Resource Use Measurement White Paper: Commenting Draft

TO: NQF Members

FR: NQF Staff

RE: Comment for Resource Use Measurement White Paper and Proposed Resource Use Evaluation Criteria

DA: September 13, 2010

In October 2009, NQF initiated a two-phase project aimed at endorsing resource use measures. Prior to the Call for Measures in Phase Two of the project, NQF convened a Steering Committee representing diverse stakeholders in an effort to understand the full implications of this endeavor for NQF and relevant stakeholders. During Phase One, the Committee was asked to identify the unique attributes of resource use measures that should be considered during evaluation of these measures.

A primary task for this Committee during Phase One was to contribute to and provide guidance to the development of the Resource Use Measurement White Paper. This paper details the resource use measure specification process and identifies the specific issues that present when developing and evaluating these measures, and ultimately informs the Resource Use Measure Evaluation Criteria (Appendix B) that will be used to evaluate the measures for endorsement in Phase Two. NQF and the Resource Use Steering Committee are seeking comment on the white paper content, including the proposed criteria in Appendix B.

Pursuant to section II.A of the Consensus Development Process v. 1.8, this draft document, along with the accompanying material, is being provided to you at this time for purposes of review and comment only—not voting. You may post your comments and view the comments of others on the NQF website. **Public comments must be submitted no later than 6:00 pm ET, October 4, 2010. NQF Member comments must be submitted no later than 6:00 pm ET, October 12, 2010.**

NQF uses a program that facilitates electronic submission of comments on this draft report. **All comments must be submitted using the online submission process.** Supporting documents related to your comments may be submitted by **e-mail to efficiency@qualityforum.org** with "Resource Use White Paper & Criteria" in the subject line and your contact information in the body of the e-mail.

Thank you for your interest in NQF's work. We look forward to your review and comments.

Table of Contents

Exhibits

Appendices

Section 1. Measuring Efficiency and Resource Use in Healthcare

Over the past several years, quality measures and quality measurement initiatives have provided important information to the healthcare community. Yet despite these ongoing efforts, information on the value provided for dollars spent in healthcare is not readily available. Development of efficiency measures is one area that has lagged behind measure development activities focused on quality. One reason for this measures gap is the lack of agreement about how to measure efficiency or how to improve it.

In its final report *Identifying, Categorizing, and Evaluating Health Care Efficiency Measures,* the Agency for Healthcare Research and Quality (AHRQ) identified the following four areas that need to be addressed to improve the measurement of efficiency in the future:

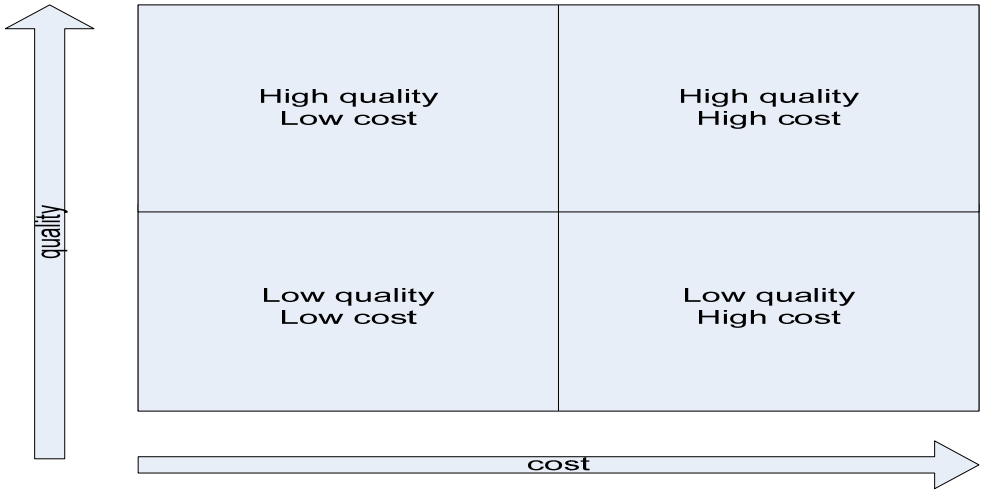
- the multiplicity of perspectives on the definition of efficiency;
- the gap between evidence-based measures and those in actual use;
- the absence of the quality dimension in efficiency measures; and
- the lack of validation or evaluation of the measures.[1]

For improvement to take place, efficiency and cost metrics must be clear, concise, and credible. Developing efficiency and cost measures, taking into account the quality domain, is an important component of transparency, which will eventually lead to improved health and efficiency across healthcare organizations.

