AMERICAN THORACIC SOCIETY DOCUMENTS

An Official American Thoracic Society Workshop Report: Developing Performance Measures from Clinical Practice Guidelines

Jeremy M. Kahn, Michael K. Gould, Jerry A. Krishnan, Kevin C. Wilson, David H. Au, Colin R. Cooke, Ivor S. Douglas, Laura C. Feemster, Richard A. Mularski, Christopher G. Slatore, and Renda Soylemez Wiener; on behalf of the ATS *Ad Hoc* Committee on the Development of Performance Measures from ATS Guidelines

This Official Workshop Report was Approved by the American Thoracic Society Board of Directors, December 2013.

Abstract

Many health care performance measures are either not based on high-quality clinical evidence or not tightly linked to patient-centered outcomes, limiting their usefulness in quality improvement. In this report we summarize the proceedings of an American Thoracic Society workshop convened to address this problem by reviewing current approaches to performance measure development and creating a framework for developing high-quality performance measures by basing them directly on recommendations from well-constructed clinical practice guidelines. Workshop participants concluded that ideally performance measures addressing care processes should be linked to clinical practice guidelines that explicitly rate the quality of evidence and the strength of recommendations, such as the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) process. Under this framework, process-based performance measures would only be developed from strong recommendations based on high- or moderate-quality evidence. This approach would help ensure that clinical processes specified in performance measures are both of clear benefit to patients and supported by strong evidence. Although this approach may result in fewer performance measures, it would substantially increase the likelihood that quality-improvement programs based on these measures actually improve patient care.

Correspondence and requests for reprints should be addressed to Jeremy M. Kahn, M.D., M.S., Associate Professor of Critical Care and Health Policy and Management, University of Pittsburgh, 602-B Scaife Hall, 3550 Terrace Street, Pittsburgh, Pennsylvania 15261. E-mail: kahnjm@upmc.edu

Ann Am Thorac Soc Vol 11, No 4, pp S186–S195, May 2014

DOI: 10.1513/AnnalsATS.201403-106ST
Internet address: www.atsjournals.org

CONTENTS

Overview

Performance measurement is an increasingly important topic in health care delivery as patients, governments, and health care payers seek more accountability in the health care system through pay-for-performance and other quality-improvement initiatives. In May, 2013 the American Thoracic Society (ATS) convened a workshop on the topic of performance measure development in health care. The purpose of this workshop was to review current approaches to performance measure development and create a framework for developing performance measures from clinical practice guidelines. Workshop participants included experts in pulmonary, critical care, and sleep medicine; guideline development; behavioral science; health services research; quality measurement; performance improvement; health economics; and health policy and management. Participants reviewed current approaches to performance measure development and created a framework for developing process-based performance measures from clinical practice guidelines. This workshop report was developed through an iterative review process, with input from key ATS assemblies and committees. The conclusions of the workshop were as follows:

- Many process-based health care performance measures are either not based on high-quality clinical evidence or not tightly linked to patient-centered outcomes, limiting their usefulness in quality improvement.
- The Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) guideline development process, adopted by the ATS in 2006, is a potentially useful

framework on which to base performance measures that overcome these problems.
- The GRADE process involves explicitly, and separately, rating the quality of evidence and strength of the recommendation based on multiple factors, including patient values and preferences.
- Process-based performance measures should ideally be based solely on strong recommendations based on high- or moderate-quality evidence, increasing the chance that they lead to improvements in outcomes.
- Process-based performance measure development should also proceed alongside guideline development, helping to ensure that guideline developers make specific, actionable recommendations that lend themselves to performance measurement.
- Future research is necessary to determine if this framework results in higher-quality performance measures than currently exist.

Introduction

Performance improvement is an increasingly important component of modern health care (1). Yet, to be successful, performance improvement initiatives must rely on high-quality performance measures that are based on strong clinical evidence and tightly linked to outcomes that are important to patients or society (2). This is rarely the case in the field of pulmonary, critical care, and sleep medicine, in which existing quality measures are based on varying levels of practice recommendations and evidence (3, 4). Consequently, quality improvement initiatives and guideline implementation projects may fail to resonate with clinicians and may not result in improved outcomes (5, 6).

Recently, the American Thoracic Society (ATS) adopted the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach to generating clinical practice guidelines (7). GRADE represents an advance over other evidence grading systems in that it provides a systematic and transparent framework for clarifying clinical questions, determining outcomes of interest, appraising and summarizing the evidence, and moving from evidence to recommendations (8). Additionally, by separating the strength of the recommendation from the quality of the evidence, GRADE allows guideline developers to easily incorporate the values and preferences of patients and other stakeholders into each recommendation (9). Because of these attributes, GRADE-based clinical practice guidelines may be a particularly useful foundation for developing performance measures that are both based on high-quality evidence and tied to patient-centered outcomes (i.e., outcomes that patients notice and care about, such as survival, function, symptoms, and health-related quality of life) (10).

To evaluate these issues, we convened a workshop in which we reviewed current approaches to performance measurement and examined ways in which the principles of GRADE may be applied to performance measure development. We specifically focused on process-based performance measures, rather than outcome-based measures, because process-based measures are directly related to treatment and diagnostic strategies that form the basis of clinical practice guidelines. In this report, we present the results of this workshop and put forth a framework for the development of performance measures based on GRADE-formulated clinical practice guidelines.

