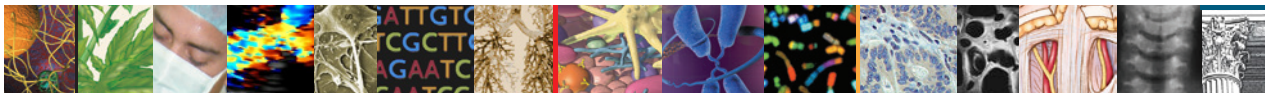

The NEW ENGLAND JOURNAL of MEDICINE

Perspective

MAY 10, 2018

Time Out — Charting a Path for Improving Performance Measurement

Catherine H. MacLean, M.D., Ph.D., Eve A. Kerr, M.D., M.P.H., and Amir Qaseem, M.D., Ph.D., M.H.A.

Performance measurement in the U.S. health care system has expanded dramatically over the past 30 years. The National Quality Measures Clearinghouse now lists more than 2500 performance measures. These measures are used in various quality-reporting, accountability, and payment programs sponsored by commercial payers, government agencies, and independent quality-assessment organizations. The Centers for Medicare and Medicaid Services (CMS) aims to base 90% of Medicare fee-for-service payments to clinicians on "value" by the end of 2018 by using performance scores.

Although most physicians view the delivery of high-quality care as a professional imperative,[1] performance-measurement activities face increasing resistance from physicians and some policymakers who believe that current measures are not meaningful.[2] In a recent survey, 63% of physicians said that current measures do not capture the quality of the care that physicians provide.[3] Yet U.S. physician practices are spending $15.4 billion each year — about $40,000 per physician — to report on performance.[3]

In response to these concerns, the Performance Measurement Committee (PMC) of the American College of Physicians (ACP) developed criteria to assess the validity of performance measures (see box). Using a modified version of the method developed at RAND and UCLA for evaluating the benefits and harms of a medical intervention, we applied the ACP criteria to the measures included in the Medicare Merit-based Incentive Payment System (MIPS)/Quality Payment Program (QPP). We hypothesized that if most of the MIPS/QPP measures assessed were deemed valid using this process, physicians could have more confidence that adherence to the measured practices would result in improved patient outcomes. Conversely, if some substantial proportion of the measures were deemed not valid, the results would suggest the need to change the process by which MIPS measures are developed and selected. (For further details, see the methods section in the Supplementary Appendix, available at NEJM.org.)

Of 271 measures in the 2017 QPP measures list, we identified and rated the validity of 86 that the committee considered relevant to ambulatory general internal medicine. Among these, 32

ACP Measure Review Criteria.

Domain 1. Importance

Meaningful clinical impact: Implementation of the measure will lead to a measurable and meaningful improvement in clinical outcomes.

High impact: Measure addresses a clinical condition that is high-impact (e.g., high prevalence, high morbidity or mortality, high severity of illness, and major patient or societal consequences).

Performance gap: Current performance does not meet best practices, and there is opportunity for improvement.

Domain 2. Appropriate Care

Overuse: Measure will promote stopping use of a test or treatment in general population or individuals where the potential harms outweigh the potential benefits.

Underuse: Measure will encourage use of a test or treatment in general population or individuals in whom the potential benefits outweigh the potential harms.

Time interval: Time interval to measure the intervention is evidence-based.

Domain 3. Clinical Evidence Base

Source: Evidence forming the basis of the measure is clearly defined with appropriate references.

Evidence: Evidence is high-quality, high-quantity, and consistent and represents current clinical knowledge.

Domain 4. Measure Specifications

Clarity — numerator and denominator clearly defined:

- For process measures, numerator includes a specific action that will benefit the patient, and denominator includes well-specified exclusions.
- For outcome measures, numerators detail an outcome that is meaningful to the patient and under the influence of medical care.
- Denominator includes well-specified and clinically appropriate exceptions to eligibility for the measure.

Clarity — all components necessary to implement measure clearly defined

Validity: The measure is correctly assessing what it is designed to measure, adequately distinguishing good and poor quality.

Reliability: Measurement is repeatable and precise, including when data are extracted by different people.

Risk adjustment: Risk adjustment is adequately specified for outcome measures.

Domain 5: Measure Feasibility and Applicability

Attribution: Level of attribution specified in the measure is appropriate (measure ties the outcomes to the appropriate unit of analysis) and is clearly stated.