For the purposes of this paper, *efficiency of care* is defined as a measure of the relationship of the cost of care associated with a specific level of performance measured with respect to the other five Institute of Medicine (IOM) aims of quality—that is, healthcare should be: safe, timely, effective, efficient, equitable, and patient centered.[2] Thus, true efficiency of care measures tend to be complex and encompass the concepts of both quality and resource use.

An illustration of the interaction of quality and cost or resource use is provided below:

Exhibit 1: Quadrant Display of Cost and Quality Dimensions



This illustration allows users to assess and compare the level of cost or resource use achieved by a provider or other entity without obscuring the level of quality; this illustrative approach adheres to the principle that quality (or health outcomes) is a dimension in the evaluation of the efficiency of care. Thus, a measurement effort that provides information for providers on both their quality outcomes and resource use or cost would consider those with high quality and low resource use as demonstrating higher efficiency than those with low quality and high resource use.

Measures of resource use are broadly applicable and comparable measures of health services counts (in terms of units or dollars) that are applied to a population or event (broadly defined to include diagnoses, procedures, or encounters). A resource use measure counts the frequency of defined health system resources; some may further apply a dollar amount (e.g., allowable charges, paid amounts, or standardized prices) to each unit of resource use—that is, *monetize* the health service or resource use units. The approach to monetizing resource use varies and often depends on the perspective and purpose of the measurement effort. Monetizing resource use is an attempt to *weight* counts appropriately. For example, a frequency count of outpatient visits would give an equal count of one to both an office visit with an evaluation and an office visit with a procedure. Monetizing this would give a larger value to the office visit with a procedure.

Because it accounts for the variation in intensity of services, it allows for resource use results to be rolled up into one measure result.

Focus of the Project

This project, funded by the Department of Health and Human Services (HHS), ultimately will result in resource use measures that can complement the quality measures NQF already has endorsed and that the healthcare community is using currently. The project initially will endorse resource use measures, which will serve as a building block for efficiency of care measures and as a signal to the measure development industry of the urgent need to endorse useable resource use measures and develop measures of efficiency of care that integrate the quality domains. Currently we know there are large numbers of resource use measures that providers and purchasers are using, including episode-based and population-based measures. The ability to which any one resource use measure brings us closer to efficiency of care (which includes outcomes), while of interest, will not be evaluated. For emerging measures, such as composites, outcomes, efficiency, and resource use measures, it is anticipated that additional guidance will be required beyond the Standard NQF evaluation criteria.

White Paper Organization

This white paper was developed with input from a variety of stakeholders and under the direction of the NQF Resource Use Steering Committee. It is intended to provide background information and identify issues associated with the evaluation of these types of measures. Further, the paper will explore key methodological issues of resource use measurement approaches, which will provide information on implementation. Overall, this paper will assist in adapting the existing NQF measure evaluation criteria to ensure that resource use measures are appropriately evaluated.

Key Terms and Definitions

The following are terms and definitions that are important to understanding the concepts presented in this paper.

Attribution: identification and assigning of a responsible provider or entity (e.g., health plan) to the care delivered to a resource unit or population.

Temporal: occurring over a sequence of time or within a particular time; refers to the timeframe and related measure logic specified in a measure.

Standardized price: pre-established uniform price for a service, typically based on historical price, replacement cost, or an analysis of completion in the market; removes variation in resource costs due to differences in negotiated prices.

