SPECIAL COMMUNICATION

Clinical Quality Measures: A Challenge for–and to–Family Physicians

Alan Drabkin, MD, FAAFP, Ronald N. Adler, MD, FAAFP, Wayne Altman, MD, FAAFP, Alan M. Ehrlich, MD, FAAFP, Alicia Agnoli, MD, MPH, MHS, and Brian S. Alper, MD, MSPH, FAAFP, FAMIA

Introduction: **Improving design, selection and implementation of appropriate clinical quality measures can reduce harms and costs of health care and improve the quality and experience of care delivery. These measures have not been evaluated for appropriateness for use in performance measurement in a systematic, reproducible, and widely accepted manner.**

Methods: **We defined 10 criteria for evaluation of measure appropriateness in 4 domains: Patient-centeredness of outcomes, specification of population measured and measure detail, reliable evidence that benefits likely outweigh harms, and independence from significant confounders. We applied these criteria to 24 measures under consideration for statewide use in Massachusetts in public and private incentive-based programs. We appraised each measure as Appropriate or Not Appropriate for such use.**

Results: **We rated 15 measures as Appropriate (62.5%). Three measures (12.5%) were considered Appropriate only if applied at a system level but not for patient-provider assessment and 6 measures (25%) were rated Not Appropriate. Reasons for designation as "Not Appropriate" included benefits not clearly outweighing harms, lack of preservation of patient autonomy, inappropriate specification of population and measure detail, confounding by locus of control, and confounding by social determinants of health.**

Conclusions: **Using this consensus-driven, 10-criteria methodology we were able to evaluate appropriateness of clinical quality measures. This methodology may improve measure design and inform selection of the most appropriate measures for use in quality measurement, financial incentives, and reporting. (J Am Board Fam Med 2022;35:427–434.)**

Keywords: **Family Medicine, Massachusetts, Pay for Performance, Quality of Health Care**

Introduction

Family physicians are routinely evaluated using clinical quality measures. Whereas such measures may inform quality improvement (QI) activities, the stakes are higher when used for public reporting or in pay-for-performance (P4P) programs. Despite the considerable measure development effort over the past 2 decades, many quality measures remain flawed. There is no universally accepted standard for measure development, evaluation, or implementation, and there is very limited evidence that these measures lead to improved health outcomes.[1,2] Implementation of flawed measures—no matter how well-intended— may have harmful and unintended consequences, including inappropriate intensification of treatment to reach arbitrary targets and opportunity costs and waste associated with a focus on measured outcomes at the expense of more important goals. Troubling ethical dilemmas are created when

This article was externally peer reviewed.
Submitted 14 July 2021; revised 12 October 2021; accepted 14 October 2021.

Dr. Drabkin and Dr. Adler are lead coauthors and contributed equally to the writing of this manuscript.

From Tufts University School of Medicine (AD, WA); Harvard Medical School (AD); University of Massachusetts Medical School (RA, AME); University of California Davis School of Medicine (AA); and Computable Publishing, Ipswich, MA (BSA).

Funding: None.
Conflict of interest: Dr. Alan Ehrlich is a full-time employee at EBSCO, publishers of DynaMed. Dr. Brian Alper was a fulltime employee at EBSCO during manuscript drafting and is the owner of Computable Publishing, LLC.

Corresponding author: Alan Drabkin, MD, FAAFP, 33 Pond Ave, Apt 305, Brookline, MA 02445 (E-mail: drabkin.alan@gmail.com).

poorly designed measures pit the interests of doctors against those of patients.

Family physicians, burdened by clinical quality measures, often experience pressure to alter their care of patients to optimize performance on measures. Such pressures may be self-imposed or may come from employers, insurers, or as a response to public reporting. Family physicians respond to these demands in different ways. Some ignore these measures, often selectively, for example, prioritizing measures based on supporting evidence and the interests of their patients. Others may engage in "gaming" to optimize performance through various manipulations, including patient selection, adjusting diagnostic coding to include or exclude certain patients from a measure, altering close-to-target blood pressure (BP) readings, and attending to idiosyncratic timing, (eg, to ensure the A1c or BP result at target is the last 1 of the calendar year for reporting reasons).

Distinguishing appropriate from inappropriate quality measures requires criteria by which to make such judgments. Whereas flawed measures may be acceptable in certain settings (such as early stages of local QI efforts), when the stakes are high (such as in P4P programs or public reporting), the measures should satisfy more rigorous criteria. Clinical quality measures have not been evaluated for appropriateness for use in performance measurement in a systematic, transparent, reproducible, and widely accepted manner other than the American College of Physicians (ACP) review.[3] This described a systematic methodology for evaluating measure validity of 86 general medicine measures. 30 (35%) were judged Not Valid and 24 (28%) as Uncertain Validity. Endorsements of measures by the National Quality Forum are influential, but their evaluation process is not openly available and reproducible.