Methods

We convened a work group composed of experts in pulmonary, critical care, and sleep medicine; guideline development; behavioral science; health services research; quality measurement; performance improvement; health economics; and health policy and management (*see* list at end of document). In assembling the work group we sought input and representation from relevant ATS committees and assemblies, including the Quality Improvement Committee, the Documents Development and Implementation Committee, the Health Policy Committee, the Behavioral Science and Health Services Research Assembly, the Critical Care Assembly, the Nursing Assembly, and the Pediatrics Assembly. Work group participants received a set of introductory materials including background on the GRADE methodology (1, 7), a recent ATS-endorsed GRADE-based guideline (2, 11), and information related to performance measure development (3, 4, 12).

Work group members met at a 1-day workshop held in San Francisco, California on May 19, 2012. The workshop consisted of presentations by content experts followed by break-out sessions focused on prioritizing guideline recommendations for performance measurement and developing performance measures based on GRADE methodology. Each didactic presentation and break-out session was followed by group discussion with a goal of defining key areas of consensus and disagreement. After the workshop, a writing committee drafted a workshop report based on an outline developed by the project co-chairs. The draft was circulated to the members of the working group and to the leadership of each of the sponsoring ATS committees and Assemblies, with revisions at each step and consensus achieved through moderated discussion. The final document was approved by the ATS Board of Directors as an Official ATS Workshop Report December 2013.

The intended audience of this report is performance measure developers and guideline developers as well as policy makers, clinicians, and researchers with an interest in performance measurement and quality improvement.

Results

Overview of Performance Measures

Performance measures are the specific, quantifiable, representation of a capacity, process, or outcome relevant to the assessment of health care quality (5, 6, 13). In general, performance measures address one of three major quality domains: the structure of care, the process of care, or the outcome of care (7, 14). Process-based and outcome-based measures are most commonly in use, and each type of measure has strengths and weaknesses. Process-based measures are generally more actionable, whereas outcome-based measures, although clearly more important to patients, do not provide specific courses for improvement. When performance measures stem from clinical practice guidelines they typically assess the process of care, because the underlying goal of guidelines is to inform practitioners and policy makers about evidence-based care

Table 1. The Agency for Healthcare Research and Quality's list of desirable attributes of a performance measure

Domain	Attribute	Description
Importance	Relevance to stakeholders	The topic area is of significant interest and is financially and strategically important to stakeholders (e.g., patients, clinicians, purchasers, public health officers, policy makers).
	Health importance	The measure addresses an important aspect of health, as defined by high prevalence or incidence and/or a significant effect on the burden of illness (i.e., effect on the mortality and morbidity of a population).
	Applicability to the equitable distribution of health care	The measure can be stratified or analyzed by subgroup to test for disparities in health or health care among a diverse population of patients.
	Potential for improvement	There is evidence of a need for the measure, either because there is poor quality overall or there are variations in quality among organizations or populations
	Susceptibility for influence by the health care system	For health care delivery measures, the results of the measure relate to the actions of the providers whose performance is being measured, so that it is possible for them to improve that performance. For public health measures, the results should be susceptible to influence by the public health system.
Scientific soundness	Explicitness of evidence	The evidence supporting the measure is explicitly stated.
	Strength of evidence	The topic area is strongly supported by the evidence (i.e., indicated to be of great importance for improving quality of care [for health care delivery measures] or improving health [for population health measures]).
	Reliability	The results of the measure are reproducible for a fixed set of conditions irrespective of who makes the measurement or when it is made.
	Validity	The measure truly measures what it purports to measure.
	Allowance for patient/ consumer factors	The measure allows for stratification or case-mix adjustment if appropriate.
	Comprehensibility	The results of the measure are understandable by the individuals who will be acting on the data.
Feasibility	Explicit specification of numerator and denominator	The measure should have explicit and detailed specifications for the numerator and denominator; statements regarding data collection requirements are understandable and implementable.
	Data availability	The necessary data sources are available and accessible within the timeframe for measurement. The costs of abstracting and collecting data are justified by the potential for improvement in care or health.

Adapted from Reference 16.

processes that directly lead to desired health outcomes (8, 15).

In developing process-based performance measures, issues to consider include the purpose of performance measurement (e.g., public reporting, local improvement efforts, regulatory requirements), the importance of the clinical problem, the ease of measurement, the variability in delivered care, the relative opportunity for improvement, and the potential for unintended consequences (9, 12). Ultimately, successful performance measures (i.e., measures that lead to improved outcomes when used in practice) are important, scientifically sound, and feasible (Table 1) (10, 16). Just because an aspect of care can be measured does not mean it should be measured, and, conversely, just because something is deemed important does not mean it can be measured well.

Once a care process is selected for performance measurement, there are several specific steps to generate a performance measure (12). These include writing an indicator statement, specifying eligibility and the desired action in plain language, defining the denominator (i.e., who is eligible for the measure), and defining the numerator (i.e., who received the care process). Additionally, writers must specify how patient preferences will be handled—options include removing refusals from the numerator, removing refusals from both the numerator and denominator, or adding refusals to the numerator and denominator as a marker of good care that indicator was offered but refused. Compared with the

option of removing refusals from the both the numerator and denominator, removing refusals from the numerator only would decrease performance, as might be desirable if refusals are due to poor health literacy rather than informed patient choices. For example, refusal to undergo resection of lung cancer is sometimes related to misperceptions about risks, such that refusals may indicate poor provider efforts at education (17). Conversely, including refusals in both the numerator and denominator would increase performance, as might be desirable when refusals are due strictly to informed patient choices. For example, palliative care consultations might be a legitimate quality measure, but given the preference sensitive nature of this intervention, the provider should not be harmed if the patient refused (18). Importantly, accurately documenting refusals and other exceptions remains a challenge but may be facilitated by advances in the electronic health record.