Monetize: to apply a dollar amount (actual charges, standard price) to a unit of resource use. Monetizing resource use is an attempt to *weight* counts or resource units appropriately. For example, a frequency count of outpatient visits would give an equal count of one to both an office visit with an evaluation and an office visit with a procedure. Monetizing this would give a larger value to the office visit with a procedure.

Efficiency of care: a measure of cost of care associated with a specified level of health outcomes. AQA defines efficiency as a measure of cost of care associated with a specified level of quality of care.

Quality of care: AQA defines quality of care as a measure of performance on IOM's six aims for healthcare: safety, timeliness, effectiveness, efficiency, equity, and patient centeredness.

Cost of care: AQA defines cost of care as the total healthcare spending, including total resource use and unit price, by payor or consumer, for a healthcare service or group of healthcare services associated with a specified patient population, time period, and unit of clinical accountability.

Value of care: AQA defines value of care as a specified stakeholder's (such as an individual patient's, consumer organization's, payor's, provider's, government's, or society's) preference-weighted assessment of a particular combination of quality and cost of care performance.

Resource use measures: broadly applicable and comparable measures of health services counts (in terms of units or dollars) applied to a population or event (broadly defined to include diagnoses, procedures, or encounters). A resource use measure counts the frequency of defined health system resources; some may further apply a dollar amount (e.g., allowable charges, paid amounts, or standardized prices) to each unit of resource use—that is, *monetize* the health service or resource use units.

Resource unit: the resources used to provide care to a patient or population. Resource units are generally identified through claims data and measured in terms of dollars, but can also include resource not captured on a claim, e.g., nursing hours.

Stratification: division of a population or resource services into distinct, independent strata, or groups of similar data, enabling analysis of the specific subgroups. This type of adjustment can be used to show where disparities exist or where there is a need to expose differences in results.

Risk adjustment: a corrective approach designed to reduce any negative or positive consequences associated with caring for patients of higher or lower health risk or propensity to require health services.

Sensitivity: the proportion of actual positives that are correctly identified as such (e.g., the percentage of people with diabetes who are correctly identified as having diabetes).

Specificity: the proportion of negatives that are correctly identified (e.g., the percentage of healthy people who are correctly identified as not having the condition). Perfect specificity would mean that the measure recognizes all actual negatives—for example, all healthy people will be recognized as healthy.

Importance to report and measure: NQF criterion focused on evaluating the extent to which the measure focus is important in exposing areas of high impact.

Scientific acceptability of measure properties: NQF criterion focused on evaluating the extent to which the measures, as specified, produce consistent (reliable) and accurate (valid) results about the cost or resources used to deliver care.

Feasibility: NQF criterion focused on evaluating the extent to which the required data are accessible and retrievable without undue burden, and the degree to which the measure can be implemented for internal improvement and public reporting.

Usability: NQF criterion focused on evaluating the extent to which the intended audiences find the information the measure produces to be meaningful, understandable, and useful both for public reporting and internal improvement.

Section 2. Designing Measures that Acknowledge the Real World While Producing Useable Output

Purchasers, health plans, providers, and policymakers want and use resource use performance measures to inform and support improvement efforts. Accurate methods of cost estimation and other key methodological and policy issues must be considered, including carefully weighed criteria for evaluating resource use measures to be used for improvement and public reporting. A gap in the measurement field exists, however, between the ideal performance measurement approach and the measures and methods that are available and implemented.[3] Ideally, the healthcare system would be subject to a comprehensive measurement approach that accurately and reliably assesses each of the six IOM aims of quality.[4]

Several recent NQF reports and ongoing projects examine various measurement issues. In 2007, NQF convened a Steering Committee to develop a framework for evaluating the efficiency of care over time, including clear definitions and a shared vision of what can be achieved around quality, cost, and value. This framework served as a foundation for the work of larger performance improvement efforts (such as the *Evaluating Efficiency Across Patient-focused Episodes of Care* framework). This report presents the NQF-endorsed® measurement framework for assessing efficiency, and ultimately value, associated with care over the course of an episode of illness and sets forth a vision to guide ongoing and future efforts. In this effort, the Steering Committee adopted the definitions of quality, cost, value, and efficiency of care used by the Ambulatory Care Quality Alliance (AQA):[5,6]