Methods

We convened a group of family physicians (with diversity in gender, age, community, and practice setting) to create a reproducible methodology for assessing their appropriateness for use in P4P programs.

By consensus, we developed a set of 10 criteria for measure appropriateness[4] (Table 1). For this pilot implementation, we classified these criteria into 4 domains: patient-centeredness, specification of outcome and population detail, evidence regarding benefits and harms, and independence from significant confounders.

Table 1. Criteria for Evaluation of Appropriateness of Clinical Quality Measures

Does it matter to patients?	1. Patient-oriented outcome: For an outcome measure, the outcome is important to patients (improves quality or quantity of life). For a process measure, the action is likely to lead to an outcome that is important to patients.
	2. Autonomy preserved (shared decision-making): Patient autonomy is preserved for decisions in which reasonable, informed patients may make different choices.
Is it appropriately specified?	3. Denominator specification: The population is clearly and adequately specified with appropriate exclusion criteria and assessment method.
	4. Numerator specification: The outcome being measured is clearly and adequately specified with appropriate timeframe and assessment method.
Is there sufficient evidence that benefits outweigh harms and costs?	5. Certainty of net benefit: There is sufficient evidence that the action(s) proposed by the quality measure generate desirable consequences that outweigh undesirable consequences.
	6. Measure implementation improves outcomes: There is sufficient evidence that actual implementation of the measure will lead to desirable consequences that outweigh undesirable consequences.
	7. Resource use: Measure implementation is likely to produce net benefits that justify the resources (human, material, and financial) expended on its implementation (care provision, measurement, and reporting).
Does the measure assess quality, independent of significant confounding factors?	8. Gaming resistance: Measure implementation is unlikely to motivate a significant number of healthcare providers to change their patient selection, clinical decision-making behavior, or reporting in ways that improve measure performance without improving health outcomes if the measure is implemented.
	9. Locus of control: The entity for whom the quality of care is being measured can have sufficient authority, influence, or capacity to affect performance on the quality measure.
	10. Social determinants of health: Social determinants of health of the population served do not unduly influence performance on the measure.

At the request of the Massachusetts Medical Society Committee on Quality, we assessed 24 measures under consideration for statewide use in public and private incentive-based programs by the Massachusetts Executive Office of Health and Human Services Quality Alignment Task Force.[5] We met 3 times for a total of 8 hours. We rated each measure as Appropriate or Not Appropriate through open dialog until reaching consensus.

Results

We rated 15 measures (62.5%) as Appropriate (Table 2). Three additional measures (12.5%), which required availability and coordination of care among systems or multiple providers, were considered Appropriate only if applied at a system level but not for patient provider assessment. We rated 6 measures (25%) as Not Appropriate. Reasons for designation as "Not Appropriate" included benefits not clearly outweighing harms, lack of preservation of patient autonomy, inappropriate specification of included population and/or measure detail, confounding by inappropriate locus of control, and confounding by social determinants of health (SDOH).

Four of the 6 measures rated Not Appropriate fail multiple criteria (Table 2).

Three measures fail to Preserve Patient Autonomy:

Two measures with specific BP targets do not provide opportunity for patients and clinicians to weigh the potential harms of additional BP lowering against a likely small benefit for BP that is already near target. The values and preferences of patients are not elicited or respected in implementing these measures.

Breast cancer screening involves highly personal decisions. There is evidence of potential benefits, but also significant potential harms (false positives, overdiagnosis, overtreatment) that vary widely in relative importance depending on patient values. Therefore, shared decision making (SDM) is most appropriate.[6] Inexplicably, this measure penalizes clinicians who engage in thoughtful collaboration with patients who then decline screening.

Two measures fail on Denominator Specification:

Two BP control measures did not exclude elderly patients (after 75 or 85 years), for whom intensive efforts to lower BP create significant risk of medication-related adverse events.

Three measures fail on Numerator Specification:

A drug dependence treatment measure requires initiation of treatment by a different clinician from that of the initial visit, or on a subsequent day from that visit. However, initiation of treatment by a primary care provider (PCP) on the same day can be clinically appropriate (even perhaps ideal).

Two outcome measures define depression remission as a PHQ-9 < 5. This is inappropriate because PHQ-9 scores of 5 to 9 are not specific for depression.[7] Factors such as fatigue and insomnia produce scores > 5 in the absence of clinical depression.