The last steps in measure development determine how the measure will be scored (e.g., as a proportion, a rate, or a standardized score), estimate the sample size necessary to identify clinically and statistically significant differences in performance, and test the reliability and validity of the measure in multiple diverse settings where it is to be deployed. The initial development should be part of a larger, iterative process wherein the measures are vetted by key stakeholders, the impact of their implementation assessed, and the measure ultimately refined based on these results (19).

Limitations of Existing Process-based Performance Measures

Process-based performance measures quantify whether specific patient populations received recommended actions to treat, diagnose, or prevent disease. Therefore, they connote certainty that the desirable consequences of the recommended course of action outweigh the undesirable consequences (7). However, existing performance measures often fall short of this expectation, advocating for interventions simply because they make intuitive sense or are easily measured. This can lead to performance measures for care practices that lack proven benefit, result in only small benefits that do not outweigh the resources required for measurement, or are based on insufficient quality evidence to make a judgment. These three limitations were highlighted in a recent analysis of Centers for Medicare and Medicaid Services/Joint Commission performance measures for patients hospitalized with community-acquired pneumonia (3).

Lack of proven benefit. Interventions assessed by existing performance measures may lack a proven benefit. As an example, consider inpatient smoking cessation counseling. A metaanalysis of 10 randomized trials found that inpatient smoking cessation counseling did not improve the smoking quit rate (15.9 vs. 15.6%; relative risk, 1.05; 95% confidence interval, 0.90–1.22) (3). Generally speaking, interventions without proven benefit are inappropriately assessed by performance measures because their measurement and implementation consume scarce resources, burden caregivers and patients, and may place patients at risk for an adverse event, without meaningful chance of benefit.

Lack of a sufficiently large benefit. Existing performance measures sometimes assess interventions that confer only small benefits to patients, which may not outweigh the costs, burdens, and risks of the intervention. As an example, consider pneumococcal vaccination. A metaanalysis of 10 randomized trials found that pneumococcal vaccination reduced the incidence of pneumococcal pneumonia only 0.5% (from 0.9 to 0.4%; meaning that 100–250 patients would need to be vaccinated to prevent one case of pneumococcal pneumonia) and reduced the incidence of invasive pneumococcal disease only 0.1% (from 0.2 to 0.1%; meaning that 500–1,000 patients would need to be vaccinated to prevent one case of invasive pneumococcal pneumonia) (3). Although these risk reductions were statistically significant, it is conceivable that health care organizations may deem such benefits to be insufficient to warrant the opportunity costs of implementation of a pneumococcal vaccination program. Patients may also not support performance measures that assess interventions with effect sizes so small that the intervention has little chance of helping them directly.

Low-quality evidence. Interventions assessed by existing performance measures are often supported by only low- or very low-quality evidence. As an example, consider blood cultures drawn in the emergency department. Only a single retrospective cohort study found that patients with community-acquired pneumonia who had blood cultures drawn in the emergency department were more likely to achieve clinical stability within 48 hours (20). A similar study found no difference in the length of stay (21), and three other observational studies found a non–statistically significant reduction in mortality (22). Such evidence is low quality and, therefore, provides little confidence that the direction and magnitude of the estimated effects are correct. When uncertainty or conflicting evidence exists, there should be other strong compelling factors that favor usefulness as a performance measure, such as strong preferences on the part of patients. At the same time, this example demonstrates that face validity alone is insufficient rationale for a performance measure.

Review of GRADE

Given the importance of the quality of evidence and clarity of benefit in performance measurement, formal systems of grading evidence may play a key role in developing performance measures. Several different systems have been developed to evaluate the quality of scientific evidence regarding the potential benefits and harms of medical interventions. GRADE provides a consistent framework for making subjective judgments about specific domains of quality (8). GRADE is explicit, systematic, and transparent, such that even when individuals disagree about quality ratings or the strength of a recommendation, they can see for themselves where the source of disagreement lies (23). For these reasons, the GRADE system was adopted by the ATS in 2006 (3).

Importantly, GRADE evaluates an entire body of evidence as it relates to a specific research question or recommendation, not just individual studies. In contrast to most other systems, GRADE rates quality of evidence outcome by outcome, including those outcomes that are considered to be important to patients. GRADE also separates the rating of the quality of evidence from the grading of the strength of the recommendation. The key question for both is the extent to which raters are certain about the conclusions. High-quality evidence is characterized by certainty about the magnitudes of benefits

and harms. Likewise, strong recommendations are characterized by certainty that the net benefits outweigh the net harms.

Under GRADE, the quality of evidence is influenced by several factors, including study design, risk of bias, consistency, precision, directness, and publication bias (Table 2). Evidence from randomized trials is presumed to be high in quality, unless compromised by one or more of these factors. If one or more of these factors is judged to be serious, the quality of evidence for that outcome is "rated down" accordingly. Evidence from observational studies, including cohort studies and case-control studies, is considered to be low in quality by default but can be "rated up" for demonstrating a large magnitude of effect, a dose–response, or results that would be even more compelling if plausible confounders are taken into account. This approach allows guideline developers to systematically and reliably rate the quality of the evidence, although it does rely on accurate and complete reporting of study methodology, which is not always available (24).

