McGlynn EA, Cleary PD. Quality of health care. Part 2: measuring quality of care. *N Engl J Med*. 1996;335:966–970.
2. Derose SF, Petitti DB. Measuring quality of care and performance from a population health care perspective. *Annu Rev Public Health*. 2003;24:363–384.
3. Hanefeld J, Powell-Jackson T, Balabanova D. Understanding and measuring quality of care: dealing with complexity. *Bull World Health Organ*. 2017;95:368–374.
4. Schold JD, Nicholas LH. Considering potential benefits and consequences of hospital report cards: what are the next steps? *Health Serv Res*. 2015;50:321–329.
5. Shwartz M, Restuccia JD, Rosen AK. Composite measures of health care provider performance: a description of approaches. *Milbank Q*. 2015;93:788–825.
6. Donabedian A. The quality of care. How can it be assessed? *JAMA*. 1988;260:1743–1748.
7. Glance LG, Joynt Maddox K, Johnson K, et al. National Quality Forum Guidelines for evaluating the scientific acceptability of risk-adjusted clinical outcome measures: a report from the National Quality Forum Scientific Methods Panel. *Ann Surg*. 2020;271:1048–1055.
8. McGlynn EA. Selecting common measures of quality and system performance. *Med Care*. 2003;41(suppl 1):I39–I47.
9. Siu AL, McGlynn EA, Morgenstern H, et al. Choosing quality of care measures based on the expected impact of improved care on health. *Health Serv Res*. 1992;27:619–650.
10. McGlynn EA, Adams JL. What makes a good quality measure? *JAMA*. 2014;312:1517–1518.
11. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep*. 1966;19:3–11.
12. O'Brien JM, Corrigan J, Reitzner JB, et al. Will performance measurement lead to better patient outcomes? What are the roles of the National Quality Forum and Medical Specialty Societies? *Chest*. 2012;141:300–307.
13. Hermann RC, Provost S. Best practices: interpreting measurement data for quality improvement: standards, means, norms, and benchmarks. *Psychiatr Serv*. 2003;54:655–657.
14. Davies HT, Crombie IK. Interpreting health outcomes. *J Eval Clin Pract*. 1997;3:187–199.
15. Guthrie B. Measuring the quality of healthcare systems using composites. *Brit Med J*. 2008;337:a639.
16. Ramasubramanian H, Joshi S, Krishnan R. Wisdom of the experts versus opinions of the crowd in hospital quality ratings: analysis of hospital compare star ratings and Google star ratings. *J Med Internet Res*. 2022;24:e34030.
17. Salerno S, Dahlerus C, Messana J, et al. Evaluating national trends in outcomes after implementation of a star rating system: results from dialysis facility compare. *Health Serv Res*. 2021;56:123–131.
18. Shwartz M, Ren J, Pekoz EA, et al. Estimating a composite measure of hospital quality from the Hospital Compare database: differences when using a Bayesian hierarchical latent variable model versus denominator-based weights. *Med Care*. 2008;46:778–785.
19. Young E, Ding Z, Kapke A, et al. ESRD QIP payment reductions are associated with mortality, utilization, and cost. *Health Serv Res*. 2020;55:96–97.
20. Jacobs R, Smith P, Goddard M. Measuring performance: an examination of composite performance indicators. Working Paper Series of University of York, 2004.
21. Johnson MA, Normand S-LT, Krumholz HM. How are our hospitals measuring up? "hospital compare": a resource for hospital quality of care. *Circulation*. 2008;118:e498–e500.
22. Salerno S, Gremel G, Dahlerus C, et al. Understanding the tradeoffs between travel burden and quality of care for in-center hemodialysis patients. *Med Care*. 2022;60:240–247.
23. Centers for Medicare and Medicaid Services.*CMS ESRD Measures Manual for the 2021 Performance Period.* Online, 2020. https://www.cms.gov/files/document/esrd-measures-manual-v61.pdf.
24. Centers for Medicare and Medicaid Services. *End-Stage Renal Disease Dialysis Facility Compare Star Ratings Technical Expert Panel: Summary Report.* In-Person Meeting, Baltimore, MD2017. https://dialysisdata.org/sites/default/files/content/ESRD_Measures/2019_ESRD_DFC_Star_Rating_TEP_Summary_Report.pdf.
25. University of Michigan Kidney Epidemiology and Cost Center. *Technical Notes on the Dialysis Facility Compare Star Rating Methodology for the October 2018 Release.* 2018. https://dialysisdata.org/sites/default/files/content/Methodology/Updated_DFC_Star_Rating_Methodology_for_October_2018_Release.pdf.

26. He K, Kalbfleisch JD, Li Y, et al. Evaluating hospital readmission rates in dialysis facilities; adjusting for hospital effects. *Lifetime Data Anal*. 2013;19:490–512.
27. Kalbfleisch J, Wolfe R, Bell S, et al. Risk adjustment and the assessment of disparities in dialysis mortality outcomes. *J Am Soc Nephrol*. 2015;26:2641–2645.
28. Wang MW, Stanley JC. Differential weighting: a review of methods and empirical studies. *Rev Educ Res*. 1970;40:663–705.
29. Brennan RL. An essay on the history and future of reliability from the perspective of replications. *J Educ Meas*. 2001;38:295–317.
30. Thissen D, Wainer H. *Test Scoring* 1st ed. Routledge; 2001.
31. Webb NM, Shavelson RJ, Haertel EH. 4 reliability coefficients and generalizability theory. *Handbook Stat*. 2006;26:81–124.
32. All-Cause Admissions and Readmissions. Battelle. Partnership for Quality Management Web site. Published 2023. Accessed June 14, 2023. https://p4qm.org/projects/All-Cause-Admissions-Readmissions
33. Austin PC, Ceyisakar IE, Steyerberg EW, et al. Ranking hospital performance based on individual indicators: can we increase reliability by creating composite indicators? *BMC Med Res Methodol*. 2019;19:131.
34. Zaslavsky AM, Shaul JA, Zaborski LB, et al. Combining health plan performance indicators into simpler composite measures. *Health Care Financ Rev*. 2002;23:101–115.
35. Rudner LM. Informed test component weighting. *Educ Meas Issues Pract*. 2001;20:16–19.
36. Friebel R, Steventon A. Composite measures of healthcare quality: sensible in theory, problematic in practice. *BMJ Qual Saf*. 2019;28:85–88.
37. Kalbfleisch JD, He K, Xia L, et al. Does the inter-unit reliability (IUR) measure reliability? *Health Serv Outcomes Res Method*. 2018; 18:215–225.