Three measures are not supported by evidence that Benefits Clearly Outweigh Harms:

Although there is evidence of net benefit for BP lowering in severe hypertension, this is uncertain for patients with mild hypertension and no cardiovascular disease.[8,9] The potential harms of medication-related adverse effects may outweigh the benefits of more intensive BP control, especially for older patients.[10]

No all-cause mortality benefit has been demonstrated for screening mammography, and the breast cancer-specific mortality benefit is extremely small (Number Needed to Screen = 1503 women aged 50 to 59 years).[11] Significant harms such as false positives and overdiagnosis are well-described and highly prevalent.[12] After engaging in SDM, many women reasonably conclude that the benefits of screening do not exceed the harms.

One measure fails to be within the clinician's Locus of Control:

A drug dependence measure that requires suitable follow up care defines a variety of visits as the index visit, including emergency department (ED) visits. Whether the ED arranges appropriate follow-up or referral is beyond a PCP's control. This measure would be appropriate only if applied at a system level, where there is control and influence over all visits and follow-up services.

This measure is also strongly influenced by SDOH, including the ability to afford certain types of care, access to care, access to transportation, and presence of an adequate social support network.

Discussion

Inappropriate quality measures cause harms and promote waste in health care. Among 24 measures evaluated by 10 criteria, we identified problems with 9 (38%). These results are similar to the ACP analysis, which deemed 35% of measures Not Valid. There are important differences in our methods and findings. The ACP developed a numeric scoring rubric, rating measures as Valid, Uncertain Validity, or Not

Table 2. Appropriateness Ratings of 24 Measures

Name	Description	Appropriateness Evaluation	Criteria Not Satisfied	ACP Review
Controlling High Blood Pressure	The percentage of members 18 to 85 years of age who had a diagnosis of hypertension (HTN) and whose BP was adequately controlled (<140/90 mm Hg) during the measurement year	Not Appropriate	2, 3, 5	Uncertain validity
Comprehensive Diabetes Care: Hemoglobin A1c (HbA1c) Poor Control (>9.0%)	Percentage of patients 18 to 75 years of age with diabetes who had hemoglobin A1c > 9.0% during the measurement period	Appropriate		Uncertain validity
CG-CAHPS (MHQP Version)	Composites: Getting Timely Appointments, Care, and Information; How Well Providers Communicate; Providers' Use of Information to Coordinate Patient Care, Helpful, Courteous, and Respectful Office Staff; Patient's Rating of Provider	Appropriate		Not rated
Depression Screening and Follow-Up for Adolescents and Adults	Percentage of members 12 years of age and older who were screened for clinical depression using a standardized tool and, if screened positive, who received follow-up care. • Depression Screening. The percentage of members who were screened for clinical depression using a standardized tool. • Follow-Up on Positive Screen. The percentage of members who screened positive for depression and received follow-up care within 30 days.	Appropriate		Uncertain validity
Depression Remission at Six or Twelve Months	Adult patients age 18 and older with major depression or dysthymia and an initial PHQ-9 score > 9 who demonstrate remission at six or twelve months defined as a PHQ-9 score less than 5. This measure applies to patients with both newly diagnosed and existing depression whose current PHQ-9 score indicates a need for treatment.	Not Appropriate	4	Not rated
Depression Remission and Response for Adolescents and Adults	Adult patients age 18 and older with major depression or dysthymia and an initial PHQ-9 score > 9 who demonstrate remission at six or twelve months defined as a PHQ-9 score less than 5.	Not Appropriate	4	Not rated

Continued

Table 2. Continued

Name	Description	Appropriateness Evaluation	Criteria Not Satisfied	ACP Review
Depression Response at Six or Twelve Months - Progress Toward Remission	Adult patients age 18 and older with major depression or dysthymia and an initial PHQ-9 score > 9 who demonstrate a response to treatment at six or twelve months defined as a PHQ-9 score that is reduced by 50% or greater from the initial PHQ-9 score.	Appropriate		Not rated
Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment	Percentage of adolescent and adult patients with a new episode of alcohol or other drug (AOD) dependence who received appropriate follow-up care: • Initiation of AOD Treatment. The percentage of patients who initiate treatment through an inpatient AOD admission, outpatient visit, intensive outpatient encounter or partial hospitalization within 14 days of the diagnosis. • Engagement of AOD Treatment. The percentage of patients who initiated treatment and who had two or more additional services with a diagnosis of AOD within 30 days of the initiation visit	Not appropriate	4,9	Not rated
Childhood Immunization Status (Combo 10)	Percentage of children that turned 2 years old during the measurement year and had specific vaccines by their second birthday	Appropriate		Not rated
Immunizations for Adolescents (Combo 2)	Percentage of adolescents that turned 13 years old during the measurement year and had specific vaccines by their 13th birthday	Appropriate		Not rated
Influenza Immunization	Percentage of patients aged 6 months and older seen for a visit between October 1 and March 31 who received an influenza immunization OR who reported previous receipt of an influenza immunization	Appropriate		Valid
Chlamydia Screening - Ages 16 to 24	Percentage of women ages 16 to 24 that were identified as sexually active and had at least one test for Chlamydia during the measurement year	Appropriate		Valid
Colorectal Cancer Screening	Percentage of adults 50 to 75 years of age who had appropriate screening for colorectal cancer	Appropriate		Valid