Once the quality of the body of evidence has been determined for each outcome, the GRADE approach uses several factors to determine the strength of the recommendation: quality of evidence, magnitude of benefits and harms, expected variation in patient preferences for the different outcomes, and resource use or cost (Table 3). Strong recommendations can be made when there is moderate- or high-quality evidence indicating that the net benefits of the intervention clearly outweigh the net harms, although variation in patient preferences and implications for resource use should also be considered. Strong recommendations can also be made even if the level of evidence is poor but values and preferences strongly support the action.

Strong recommendations mean that there is certainty that the desirable consequences of the intervention exceed the undesirable consequences, that further research is unlikely to change the level of certainty, and that nearly all patients would desire the intervention when given the choice. GRADE therefore sets a high bar for making strong recommendations. When using GRADE, weak recommendations are common, reflecting residual uncertainty about the quality of evidence, the ratio of benefits to harms, and/or variation in patient preferences. Weak recommendations, though often disappointing to guideline developers, are nonetheless important, because they explicitly acknowledge uncertainty while still providing evidence-based guidance to clinical practice. As developers refrain from making strong recommendations that are not warranted, they advance the field by highlighting gaps in evidence and enhance patient care by recognizing that some variation in practices may be justified.

A Framework for Performance Measure Development Based on GRADE

Based on this background, the workshop participants developed several guiding principles for creating performance measures from GRADE-based clinical practice guidelines.

First, process-based performance measures should be developed from only strong recommendations. Process-based performance measures indicate an imperative to implement the described process; only strong recommendations signal the level of certainty necessary to form such an imperative. Weak recommendations, although useful for informing clinical practice, indicate uncertainty that the benefits outweigh the harms and may not justify the costs. Weakly recommended treatments benefit some patients, but not all, and may even harm a substantial minority. Weak recommendations also signify areas that require more scientific evidence to achieve the level of certainty required for a performance measure. For these reasons, weak recommendations, despite their potential face validity, should not be the focus of performance measures.

Second, process-based performance measures should be developed from only recommendations based on high- or moderate-quality evidence. Although GRADE permits strong recommendations based on low- or very low-quality evidence

Table 2. Grading of Recommendations Assessment, Development, and Evaluation approach for assessing the quality of a body of evidence

Domain	Factor	Comment
Factors to consider when rating down the quality of evidence from randomized controlled trials	Risk of bias	Unclear allocation concealment, incomplete blinding, selective reporting of results, incomplete accounting of outcomes
	Heterogeneity	Inconsistent results across studies
	Imprecision	Confidence intervals range from substantial benefit to little or no benefit
	Indirectness	The best available evidence is taken from studies of a population, intervention, comparator, or outcome other than that specified in the question of interest
	Publication bias	Studies showing negative or conflicting results may exist but are missing from the literature
Factors to consider when rating up the quality of evidence from observational studies	Magnitude of effect	The size of benefit is particularly strong
	Dose–response	The benefit is strongest in areas where it is suspected to be strongest, as in when the level of exposure is high
	Plausible confounding effect	Beneficial effect of intervention even though intervention group sicker at baseline

Table 3. Grading of Recommendations Assessment, Development, and Evaluation approach for determining the strength of the recommendation

Factor	Comment
Balance between desirable and undesirable effects	The larger the difference between the desirable and undesirable effects, the higher the likelihood that a strong recommendation is warranted. The narrower the difference, the higher likelihood that a weak recommendation is warranted
Quality of evidence	The higher the quality of evidence, the higher the likelihood that a strong recommendation is warranted
Values and preferences	The more values and preferences vary, or the greater the uncertainty in values and preferences, the higher the likelihood that a weak recommendation is warranted
Costs (resource allocations)	The higher the costs of an intervention—that is, the greater the resources consumed—the lower the likelihood that a strong recommendation is warranted.

in exceptional situations, such recommendations are likely to require modification as the estimated benefits and harms change with further research. Additionally, strong recommendations based on low- or very low-quality evidence typically reflect existing common clinical practices and therefore may already be widely implemented. For example, recent guidelines on the care of patients with idiopathic pulmonary fibrosis (IPF) make a strong recommendation based on low-quality evidence that patients with IPF and resting hypoxemia receive long-term oxygen therapy (25). It is unlikely that any patient with IPF and hypoxemia does not already receive long-term oxygen, limiting the usefulness of a performance measure on this topic. Last, strong recommendations based on low- or very low-quality evidence are likely to greatly depend on the values and preferences of individual patients, thus making them poor performance measures. For example, the IPF guidelines make a strong recommendation based on low-quality evidence that appropriate patients with IPF undergo lung transplantation. This recommendation would make a poor performance measure because of the critical role of patient preferences and social support in the decision to undergo lung transplant.

Third, guideline recommendation statements should be written taking performance measurement into account. To be useful from a performance measurement standpoint, guideline statements should ideally include clear recommendations for action in measurable populations (26). Actionable recommendation statements are those that clearly define the population and the recommended action in explicit terms. Without such recommendation statements it will be extremely difficult to apply guidelines to performance improvement in any measurable way.

