Published in final edited form as:

Am J Med Qual. 2018 ; 33(3): 321–322. doi:10.1177/1062860617721645.

Exploring the Evidence Base Behind Quality Measures

Ezinne Eze-Ajoku, MBBS, MPH,

Johns Hopkins Bloomberg School of Public Health

Melissa Lavoie, and

Johns Hopkins University School of Medicine

Matthew DeCamp, FACP MD PhD¹

Assistant Professor, Johns Hopkins Berman Institute of Bioethics and Division of General Internal Medicine, Deering Hall, 1809 Ashland Avenue, Baltimore, MD 21205

Quality measures pervade medicine. Thousands exist and are used to rate hospital quality, to evaluate physician performance, and to determine financial reimbursement, among other uses.¹ Quality measurement is therefore inextricably linked with clinical practice. Yet recently, half of primary care physicians perceived quality measures as negatively impacting the care they deliver.²

Colloquially among physicians, one hears various explanations for this negative sentiment (e.g., “reporting burden,” “measures don’t apply to my panel,” “measures don’t reflect real quality,” and so on). One sometimes hears that measures are not “evidence-based.” If true, this would be a major obstacle to physician buy-in to quality measurement. The stronger the evidence base behind a measure, the stronger the ethical obligation for physicians to change practice patterns, motivate specific decisions or actions by their patients, and support quality improvement activities locally and nationally.³ It therefore seems reasonable to ask: What is the evidence base behind quality measures?

To help answer this question, we examined the strength of evidence behind quality measures used in Medicare’s 2016 Shared Savings Program. These measures apply to over 430 accountable care organizations (ACOs) covering nearly 8 million Medicare beneficiaries. Medicare prefers quality measures endorsed by the National Quality Forum (NQF), which are considered “best in class.”⁴ When measure developers submit to NQF, they include evidence forms documenting evidence in support of the measure’s use.

We collected evidence forms for each of the 34 measures. Each form was reviewed independently (by EEA and ML), and the evidence grade(s) behind a measure were collected. We sought to categorize grades charitably, favoring the highest applicable grades. Areas of disagreement were resolved by consensus among all authors.

Reviewers also collected the system or systems used to grade that evidence. NQF encourages, but does not require, use of a formal grading system, such as the United States

¹Corresponding author: (410) 614-2187 (phone), mdecamp1@jhmi.edu.

COI: The authors have no potential conflicts of interest to report related to the content of this manuscript.

Preventive Services Task Force (USPSTF). When a formal grading system was not used, we collected the developer's overall assessment of the evidence (high, moderate, low, or inconsistent).

Next we categorized the reported evidence grades into high, moderate, or mixed categories using the following rubric: "High" meant that the reported evidence was unequivocally at the top of the scale relative to the grading system used. "Moderate" meant that the reported evidence was moderate or inconsistent. "Mixed" meant that reported evidence differed in grades by more than one grade level. For illustration, the USPSTF grading system was used for several measures. A USPSTF grade of "A" would qualify as "High" overall strength in our analysis, because "A" is the top of the USPSFT scale (representing high certainty of substantial benefit). A grade of "B" would correspond to "Moderate" overall strength, because there is either high certainty of moderate benefit or moderate certainty of substantial benefit. Reported grades of "A" and "C" would be categorized as "Mixed" strength. Although we initially included a category of "Low" strength, no measures fit this category.

Of 34 measures (32 individual measures and a composite of two diabetes measures), 23 were NQF endorsed. The 11 non-endorsed measures included certain patient experience survey measures (n=5); claims-based measures related to unplanned admissions and meaningful use of electronic records (n=4); and preventive measures related to blood pressure screening and statin therapy (n=2). Developer-reported evidence assessments were available for 16/23 endorsed measures. Most were process measures, because outcome measures do not require formal assessment.

Differences existed in grading systems used and the evidentiary strength. Relative to the method of grading, we categorized 7 as "High" strength, 6 as "Moderate," and 3 as "Mixed."

Based on average ACO performance, performance appeared to be lower in the moderate evidence category (overall average, 61%, range 44–85%) compared to the high evidence category (overall average, 77%; range 63–91%).

On the one hand, our observations reassure practicing physicians. NQF-endorsed measures undergo formal evaluation that involves assessment of the evidence base, and most are graded moderate or above. On the other hand, physicians may be concerned that not all measures are NQF-endorsed or formally graded and that the grading systems differ or may be unfamiliar.