Continued

Table 2. Continued

Name	Description	Appropriateness Evaluation	Criteria Not Satisfied	ACP Review
Breast Cancer Screening	Percentage of women 50 to 74 years of age who had a mammogram to screen for breast cancer	Not Appropriate	2,5	Valid
Cervical Cancer Screening	Percentage of women 21 to 64 years of age, who received one or more Pap tests to screen for cervical cancer	Appropriate		Valid
Asthma Medication Ratio	Percentage of patients 5 to 64 years of age who were identified as having persistent asthma and had a ratio of controller medications to total asthma medications of 0.50 or greater during the measurement year	Appropriate		Not rated*
Comprehensive Diabetes Care: Eye Exam	Percentage of patients 18 to 75 years of age with diabetes who had a retinal or dilated eye exam by an eye care professional during the measurement period or a negative retinal exam (no evidence of retinopathy) in the 12 months before the measurement period	Appropriate		Not rated
Comprehensive Diabetes Care: Blood Pressure Control (<140/90 mm Hg)	Percentage of members 18 to 75 years of age with diabetes (type 1 and type 2) whose most recent blood pressure (BP) reading is < 140/90 mm Hg during the measurement year	Not Appropriate	2,3,5	Uncertain validity
Child and Adolescent Major Depressive Disorder: Suicide Risk Assessment	Percentage of patient visits for those patients aged 6 through 17 years with a diagnosis of major depressive disorder with an assessment for suicide risk	Appropriate		Not rated
Follow-Up After Hospitalization for Mental Illness (30-Day)	Percentage of discharges for members 6 years of age and older who were hospitalized for treatment of selected mental health disorders and who had an OP visit, an intensive OP encounter, or partial hospitalization with a mental health practitioner. Two rates are reported (1) the percentage of members who received follow-up within 30 days of discharge, 2) the percent of members who received follow-up within 7 days of discharge	Appropriate at system level of application but Not Appropriate at individual practitioner level		Not rated
Follow-Up After Hospitalization for Mental Illness (7-Day)	Percentage of discharges for members 6 years of age and older who were hospitalized for treatment of selected mental health disorders and who had an OP visit, an	Appropriate at system level of application but Not Appropriate at individual		Not rated

Continued

Table 2. Continued

Name	Description	Appropriateness Evaluation	Criteria Not Satisfied	ACP Review
	intensive OP encounter, or partial hospitalization with a mental health practitioner. Two rates are reported: 1) the percentage of members who received follow-up within 30 days of discharge, 2) the percent of members who received follow-up within 7 days of discharge	practitioner level		
Follow-up After Emergency Department Visit for Mental Health (7-Day)	The percentage of emergency department (ED) visits for members 6 years of age and older with a principal diagnosis of mental illness, who had a follow-up visit for mental illness. Two rates are reported:1. The percentage of ED visits for which the member received follow-up within 30 days of the ED visit (31 total days).2. The percentage of ED visits for which the member received follow-up within 7 days of the ED visit (8 total days).	Appropriate at system level of application but Not Appropriate at individual practitioner level		Not rated
Continuity of Pharmacotherapy for Opioid Use Disorder	Percentage of adults 18 to 64 years of age with pharmacotherapy for opioid use disorder (OUD) who have at least 180 days of continuous treatment	Appropriate		Not rated
Use of Imaging Studies for Low Back Pain	Percentage of patients 18 to 50 years of age with a diagnosis of low back pain who did not have an imaging study (plain Radiograph, MRI, CT scan) within 28 days of the diagnosis	Appropriate		Valid

*Merit-based Incentive Payment System (MIPS) measure 444 is an alternative, and is rated Valid by the ACP

Criteria not satisfied (from Table 1)

2. Autonomy not preserved
3. Denominator not appropriately specified
4. Numerator not appropriately specified
5. Benefits do not clearly outweigh harms
9. Confounders such as Locus of Control

Abbreviations: PHQ, patient health questionnaire; OUD, opioid use disorder; AOD, alcohol or other drug dependence; ACP, american college of physicians.