To highlight how this framework might work in action, the work group applied the early steps of drafting a performance measure to a recent GRADE-based guidelines for the management of patients with stable chronic obstructive pulmonary disease (COPD) (27), monitoring for delirium in critically ill adults (28), and the evaluation and treatment of sleepiness in noncommercial drivers (29). These guidelines were chosen because together they represent the three pillars of the ATS: pulmonary, critical care, and sleep medicine. A general framework for this process is shown in Figure 1, and examples from each guideline are shown in Table 4. Several key elements of this process are noteworthy. Most importantly, the process shows how the indicator statement is the key intermediate step when moving from clinical practice guideline recommendation to a performance measure. Furthermore, it shows how a GRADE recommendation statement can be deconstructed into an indicator statement that includes a numerator and a denominator as well as any important exclusion criteria.

These examples also highlight the practical aspects of performance measure development, including the role that potential data sources play in developing measures. For example, the indicator statement from the COPD guideline describes the "proportion of patients 40 years and older with a diagnosis of COPD experiencing an exacerbation of COPD in the past 6 months prescribed a long-acting anticholinergic or long-acting inhaled β-agonist." Notably, the guideline recommends the use of long-acting inhaled bronchodilators for the treatment of all symptomatic patients with COPD and $FEV_1 < 60\%$ predicted (11). However, identifying "symptomatic" patients and obtaining pulmonary function test data would require abstraction of the medical record, which is not feasible on a large scale. Limiting the denominator statement to only those patients with a recent

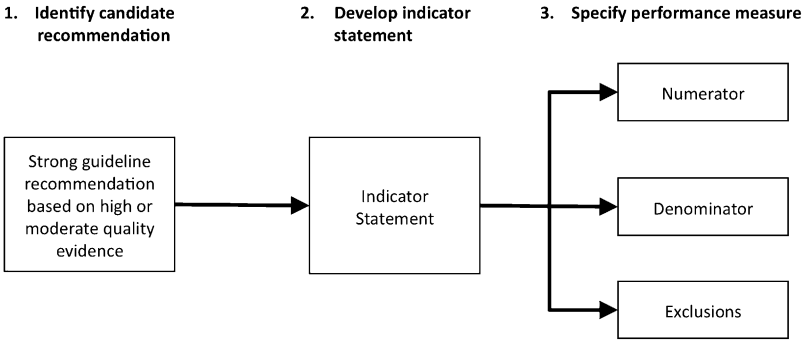

Figure 1. An illustration of the process by which a strong recommendation based on high- or moderate-quality evidence from a Grading of Recommendations Assessment, Development, and Evaluation–developed clinical practice guideline can be used to write an indicator statement for valid and useful performance measure.

Table 4. Examples of how strong recommendations based on moderate- or high-quality evidence can inform a performance measure's numerator, denominator, and exclusion criteria

Clinical Domain	Step 1. Identify Candidate Recommendation	Step 2. Develop Indicator Statement	Step 3. Specify Performance Measure
Pulmonary	We "recommend that clinicians prescribe monotherapy using either long-acting inhaled anticholinergics or long-acting inhaled β-agonists (LABA) for symptomatic patients with COPD and $FEV_1 < 60\%$ predicted"; strong recommendation, moderate-quality evidence (11)	Proportion of patients aged 40 years and older with a diagnosis of COPD experiencing an exacerbation of COPD in the past 6 months prescribed a long-acting inhaled anticholinergic or LABA	Numerator Number of patients filling a prescription for a long-acting inhaled anticholinergic or LABA Denominator Number of patients aged 40 years and older with diagnosis of COPD experiencing an exacerbation of COPD in the past 6 months Exclusions Patients with comorbid diagnosis of asthma Patients with documented adverse reaction to LABA and/or inhaled anticholinergics
Critical care	"We recommend routine monitoring for delirium in ICU patients" at moderate to high risk; strong recommendation, moderate-quality evidence (28)	Proportion of moderate- to high-risk ICU patient-days in which the patient was screened for delirium using a validated screening tool.	Numerator Patient-days in which the patient was screened for delirium using either the Confusion Assessment Method of the ICU or the Intensive Care Delirium Screening Checklist Denominator ICU patient-days in which the patient has either a baseline history of alcoholism, cognitive impairment, or hypertension; has severe sepsis or septic shock; is on mechanical ventilation; or is receiving parenteral sedative or opioid medications. Exclusions Patients receiving neuromuscular blockers Patients receiving comfort measures only
Sleep	"In patients with confirmed OSA who have been deemed high-risk drivers…we recommend CPAP therapy to reduce driving risk, rather than no treatment"; strong recommendation, moderate-quality evidence (29)	Proportion of adult patients with polysomnography-confirmed OSA, daytime sleepiness, and a recent history of an adverse driving incident.	Numerator Number of patients who are prescribed CPAP therapy Denominator Number of patients aged 18 years or older with polysomnography-confirmed OSA, daytime sleepiness defined as an Epworth Sleepiness Scale measurement > 10, and a documented adverse incident related to driving. Exclusions Patients with a contraindication to CPAP therapy Patients who do not drive or who agree to refrain from driving

Definition of abbreviations: COPD = chronic obstructive pulmonary disease; CPAP = continuous positive airway pressure; ICU = intensive care unit; LABA = long-acting inhaled β-agonists; OSA = obstructive sleep apnea.
Examples are provided in three clinical domains: pulmonary, critical care, and sleep medicine

exacerbation of COPD and who are aged 40 years or older results in a performance measure that can be implemented efficiently by using administrative data, albeit with tradeoffs surrounding sensitivity and specificity. As the goal is to capture patients who are most likely to benefit from the addition of long-acting inhaled therapies for benchmarking purposes, high specificity is desirable and preferable over a more sensitive, but less specific, measure. However, not all patients aged 40 years or older with a recent COPD exacerbation will have pulmonary function test data that meet the guideline recommended criteria for COPD diagnosis, demonstrating that just because a definition seems more specific does not mean that this actually is the case (30).