- **Quality of care** is a measure of performance on IOM's six aims for healthcare: safety, timeliness, effectiveness, efficiency, equity, and patient centeredness.
- **Cost of care** is a measure of the total healthcare spending, including total resource use and unit price(s), by payor or consumer, for a healthcare service or group of healthcare services associated with a specified patient population, time period, and unit(s) of clinical accountability.
- **Efficiency of care** is a measure of cost of care associated with a specified level of quality of care. Efficiency of care is a measure of the relationship of the cost of care associated

with a specific level of performance measured with respect to the other five IOM aims of quality.

- **Value of care** is a measure of a specified stakeholder's (such as an individual patient, consumer organization, payor, provider, government, or society's) preference-weighted assessment of a particular combination of quality and cost of care performance.

Additionally, the Composite Measure Evaluation Framework provides the background, rationale, and evaluation criteria for composite measures. A composite measure is a combination of two or more individual measures in a single measure that results in a single score. NQF has also engaged in a comprehensive effort culminating in an upcoming report on the *Measurement Implications of Payment Reform Models*, to be published in October 2010, that will discuss how current performance measures should be applied to new payment models, such as accountable care organizations (ACOs) and medical homes, and suggest areas for measure development to support these new models. Further, NQF initiated a project to develop a measurement framework for multiple chronic conditions that will serve as a foundation for the future endorsement of performance measures that explicitly address multiple chronic conditions. Measure developers have pursued various paths toward meeting the goal of performance improvement, with each seeking to strike a balance between the perfect measurement approach and the reality of developing and implementing feasible measures of resource use.

Current approaches for measuring resource use range from broadly focused measures, such as per capita measures, which address total healthcare spending (or resource use) per person, to those with a more narrow focus, such as measures dealing with the healthcare spending or resource use of an individual procedure, e.g., a hip replacement (see Exhibit 2).

Exhibit 2: Spectrum of Resource Use Measurement Approaches

Examining the spectrum above, we see there are many types of resource use measures that by design are narrower in focus. The fundamental tradeoff among these approaches lies in their

degree of specificity or focus and the care delivery being measured. A highly specific or narrow resource use measure—for example, the cost of cataract surgeries performed by ophthalmologists (not subspecialists or those affiliated with a teaching hospital) on women aged 65 to 75 years old with hypertension, diabetes, and no other comorbidities—results in a highly homogeneous measure of analysis. This tightly defined measure of analysis increases direct comparison because of its high degree of specificity, but this specificity results in few instances for each measure's provider-patient combination.

Alternatively, broad measures, which trade sensitivity for specificity, identify more services and patients. While these broad measures allow users more flexibility in examining services across combinations of conditions, providers, and settings, this reduction in specificity results in a more heterogeneous measure of analysis that requires more sophisticated risk adjustment. Using both types of measures simultaneously may be ideal, providing users with a comprehensive understanding and broad view of the resources being used along with the ability to identify specific sources of high or low resource use that require action. As an example, the implications of preventive services on hospital admissions are often discussed and examined. A highly specific or narrow measure evaluating a preventive service (e.g., outpatient imaging resource use) would not capture broader implications of the imaging studies if the use of some advanced imaging led to fewer hospital admissions. Implementing resource use measures and providing results that are actionable is critical and a key criterion for NQF evaluation of a measure. (We will discuss the evaluation criteria in Section 6.) Specifically, the results must be interpretable and target the appropriate and relevant audience, and they must be able to be used to take action.

