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, physicians and some policymakers who believe that current measures are not meaningful. In a recent survey, 63% of physicians said that current measures do not capture the quality of the care that physicians provide. Yet U.S. physician practices are spending $15.4 billion each year — about $40,000 per physician — to report on performance.

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. 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.

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...
(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

<table>
<thead>
<tr>
<th>Measure</th>
<th>NQF-Endorsed Steward Measure</th>
<th>Importance</th>
<th>Appropriateness</th>
<th>Clinical Evidence</th>
<th>Specifications</th>
<th>Feasibility Rationale</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avoidance of Antibiotic Treatment in Adults with Acute Bronchitis</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Based on appropriate exclusion criteria for patients with COPD or immunocompromised patients. Clinically important process with known performance-appropriate adverse events by excluding patients receiving inappropriate antibiotic therapy.</td>
<td></td>
</tr>
<tr>
<td>Chronic Stable Coronary Artery Disease: Antiplatelet Therapy: MIPS 006 (NQF 0067)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Clinically important process with known performance-appropriate adverse events by excluding patients receiving inappropriate antiplatelet therapy.</td>
<td></td>
</tr>
<tr>
<td>Stroke and Stroke Rehabilitation: Discharged on Anticoagulant Therapy: MIPS 02</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>Indicates a time frame that contradicts evidence-based clinical recommendations. Does not consider patient preferences for switching to alternative treatments.</td>
<td></td>
</tr>
<tr>
<td>Anti-depressant Medication Management: MIPS 009 (NQF 0105)</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>±</td>
<td>Indicates a time frame that contradicts evidence-based clinical recommendations. Does not consider patient preferences for switching to alternative treatments.</td>
<td></td>
</tr>
<tr>
<td>Pain Assessment and Follow-Up: MIPS 032 (NQF 0420)</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>+</td>
<td>Insufficient evidence that screening reduces harm to warrant routine screening. Target population too broad.</td>
<td></td>
</tr>
<tr>
<td>Elder Maltreatment Screen and Follow-Up: MIPS 131 (NQF 0420)</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>−</td>
<td>Insufficient evidence that screening reduces harm to warrant routine screening. Target population too broad.</td>
<td></td>
</tr>
<tr>
<td>Controlling High Blood Pressure: MIPS 236 (NQF 0018)</td>
<td>+</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>Adult Sinusitis: Antibiotic Prescribed for Acute Sinusitis: MIPS 052 (NQF 0102)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Measure should exclude patients who have severe or worsening symptoms within 10 days of onset and who would benefit from earlier, appropriate antibiotic treatment.</td>
<td></td>
</tr>
</tbody>
</table>

* Measures were rated on a 9-point scale according to whether they met criteria: higher scores were better. A plus sign indicates that the measure meets criteria (rating was 8 or 9); a minus sign indicates that the measure meets some criteria (rating was 4 or 5). AAN denotes American Academy of Neurology, AAO-HNS denotes American Academy of Otolaryngology-Head and Neck Surgery, ATS denotes American Thoracic Society, CMS denotes Center for Medicare and Medicaid Services, MIPS denotes Medicare-Based Incentive Payment System, NCQA denotes National Committee for Quality Assurance, NQF denotes National Quality Forum, and QPP denotes Quality Payment Program.

---

The New England Journal of Medicine

Downloaded from nejm.org on August 9, 2018. For personal use only. No other uses without permission.

Copyright © 2018 Massachusetts Medical Society. All rights reserved.
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.

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 inter-

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 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 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 appropriateness method does not classify measures as valid when there are significant disagreements among the panelists.

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
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. 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