Benefits of this Approach

The framework presented here should help ensure that performance measures, and the programs that use them, lead to improvements in patient-centered outcomes. This should reduce the likelihood that performance improvement programs fail to improve outcomes and waste scarce health care resources (31, 32). In addition, this approach creates a rigorous foundation for performance measures that until this point has been lacking (4). Rather than crafting performance measures based on intuition or convenience, performance measures would be based on clear processes of care that have been vetted in an organized, systematic fashion, taking into account the quality of evidence and values of key stakeholders.

This framework also helps ensure that performance measures are developed in an open and transparent way. A key strength of GRADE is its transparency, in that it allows guideline users to understand and track exactly how and why recommendations were made. Similarly, performance measure development must follow a transparent process with broad stakeholder involvement to ensure relevance to those with an interest in quality assessment (33). Basing performance measures on rigorously vetted guideline recommendations limits the possibility that stakeholders with narrow objectives influence measure development by advancing measures that are designed to achieve specific end results regardless of the evidence base.

Limitations of this Approach

There are several important limitations to performance measures developed using the process we propose above. First, although this framework precludes the translation of weak recommendations into performance measures, it is possible that some weak recommendations could improve patient outcomes if broadly applied. It is possible that weakly recommended treatments may improve outcomes on average, even if they harm some patients. Furthermore, there may be inherent value in simply reducing variation, even if those standards are based on weak recommendations. This approach carries the risk that meaningful opportunities for performance improvement are missed. However, codifying weak recommendations as performance measures carries greater risks, including wasted resources invested in quality-improvement activities that do not improve care and the chance that legitimate practice variation may be punished.

Second, our suggested approach applies primarily to creation of performance measures intended for broad application across diverse sites; measures developed locally to track quality-improvement efforts within a site may not need to follow such stringent criteria. Nonetheless, even in the local setting, measure developers should be cognizant of opportunity costs—creating measures intended to reduce variation by uniformly implementing weak recommendations (in which there is uncertainty about the balance between benefits and harms) may not be worth pursuing.

Third, although GRADE-based recommendations are transparent, they are subjective and conditional on the varying judgments of the developers. Nonetheless, a strength of the GRADE system is that it explicitly requires a detailed and transparent description of the underlying judgments and values that influence a recommendation (34). Thus, performance measure developers can consider these judgments and values when considering which recommendations to turn into performance measures.

Future Directions

Several steps can be taken now to further this process. First, as new guidelines are written there should be stricter integration with performance measurement—indeed, the two efforts should proceed synchronously. Guideline developers should think *a priori* about performance measurement, and quality experts should help ensure that recommendation statements are phrased in a way to aid translation into performance measures. Some investigators are already experimenting with informatics approaches to standardize this process (35). Guideline and measure development committees engaged in this process should be multidisciplinary, with input not only from clinical experts, guideline methodologists, and quality professionals but also patients, payers, quality improvement groups, and other key stakeholders.

Second, future work is needed to address how to identify the highest leverage performance measures for a given disease state. This is particularly important for patients with multiple chronic conditions, for which adherence to all possible performance metrics may not be achievable (36). One approach would be to systematically integrate GRADE-based guideline recommendations with outcomes studies demonstrating the strengths of association, costs, and the opinions of key stakeholders. Work is also needed on how individual recommendations can or should be aggregated into a composite measure or bundle. Bundling together multiple weak recommendations may lead to improvements in outcomes (37). Research is needed into how to create these bundles effectively and transparently.

Third, future work is needed to determine the optimal method to validate performance measures before their widespread use. Ideally, validation studies would determine if the measure includes the appropriate patient population, is operationalized correctly, and can be implemented without untoward adverse consequences. Current practice frequently includes evaluating performance measures in demonstration projects, but this process is inconsistent. However, it may be overly onerous to require that measure developers ensure that proposed measures do more good than harm before clinical application. A rigorous framework for measure development like the one we propose may help prevent adverse consequences, making it more tenable to evaluate measure impact after implementation, rather than before.

Finally, research is needed to empirically compare performance measures developed using this process to performance measures developed *ad hoc*. Although we believe that this framework leads to more useful and valid performance measures and minimizes the creation of unsuccessful measures, this hypothesis should be tested. Additionally, the acceptability of both the approach and the resulting sample performance measures should be vetted with patients, payers, and other stakeholders not represented in our ATS workshop. Nonetheless, we believe the proposed approach could represent a significant advancement in the creation of performance measures that are evidence based, clinically relevant, and patient centered.

Conclusions

In this workshop report we propose a framework for developing performance measures using GRADE-based clinical practice guidelines. Under this framework, process-based performance measures would only be translated from strong

recommendations based on high- or moderate-quality evidence. Ideally, this framework would overcome the limitations of many existing performance measures that are often based on low-quality evidence and lead to interventions resulting in minimal, if any, improvement in quality (3). Although this process may add the burden of measure development, the tradeoff would be stronger measures more likely to change practice. By transparently linking the strength of recommendations to the quality of the evidence and values surrounding the outcome, GRADE can uniquely lend itself to the creation of meaningful and useful performance measures. This approach does not resolve all the problems with health care performance measurement and may create some new challenges. Thus, it should be followed by validation studies to evaluate its value. Nonetheless, it represents a substantial step forward in the effort to make sure that health care delivery aligns with patient preferences and goals. ■

This official ATS Workshop Report was prepared by an *ad hoc* subcommittee of the Quality Improvement Committee, Document Development and Implementation Committee, and Behavioral Sciences and Health Services Research Assembly.