These observations suggest ways to improve performance measurement. Our findings echo prior calls for measure endorsers to require evaluation by a standardized grading system.^{5,6} GRADE may be perceived as more rigorous and better tailored to translating evidence into practice recommendations,⁷ though other systems, including the USPSTF, have advantages as well.⁸ As measures increase in number and their impact grows, the case for more standardized and rigorous assessment gets stronger.

In addition, trust in performance measurement might be improved were an overall strength of evidence statement or assessment available for individual measures in a concise, user-friendly format both publicly and at the point-of-care. Although we were able to attain

information via the raw evidence forms, this was arduous. When involved in quality improvement activities, the busy physician should be able to trust endorsement.

We hesitate to draw firm conclusions. After all, we collected information only within a single program, albeit a large and influential one. We relied on measure developers' assessment of the evidence, which may overestimate current measure evidence, and evidence is only one component of measure evaluation. NQF also evaluates measure importance, feasibility, usability, and the presence of competing measures. Past some minimal threshold, if a measure addresses a very important health issue, physicians might act on a lower level of evidentiary support. And we cannot infer causation between evidence and ACO performance, especially in light of the many behavioral influences beyond evidence alone (e.g., organizational priorities, incentives, task complexity of particular measures, patient choice,⁹ and so on).

Nevertheless, we believe our observations should motivate research to understand physicians' perceptions of quality measures. Prior surveys suggested that physicians question whether measures accurately reflect quality of care.¹⁰ Explicit attention should be given to how physicians' perceptions of the evidence affects their practice patterns and their sense of professional obligation to conform to a particular measure. Quality measurement has the potential to improve patient care, but only if everyone shares a commitment to, and trust in, the measures used.

Acknowledgments

FUNDING: This work was supported by the Agency for Healthcare Research and Quality (5K08HS023684).

References

1. Blumenthal, D., Malphrus, E., McGinnis, JM., Committee on Core Metrics for Better Health at Lower Cost. Vital Signs: Core Metrics for Health and Health Care Progress. Washington, DC: National Academies Press; 2015.
2. Commonwealth Fund/Kaiser Family Foundation. National Survey of Primary Care Providers. Primary Care Providers' Views of Recent Trends in Health Care Delivery and Payment. 2015. Available from: https://kaiserfamilyfoundation.files.wordpress.com/2015/08/1831_commonwealth_kaiser_primary_care_survey_final1.pdf. Accessed May 25, 2017
3. Qaseem A, Snow V, Gosfield A, Gregg D, Michl K, Wennberg D, et al. Pay for Performance Through the Lens of Medical Professionalism. *Ann Intern Med*. 2010; 152:366–369. [PubMed: 20231567]
4. National Quality Forum. Field Guide to NQF Resources. Available from: http://www.qualityforum.org/Field_Guide/NQF_Endorsement.aspx. Accessed May 25, 2017
5. Baker DW, Qaseem A, for the American College of Physicians' Performance Measurement Committee. Evidence-Based Performance Measures: Preventing Unintended Consequences of Quality Measurement. *Ann Intern Med*. 2011; 155:638–640. [PubMed: 22041954]
6. National Quality Forum. Guidance for Evaluating the Evidence Related to the Focus of Quality Measurement and Importance to Measure and Report. Jan, 2011. Available at: https://www.qualityforum.org/Measuring_Performance/Improving_NQF_Process/Evidence_Task_Force_Final_Report.aspx. Accessed May 25, 2017
7. Cuello-García CA, Dávalos-Rodríguez ML. Perceptions and attitudes towards different grading systems from clinical guidelines developers. *J Eval Clin Pract*. 2009; 15(6):1074–1076. DOI: 10.1111/j.1365-2753.2009.01319.x [PubMed: 20367707]

8. Gopalakrishna G, Langendam MW, Scholten RJ, Bossuyt PM, Leeflang MM. Guidelines for guideline developers: a systematic review of grading systems for medical tests. *Implement Sci.* 2013; 8:78.doi: 10.1186/1748-5908-8-78 [PubMed: 23842037]
9. Walter LC, Davidowitz NP, Heineken PA, Covinsky KE. Pitfalls of converting practice guidelines into quality measures: Lessons learned from a VA performance measure. *JAMA.* 2004; 291(20): 2466–2470. DOI: 10.1001/jama.291.20.2466 [PubMed: 15161897]
10. Casalino LP, Alexander GC, Jin L, Konetzka RT. General internists' views on pay-for-performance and public reporting of quality scores: a national survey. *Health Aff (Millwood).* 2007; 26:492–9. [PubMed: 17339678]