Physician's control: Performance measure addresses an intervention that is under the influence of the physician being assessed.

Usability: Results of the measure provide information that will help the physician to improve care.

Burden: Data collection is feasible and burden is acceptable (low, moderate, or high)

(37%) were rated as valid by our method, 30 (35%) as not valid, and 24 (28%) as of uncertain validity. We also determined the proportion of the measures that had been developed by the National Committee for Quality Assurance (NCQA) or endorsed by the National Quality Forum (NQF) that were rated as valid by our method. As compared with measures that were not endorsed by these organizations, greater percentages of NCQA-developed and NQF-endorsed measures were deemed valid (59% and 48%, respectively, vs. 27% for nonendorsed measures), and smaller percentages were deemed not valid (7% and 22%, vs. 49% for nonendorsed measures). (For further details on the measure review results, see the tables in the Supplementary Appendix.)

For each measure, the committee rated validity with respect to five domains: importance, appropriateness, clinical evidence, specifications, and feasibility and applicability. Examples of the overall and domain ratings given to individual measures judged to be valid, not valid, and of uncertain validity are shown in the table.

Notably, among the 30 measures rated as not valid, 19 were judged to have insufficient evidence to support them. For example, MIPS measure 181, "Elder Maltreatment Screen and Follow-Up," requires the completion of the Maltreatment Screening tool on the date of an encounter and a documented follow-up plan for all patients 65 years of age or older. Although elder abuse is a serious problem that physicians should appropriately diagnose and report, the U.S. Preventive Services Task Force has found insufficient evidence to warrant routine screening. We believe the substantial resources required to screen large populations of elderly patients for maltreatment and to track follow-up would be better directed at care processes whose link to improved health is supported by more robust evidence.

Another characteristic of measures that were not rated as valid by our method was inadequately specified exclusions, resulting in a requirement that a process or outcome occur across broad groups of patients, including pa-

Ratings for a Sample of Measures.*

Rating	NQF-Endorsed	Steward	Measure	Importance	Appropriateness	Clinical Evidence	Specifications	Feasibility	Rationale
Valid	Yes	NCQA	Avoidance of Antibiotic Treatment in Adults with Acute Bronchitis Numerator: MIPS 116 (NQF 0058)	+	+	+	+	+	Based on appropriate evidence; specifications include appropriate exclusion criteria for patients with COPD or immunocompromised patients.
	Yes	AHA	Chronic Stable Coronary Artery Disease: Antiplatelet Therapy: MIPS 006 (NQF 0067)	+	+	+	+	+	Clinically important process with known performance gap. Specifications limit potential for preventable adverse events by excluding patients receiving warfarin therapy.
	No	AAN	Stroke and Stroke Rehabilitation: Discharged on Antithrombotic Therapy: MIPS 032	+	+	+	+	+	Aligns with principles of high-value care; contributes to improved outcomes; based on high-quality evidence.
Not valid	Yes	NCQA	Anti-depressant Medication Management: MIPS 009 (NQF 0105)	–	–	–	±	±	Indicates a time frame that contradicts evidence-based clinical recommendations. Does not consider patient preferences for switching to alternative treatments. Designed to assess performance by health plans, which have access to clinical management data that clinicians lack.
	Yes	CMS	Pain Assessment and Follow-Up: MIPS 131 (NQF 0420)	±	±	–	±	+	Insufficient evidence to require screening. Implementation could unintentionally promote overuse of opioid therapy. Specifications do not exclude patients who are at risk for the development of opioid use disorders (e.g., patients with a history of substance abuse or alcohol use disorders). Referral to a pain management specialist is not practical in every area of the country.
	No	CMS	Elder Maltreatment Screen and Follow-Up: MIPS 181	±	±	–	±	–	Insufficient evidence that screening reduces harm to warrant routine screening. Target population too broad.
Uncertain validity	Yes	NCQA	Controlling High Blood Pressure: MIPS 236 (NQF 0018)	+	±	±	±	±	Does not stratify patients into well-defined risk groups (by age, coexisting diseases). Defines office measurements as preferred monitoring method, but ambulatory monitoring is preferred for assessing blood-pressure control.
	Yes	ATS	Chronic Obstructive Pulmonary Disease: Long-Acting Bronchodilator Therapy: MIPS 052 (NQF 0102)	+	±	±	±	±	Unclear evidence to differentiate benefit of specific bronchodilator therapy for COPD outcomes. Measure lacks specificity regarding reported symptoms, level of COPD severity to which it applies, and details of recommended bronchodilator therapy. Testing results included have weak reliability. Electronic databases may not capture active symptoms; clinicians may encounter barriers to data retrieval because of proprietary information systems.
	No	AAO-HNS	Adult Sinusitis: Antibiotic Prescribed for Acute Sinusitis: MIPS 331	+	+	+	+	+	Measure should exclude patients who have severe or worsening symptoms within 10 days after onset and who would benefit from earlier, appropriate antibiotic treatment.