Members of the Writing Committee are as follows:
Jeremy A. Kahn, M.D., M.S. (Co-Chair)
Michael K. Gould, M.D., M.S. (Co-Chair)
Jerry A. Krishnan, M.D., Ph.D. (Co-Chair)
Kevin C. Wilson, M.D. (Co-Chair)
David H. Au, M.D., M.S.
Laura C. Feemster, M.D., M.S.
Colin R. Cooke, M.D., M.S.
Ivor S. Douglas, M.D.
Richard A. Mularski, M.D., M.S.H.S., M.C.R.
Christopher G. Slatore, M.D., M.S.
Renda Soylemez Wiener, M.D., M.P.H.

Author Disclosures: J.M.K. has received research support from Cerner Corp. ($1–$4,999). M.K.G. has received research support from Archimedes, Inc. ($10,000–$49,999). J.A.K. has consulted for Adelphi Values ($1–$4,999), eMAX Health ($1–$4,999), and UpToDate ($1–$4,999). K.C.W. has investment accounts with State Street Bank that are independently managed by Moody Lynn and Co. and previously may have included healthcare-related holdings. D.H.A. has received research support from Gilead ($100,000–$249,999), and consulted for Bosch Healthcare ($1–$4,999) and Nexcura, LLC (no payment received). C.R.C., I.S.D., L.C.F., R.A.M., C.G.S. and R.S.W. reported no relevant financial relationships with commercial interests.

Workshop Attendees:
Jeremy A. Kahn, M.D., M.S.
Michael K. Gould, M.D., M.S.
Jerry A. Krishnan, M.D., Ph.D.
Kevin C. Wilson, M.D.
David H. Au, M.D., M.S.
Safwan Badr, M.D.
Margaret Carno, Ph.D., M.B.A., R.N., C.P.N.P.
Laura C. Feemster, M.D., M.S.
Colin R. Cooke, M.D., M.S.
Ivor S. Douglas, M.D.
Christine Fukui, M.D.
John Heffner, M.D.
David Hiestand, M.D., Ph.D.
Robert Hyzy, M.D.
Suzanne Lareau, M.D.
Richard A. Mularski, M.D., M.S.H.S., M.C.R.
Adrienne Prestridge, M.D.
Christopher G. Slatore, M.D., M.S.
Marianna Sockrider, M.D.
Holger Shunemann, M.D., M.Sc., Ph.D.
Gerard Turino, M.D.
Curtis Weiss, M.D., M.S.
Renda Soylemez Wiener, M.D., M.P.H.

Acknowledgment: The authors thank Jessica Wisk for her efforts in coordinating the workshop and assisting with the document's development. They also thank Elizabeth McGlynn, Ph.D. for her insight and advice at the workshop.

References

1 Blumenthal D. Performance improvement in health care—seizing the moment. *N Engl J Med* 2012;366:1953–1955.
2 Shojania KG, Grimshaw JM. Evidence-based quality improvement: the state of the science. *Health Aff (Millwood)* 2005;24:138–150.
3 Wilson KC, Schünemann HJ. An appraisal of the evidence underlying performance measures for community-acquired pneumonia. *Am J Respir Crit Care Med* 2011;183:1454–1462.
4 Heffner JE, Mularski RA, Calverley PM. COPD performance measures: missing opportunities for improving care. *Chest* 2010;137: 1181–1189.
5 Werner RM, Bradlow ET. Relationship between Medicare's hospital compare performance measures and mortality rates. *JAMA* 2006; 296:2694–2702.
6 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;26:492–499.
7 Schünemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, Ernst A, Fahy BF, Gould MK, Horan KL, Krishnan JA, *et al*.; ATS Documents Development and Implementation Committee. An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. *Am J Respir Crit Care Med* 2006;174:605–614.
8 Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924–926.
9 Eddy DM. Clinical decision making: from theory to practice. Anatomy of a decision. *JAMA* 1990;263:441–443.
10 Methodology Committee of the Patient-Centered Outcomes Research Institute (PCORI). Methodological standards and patient-centeredness in comparative effectiveness research: the PCORI perspective. *JAMA* 2012;307:1636–1640.
11 Qaseem A, Wilt TJ, Weinberger SE, Hanania NA, Criner G, van der Molen T, Marciniuk DD, Denberg T, Schünemann H, Wedzicha W, *et al*.; American College of Physicians; American College of Chest Physicians; American Thoracic Society; European Respiratory Society. Diagnosis and management of stable chronic obstructive pulmonary disease: a clinical practice guideline update from the American College of Physicians, American College of Chest Physicians, American Thoracic Society, and European Respiratory Society. *Ann Intern Med* 2011;155:179–191.
12 McGlynn EA, Asch SM. Developing a clinical performance measure. *Am J Prev Med* 1998;14:14–21.
13 Brook RH, McGlynn EA, Shekelle PG. Defining and measuring quality of care: a perspective from US researchers. *Int J Qual Health Care* 2000;12:281–295.
14 Donabedian A. The quality of medical care. *Science* 1978;200: 856–864.
15 Rubin HR, Pronovost P, Diette GB. From a process of care to a measure: the development and testing of a quality indicator. *Int J Qual Health Care* 2001;13:489–496.
16 National Quality Measures Clearinghouse. Desirable attributes of a quality measure [accessed 2014 Apr 10]. Available from: http://www.qualitymeasures.ahrq.gov/tutorial/attributes.aspx.
17 Armstrong K, Hughes-Halbert C, Asch DA. Patient preferences can be misleading as explanations for racial disparities in health care. *Arch Intern Med* 2006;166:950–954.
18 Kahn JM. Quality improvement in end-of-life critical care. *Semin Respir Crit Care Med* 2012;33:375–381.
19 Ferris TG, Vogeli C, Marder J, Sennett CS, Campbell EG. Physician specialty societies and the development of physician performance measures. *Health Aff (Millwood)* 2007;26:1712–1719.
20 Dedier J, Singer DE, Chang Y, Moore M, Atlas SJ. Processes of care, illness severity, and outcomes in the management of community-acquired pneumonia at academic hospitals. *Arch Intern Med* 2001; 161:2099–2104.