Valid. In contrast to the ACP scoring rubric, our criteria were developed and applied by group consensus regarding key elements of appropriateness, without using numeric cutoffs. Distinct from the ACP, our criteria included or emphasized the elements of preservation of patient autonomy, assessment of certainty of net benefit, evaluation of resistance to gaming, and limiting potential confounding by SDOH.

The qualitative aspect of our rating process may be considered a limitation but also allows flexibility in implementation to meet local priorities. Quantitative approaches may be developed in future iterations of this approach, but care must be taken to determinate

that quantification does not result in false precision or inconsistent results. We are confident that our criteria would allow other representative groups of stakeholders to reach similar conclusions to us, but demonstration of external validity is beyond the scope of this first pilot.

The ACP did not rate 12 of the 24 measures we evaluated (primarily measures relevant to children or mental health). Among the 12 measures rated by both groups, 3 measures were rated differently. We judged the breast cancer screening measure "Not Appropriate" (due to absence of certainty of net benefit and failure to preserve patient autonomy), whereas the ACP deemed this measure Valid (and does not include patient autonomy as a key criteria). 2 other measures were rated differently because the ACP includes a third rating category of rating (Uncertain Validity) whereas we do not. Results are summarized in Table 2.

Conclusion

Clinical quality measures influence behavior, especially when tied to P4P, but they may induce harms and waste through unintended consequences, especially when poorly designed or implemented. Identifying flawed clinical quality measures using specific criteria can illuminate the nature of their flaws and facilitate replacement or improvement.

Inappropriate quality measures should be retired or improved. More meaningful measures (eg, the Person-Centered Primary Care Measure[13]) should be developed to promote improved quality and experience of care for patients and clinicians. Family physicians are ideally positioned to influence decisions regarding selection and prioritization of performance measures, which often occur at local and regional levels. This would promote aligninment of allocation of effort and resources to achieve outcomes that truly matter to patients.

We thank the many students from Tufts University School of Medicine and Harvard Medical School who participated in previous analyses using earlier versions of the criteria set.

To see this article online, please go to: http://jabfm.org/content/35/2/427.full.

References

1. Saver BG, Martin SA, Adler RN, et al. Care that Matters: Quality Measurement and Health Care. PLoS Med 2015;12:e1001902. Nov 17.
2. Rathi VK, McWilliams JM. First-Year Report Cards From the Merit-Based Incentive Payment System (MIPS): What Will Be Learned and What Next? JAMA 2019;321:1157–8.
3. MacLean CH, Kerr EA, Qaseem A. Time Out - ``Charting a Path for Improving Performance Measurement. N Engl J Med 2018 May 10;378:1757–76.
4. Adler R, Hamdan S, Scanlon C, Altman W. Quality Measures: How to Get Them Right. Fam Pract Manag 2018;25:23–8.Jul-Aug.
5. Massachusetts Aligned Measure Set for Global Budget-Based Risk Contracts 2020 Measures and Implementation Parameters, May 3, 2019. Massachusetts Executive Office of Health and Human Services Quality Alignment Taskforce. Available at https://www.mass.gov/doc/2020-measures-and-implementation-parameters-updated-as-of-050319-0/download. Accessed October 31, 2020.
6. Keating NL, Pace LE. Breast Cancer Screening in 2018: Time for Shared Decision Making. JAMA 2018;319:1814–5.May 1.
7. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16:606–13.
8. Diao D, Wright JM, Cundiff DK, Gueyffier F. Pharmacotherapy for mild hypertension. Cochrane Database Syst Rev 2012;15:CD006742. Aug.
9. Martin SA, Boucher M, Wright JM, Saini V. Mild hypertension in people at low risk. BMJ 2014;349:g5432.
10. Tinetti ME, Han L, Lee DS, et al. Antihypertensive medications and serious fall injuries in a nationally representative sample of older adults. JAMA Intern Med 2014;174:588–95.
11. Myers ER, Moorman P, Gierisch JM, et al. Benefits and Harms of Breast Cancer Screening: A Systematic Review. JAMA 2015;314:1615–34.
12. Nelson HD, Pappas MA, Cantor A, Griffin MS, Daeges M, Humphrey L. Harms of Breast Cancer Screening: Systematic Review to Update the 2009 U. S. Preventive Services Task Force Recommendation. Ann Intern Med 2016;164:256–67.
13. Etz RS, Zyzanski SJ, Gonzalez MM, Reves SR, O'Neal JP, Stange KC. A new comprehensive measure of high-value aspects of primary care. Ann Fam Med 2019;17:221–30.