As previously stated, an advantage of per capita measurement is that it measures all costs for each person in a population, thus providing a comprehensive view of health service resource use. Without additional adjustments or detail, however, the user's ability to interpret and take action to effect results is called into question. For example, it may be difficult to explain and identify causes for differences in total spending. Are they due to patient characteristics, provider differences, patient preferences, or differences in practice patterns? Therefore, to make measures more specific, measure developers often include further detail (e.g., splitting out total resource use by type of health service) or adjustments, such as risk adjustment or stratification, to make

the information more comparable and actionable. Advantages of episode-based measurement, which are farther to the right on the spectrum in Exhibit 2, include the fact that they are more specific, resulting in fairer, direct comparisons that are often considered more readily actionable by providers. For example, an episode-based measure will examine the resource use associated with a particular episode of illness or around a particular event. This more granular and focused resource use measurement approach, while often still requiring risk adjustment or stratification, provides users with more readily actionable results than per capita measure alone. For example, a per capita or per patient measure demonstrating a provider network as having high pharmacy resource use would require more information to take clinically sound and reasonable action. In contrast, a provider network demonstrating relatively high pharmacy use for an episode of chronic heart failure (CHF) would know to assess its prescribing patterns for patients presenting with a diagnosis of heart failure.

However, these strengths are also limitations; the episode-based measurement approach entails parsing out each patient's care into appropriate and often multiple episode measures (e.g., multiple episodes). Thus, while a provider network may do an optimal job in managing the resources for episodes of CHF, the same provider network might be less effective in managing resources for hip fracture. Episodes traditionally have been constructed on a condition-by-condition basis. Further, many patients have multiple conditions, and the resources used for their care are measured among multiple episode measures. Generally, multiple episodes are not designed to relate to one another and also do not necessarily add up to measure total resources used for the whole patient. Further, not all diagnoses, encounters, or events will be tied to a defined episode despite the potential association with the patient's resource use. Episode-based measure developers generally have tried to balance this condition-by-condition episode measurement tradeoff in two ways. First, some have opted to maintain the condition-by-condition approach but apply risk adjustment to each episode to account for patients with comorbid conditions that may or may not be a part of another episode measure. Second, some have developed an approach that allows for the comparison of total resource use patient to patient by matching patients based on a primary condition. One could argue that this latter adjustment falls somewhere between per patient and per episode measurement on the spectrum model.

In addition to per capita and per episode measurement, there are multiple options for service-specific measurement, usually focused on an admission (e.g., hospitalization) or a procedure (on the right of the spectrum in Exhibit 2). The highly specific design of these measures can provide users with results that require little further manipulation, while still addressing services that account for a substantial share of total healthcare spending (e.g., inpatient resource use). For some providers, such as surgeons and hospitalists, the results from these types of measures may be the most actionable. Again, these advantages also can be drawbacks. Because these measures examine an individual service or admission and only directly related services, such as a hospitalization and healthcare services 30 days post-discharge, they often miss services or conditions that led to the hospitalization, do not adjust for comorbidities, and are often short-term measures. Thus, these highly specific measures do not include important information about the conditions or services leading up to the occurrence, the need for the services, or the repercussions stemming from them at any length. Lastly, this approach provides no insight in approaches to optimizing the mix of health services—critical information to moving the system to more optimal resource use. To the extent that measure developers try to add this context, service-specific measures move to the left on the spectrum.

For provider profiling or reporting applications, different resource use measures can produce important differences in results. This is true when providers are being examined based on different types of resource use measures or when users are applying different options in the methodology to the same resource use measure to the same provider. This is an important complexity encountered when implementing resource use measures of all types and has caused a substantial amount of confusion, frustration, and anxiety for providers and those who implement these measures. Current methods often allow user discretion regarding specification of the measurement options (e.g., outlier, thresholds, or peer group decisions), and the degree of discretion varies by measure developer and by the type of resource use measure. This variance can result in the same provider or provider network having different final resource use results for the same (or seemingly the same) resource use measure. In one study, Thomas et al. compared the predictive accuracy and consistency of methods used for provider profiling, finding that while there was much consistency overall, different software identified different providers as

relatively high cost or low cost. [7] This situation also occurs when physicians and providers are measured by different payors, which have access to only some of the providers' claim data and thus cannot examine the practice patterns in whole. A critical challenge and consideration is how to distinguish between factors influenced by physician's or provider's decisions and those factors that are beyond the control of the provider. While some measures by design attempt to rectify this, e.g., comparing costs for services linked to a specific episode among the same physician specialists, some differences outside their control will still exist.[8] This challenge is especially apparent when measuring and reporting results for individual physicians.