21 Meehan TP, Fine MJ, Krumholz HM, Scinto JD, Galusha DH, Mockalis JT, Weber GF, Petrillo MK, Houck PM, Fine JM. Quality of care, process, and outcomes in elderly patients with pneumonia. *JAMA* 1997;278:2080–2084.
22 Lee JS, Primack BA, Mor MK, Stone RA, Obrosky DS, Yealy DM, Fine MJ. Processes of care and outcomes for community-acquired pneumonia. *Am J Med* 2011;124:1175.e9–17.
23 Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, DeBeer H, *et al*. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol* 2011;64:383–394.
24 Altman DG, Moher D, Schulz KF. Improving the reporting of randomised trials: the CONSORT Statement and beyond. *Stat Med* 2012;31:2985–2997.
25 Raghu G, Collard HR, Egan JJ, Martinez FJ, Behr J, Brown KK, Colby TV, Cordier J-F, Flaherty KR, Lasky JA, *et al*.; ATS/ERS/JRS/ALAT Committee on Idiopathic Pulmonary Fibrosis. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am J Respir Crit Care Med* 2011;183:788–824.
26 Shiffman RN, Dixon J, Brandt C, Essaihi A, Hsiao A, Michel G, O'Connell R. The GuideLine Implementability Appraisal (GLIA): development of an instrument to identify obstacles to guideline implementation. *BMC Med Inform Decis Mak* 2005;5:23.
27 Qaseem A, Wilt TJ, Shekelle P. A clinical practice guideline update on the diagnosis and management of stable COPD. *Ann Intern Med* 2012;156:69.
28 Barr J, Fraser GL, Puntillo K, Ely EW, Gélinas C, Dasta JF, Davidson JE, Devlin JW, Kress JP, Joffe AM, *et al*.; American College of Critical Care Medicine. Clinical practice guidelines for the management of pain, agitation, and delirium in adult patients in the intensive care unit. *Crit Care Med* 2013;41:263–306.
29 An official American Thoracic Society Clinical Practice Guideline. Sleep apnea, sleepiness, and driving risk in non-commercial drivers. An update of a 1994 Statement *Am J Respir Crit Care Med* 2013; 187:1259–1266.
30 Stein BD, Charbeneau JT, Lee TA, Schumock GT, Lindenauer PK, Bautista A, Lauderdale DS, Naureckas ET, Krishnan JA. Hospitalizations for acute exacerbations of chronic obstructive pulmonary disease: how you count matters. *COPD* 2010;7:164–171.
31 Roski J, Jeddeloh R, An L, Lando H, Hannan P, Hall C, Zhu S-H. The impact of financial incentives and a patient registry on preventive care quality: increasing provider adherence to evidence-based smoking cessation practice guidelines. *Prev Med* 2003;36:291–299.
32 Glickman SW, Ou F-S, DeLong ER, Roe MT, Lytle BL, Mulgund J, Rumsfeld JS, Gibler WB, Ohman EM, Schulman KA, *et al*. Pay for performance, quality of care, and outcomes in acute myocardial infarction. *JAMA* 2007;297:2373–2380.
33 Campbell SM, Braspenning J, Hutchinson A, Marshall M. Research methods used in developing and applying quality indicators in primary care. *Qual Saf Health Care* 2002;11:358–364.
34 Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, Schünemann HJ; GRADE Working Group. Going from evidence to recommendations. *BMJ* 2008;336:1049–1051.
35 Shiffman RN, Michel G, Krauthammer M, Fuchs NE, Kaljurand K, Kuhn T. Writing clinical practice guidelines in controlled natural language. In: Fuchs NE, editor. Controlled natural language. Heidelberg: Springer; 2010. pp. 264–280.
36 Fabbri LM, Boyd C, Boschetto P, Rabe KF, Buist AS, Yawn B, Leff B, Kent DM, Schünemann HJ; ATS/ERS Ad Hoc Committee on Integrating and Coordinating Efforts in COPD Guideline Development. How to integrate multiple comorbidities in guideline development: article 10 in Integrating and coordinating efforts in COPD guideline development. An official ATS/ERS workshop report. *Proc Am Thorac Soc* 2012;9:274–281.
37 Levy MM, Dellinger RP, Townsend SR, Linde-Zwirble WT, Marshall JC, Bion J, Schorr C, Artigas A, Ramsay G, Beale R, *et al*.; Surviving Sepsis Campaign. The Surviving Sepsis Campaign: results of an international guideline-based performance improvement program targeting severe sepsis. *Crit Care Med* 2010;38:367–374.