* Measures were rated on a 9-point scale according to whether they meet criteria; higher scores were better. A plus sign indicates that the measure meets criteria (rating was 7, 8, or 9); a minus sign indicates that the measure does not meet criteria (rating was 1, 2, or 3); and a plus–minus sign indicates that the measure meets some criteria (rating was 4, 5, or 6). AAN denotes American Academy of Neurology, AAO-HNS American Academy of Otolaryngology-Head and Neck Surgery, AHA American Heart Association, ATS American Thoracic Society, CMS Centers for Medicare and Medicaid Services, MIPS Medicare Merit-based Incentive Payment System, NCQA National Committee for Quality Assurance, NQF National Quality Forum, and QPP Quality Payment Program.

The New England Journal of Medicine

tients who might not benefit. MIPS measure 236, "Controlling High Blood Pressure," for instance, requires that a blood pressure of 140/90 mm Hg or lower be achieved in the clinic setting for all patients. Forcing blood pressure down to this threshold could harm frail elderly adults and patients with certain coexisting conditions.

We also identified measures that were directed at important, evidence-based quality concepts but had poor specifications that might misclassify high-quality care as low-quality care. For example, MIPS measure 009, "Antidepressant Medication Management," assesses whether patients who started taking an antidepressant medication continued taking one at 3 and 6 months after initiation. This measure does not consider patients' reasonable preferences for switching to alternative, evidence-based interventions such as psychotherapy or electroconvulsive therapy after experiencing side effects of antidepressants.

Our analysis identified troubling inconsistencies among leading U.S. organizations in judgments of the validity of measures of physician quality. Although the ACP assessment was limited to a defined set of measures, that set was large and included the vast majority of measures that will be applied to ambulatory care internists as part of the United States' largest physician quality-assessment program for the purpose of accountability. Our findings are striking given that the criteria we used were similar to those used by NQF and CMS. Why the disconnect?

Possible explanations include the methods used to assess measures and the characteristics of the experts who did the assessing. The RAND–UCLA appropriateness method does not classify measures as valid when there are significant disagreements among the panelists. In contrast, the NQF threshold for endorsement is close to a simple majority of panelists (60%). The ACP method thus sets a higher standard for validity. In addition, we would argue that the RAND–UCLA method can be considered more evidence-based than other methods, since favorable clinical outcomes have been demonstrated for patients treated according to standards developed with this method.[4,5]

Our analysis identified troubling inconsistencies among leading U.S. organizations in judgments of the validity of measures of physician quality.

It is also possible that the perspectives of the groups doing the rating contribute to differences in validity ratings. Specifically, NQF convenes multistakeholder groups, whereas the ACP committee is composed exclusively of physicians with expertise in clinical medicine and research. However, analyses of the RAND–UCLA method in which multiple panels were convened to rate identical criteria have demonstrated high levels of agreement across panels for necessary care. Hence, although changing the panel composition might result in some differences in ratings, we would not expect the variation to be large enough to explain why so many NQF-endorsed measures were rated as not valid by the ACP committee.

The fact that only 37% of measures proposed for a national value-based purchasing program were found to be valid with a standardized method has implications for physician-level performance measurement. The use of flawed measures is not only frustrating to physicians but also potentially harmful to patients. Moreover, such activities introduce inefficiencies and administrative costs into a health system widely regarded as too expensive. If developers, assessors, and public and private payers adopted a more rigorous method of assessing measures' validity, potential problems could be identified before the measures were launched. It makes sense for practicing clinicians to participate in the development and review of measures. At the same time, a single set of standards (like those put forth by the National Academy of Medicine for clinical practice guidelines) could be developed that would allow others to evaluate the trustworthiness of performance measures.