Table

Evidence Supporting Measures Used in Medicare Accountable Care Organizations

ACO Measure	NQF Number	Description	Type	Evidence Grade		ACO Average Performance [†]
				Formal Grade	Developer's Assessment [*]	
ACO-15 Preventive Care	NQF0043	Pneumococcal vaccination	Process		High	68%
ACO-16 Preventive Care	NQF0421	Adult weight screening and follow-up	Process	SORT: A		71%
ACO-17 Preventive Care	NQF0028	Tobacco use assessment and cessation	Process	US PHS: A		91%
ACO-19 Preventive Care	NQF0034	Colorectal cancer screening	Process	USPTF: A (high)		63%
ACO-39 Care Coordination & Patient Safety	NQF0419	Documentation of current medications	Process	GRADE: High		84%
ACO-31 At-risk Population	NQF0083	Beta-blocker for left ventricular systolic dysfunction (LVSD)	Process	AHA: A		88%
ACO-33 At-risk Population	NQF0066	ACE inhibitor or ARB therapy for Diabetes or LVSD	Process	ACC/AHA: A		79%
ACO-10 Care Coordination & Patient Safety	NQF0277	Ambulatory Sensitive Conditions: Heart Failure Admissions	Process	UCSF-Stanford EPC: 14 (range: 0–26)		–
ACO-30 At-risk Population	NQF0068	Ischemic Vascular Disease: aspirin or other antithrombotic	Process	ACC/AHA: A-B		85%
ACO-14 Preventive Care	NQF0041	Influenza immunization	Process		Moderate	64%
ACO-18 Preventive Care	NQF0418	Depression screening	Process	SORT: B		46%
ACO-20 Preventive Care	NQF2372	Mammography screening	Process	USPSTF: B (fair)		68%
ACO-41 At-risk Population	NQF0055	Diabetes: Eye Exam	Process	AGS: B ADA: A AAOI: I		44%
ACO-27 At-risk Population	NQF0059	Diabetes: A1C poor control	Outcome	VA/DOD: A, C ACE: A, D AGS: B, ADA: B, C		82% (18% poor control)
ACO-13 Care Coordination & Patient Safety	NQF0101	Falls: Screening for fall risk	Process	USPSTF: B, C (fair) AGS: A, B		58%

ACO Measure	NQF Number	Description	Type	Evidence Grade		ACO Average Performance [†]
				Formal Grade	Developer's Assessment [*]	
ACO-28 At-risk Population	NQF0018	Controlling high blood pressure	Outcome	GRADE: A-E JNC 8: high-moderate		70%

^{*} **Developer's Assessment** of the quality, quantity and consistency of the evidence is reported when no formal grading system was used.

[†] **ACO Performance** is the percentage of Medicare Shared Savings Program patients who met the measure requirements in the 2015 performance year calculated as a weighted average of ACP data from <http://data.cms.gov>. Because ACO-10 is a rate, average performance was not calculated.

	Rating of evidence is high, strong or consistently good, consistently at the highest level of evidence based on the system used or developer's assessment is high in the absence of a formal grade.
	Rating of evidence is moderate, limited, or inconsistent, based on the system used, or developer's assessment is moderate in the absence of a formal grade.
	Rating of evidence is mixed, i.e., reported ratings of evidence differ by more than 1 grade category.

Note: AAO = American Academy of Ophthalmology; ACC = American College of Cardiology; ACE = American College of Endocrinology; ADA = American Diabetes Association; AGS = American Geriatrics Society; AHA = American Heart Association; AMA = American Medical Association; GRADE = Grading of Recommendations Assessment, Development and Evaluation; JNC = Joint National Committee; SORT = Strength of Recommendation Taxonomy; UCSF-Stanford EPC = University of California San Francisco – Stanford Evidence-based Practice Center; USPSTF = United States Preventive Services Task Force, US PHS = United States Public Health Service; VA/DOD = Veterans' Affairs/Department of Defense.