We believe that the next generation of performance measurement should not be limited by the use of easy-to-obtain (e.g., administrative) data or function as a stand-alone, retrospective exercise. Instead, it should be fully integrated into care delivery, where

it would effectively and efficiently address the most pressing performance gaps and direct quality improvement. For now, we need a time-out during which to assess and revise our approach to physician performance measurement.

An audio interview with Dr. MacLean is available at NEJM.org

Disclosure forms provided by the authors are available at NEJM.org.

From the Center for the Advancement of Value in Musculoskeletal Care, Hospital for Special Surgery, New York (C.H.M.); the University of Michigan Department of Internal Medicine and Institute for Healthcare Policy and Innovation and the Veterans Affairs Ann Arbor Center for Clinical Management and Research, Ann Arbor (E.A.K.); and the American College of Physicians, Philadelphia (A.Q.); and the ACP Performance Measurement Committee (C.H.M., E.A.K.). The other members of the Performance Measurement Committee were J. Thomas Cross, Jr., Eileen Barrett, Robert Centor, Andrew Dunn, Nick Fitterman, Bruce Leff, Ana María López, Mark Metersky, Robert Pendleton, Stephen D. Persell, Edmondo J. Robinson, Sameer D. Saini, Paul Shekelle, and from the American College of Physicians, Sarah Dinwiddie.

This article was published on April 18, 2018, at NEJM.org.

1. Qaseem A, Snow V, Gosfield A, et al. Pay for performance through the lens of medical professionalism. Ann Intern Med 2010;152:366-9.
2. Berwick DM. Era 3 for medicine and health care. JAMA 2016;315:1329-30.
3. Casalino LP, Gans D, Weber R, et al. US physician practices spend more than $15.4 billion annually to report quality measures. Health Aff (Millwood) 2016;35:401-6.
4. Higashi T, Shekelle PG, Adams JL, et al. Quality of care is associated with survival in vulnerable older patients. Ann Intern Med 2005;143:274-81.
5. Hemingway H, Crook AM, Feder G, et al. Underuse of coronary revascularization procedures in patients considered appropriate candidates for revascularization. N Engl J Med 2001;344:645-54.

DOI: 10.1056/NEJMp1802595

Deployment of Preventive Interventions — Time for a Paradigm Shift

Katherine Pryor, M.D., and Kevin Volpp, M.D., Ph.D.

In 2002, Knowler et al. reported results of a landmark study — a large, randomized, controlled trial comparing a behavioral intervention with medical therapy in the prevention of diabetes.[1] Over a mean follow-up period of 2.8 years, the lifestyle-modification program, known as the Diabetes Prevention Program (DPP), reduced the incidence of diabetes by 58% as compared with placebo among people with elevated fasting and post-load plasma glucose concentrations. Metformin reduced the incidence of diabetes by 31% as compared with placebo.

Despite these findings, insurers have been slow to provide coverage for DPP-like interventions. In 2016, the Centers for Medicare and Medicaid Services piloted the program and determined that it improved the quality of patient care and reduced net Medicare spending, prompting a goal of expanding the DPP nationwide by 2018. Although coverage of metformin has been ubiquitous since it was introduced in the United States in 1995, many private insurers started covering the DPP only recently.

Financial incentives for tobacco cessation during pregnancy provide another example of an effective behavioral intervention that hasn't been translated into practice. Smoking during pregnancy is a leading cause of maternal and neonatal morbidity and mortality, particularly among socially disadvantaged women and their children, and has long been a public health target. In the United States, such smoking rates have decreased only marginally in recent decades. A Cochrane review concluded that financial incentives are the most effective intervention in this population and can lead to quit rates up to four times higher than those achieved with other interventions. But such incentives haven't been implemented in routine care of pregnant women.

Why are highly effective preventive interventions adopted slowly, if at all? The first issue is that, historically, far more resources have been devoted to treating disease than to preventing it; in 2015, only 3% of health care dollars were spent on preventive services. However, ongoing shifts in health financing are creating incentives for providers to pay more attention to modifiable risks such as antenatal smoking. Hospitals participating in accountable care organizations, for example, save thousands of dollars for each neonatal intensive care unit stay they prevent.

Second, treatments determined by the Food and Drug Administration (FDA) to be safe and effective are usually covered by insurers regardless of their cost, but preventive services have been