Taking into consideration all the advantages and limitations discussed, there are reasons to opt for simultaneously implementing measures that are broad and incorporate many conditions or patients and measures that are narrower in focus. Regardless of the type of resource use measure that is developed and implemented, all should meet measurement properties and criteria discussed in this paper. Specifically, they should contribute to understanding the current state of the healthcare system, have been thoroughly vetted with experts and empirically tested to establish their credibility, and support decision making, and they should not be prohibitive for users to implement. The next sections will address resource use measurement approaches in greater detail and discuss some of the challenges they may encounter in meeting the identified measurement properties or criteria. The final section discusses in detail the proposed NQF Resource Use Evaluation Criteria, which is based on the current NQF criteria.

Section 3: Perspective and Types of Resource Use Measures

This section discusses the importance of *perspective* and defines the main types of resource use measures—per capita, per patient, per episode, per admission, and per procedure measures of resource use. The descriptions are provided to facilitate discussion about the major criteria for evaluating resource use measures; the evaluation of resource use measures is discussed in more detail in Section 6. Although comparisons may be drawn among the different measure types, the objective is not to pick one best type, but rather to elucidate some distinct features of each. As previously discussed, one measure type alone may not be the best option for assessing and addressing resource use. Further, to drive performance improvement, measures should provide fair and meaningful comparisons across providers and account for the diversity of the population—taking into account various ages, races, ethnicities, genders, disabilities, socioeconomic conditions, geographic locations, and multiple chronic conditions. In this section we will discuss related recent and ongoing efforts NQF has undertaken, a conceptual model displaying the spectrum of resource use measures, and the implications of this model.

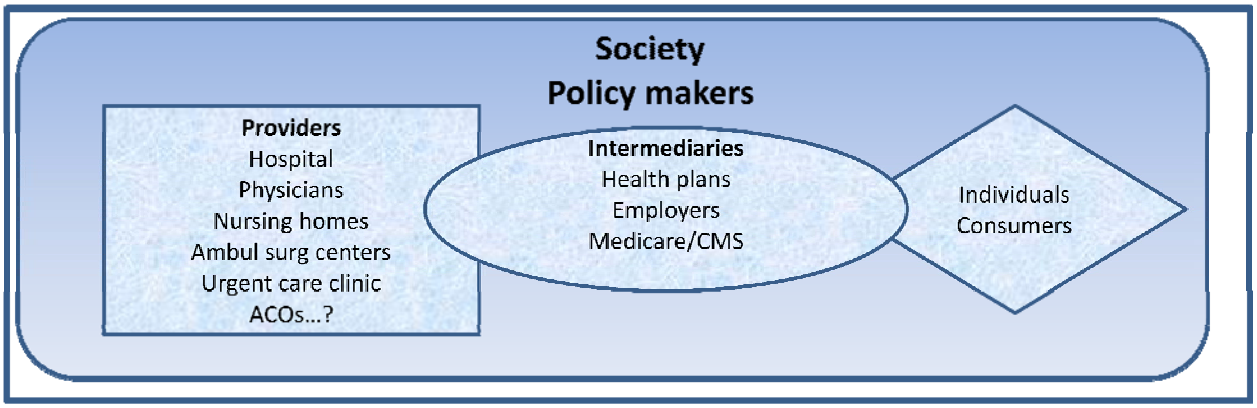
In the *Identifying, Categorizing, and Evaluating Health Care Efficiency Measures* report produced for the Agency for Healthcare Research and Quality (AHRQ) in 2008, the authors identify *perspective* (i.e., who is evaluating what and for what purpose) as one of the key levels of their typology for efficiency measures.[9] Arguably, this typology level applies to most healthcare measures, including resource use measures. Adapting their typology to resource use measures, we also identify four types of entities that encompass the perspectives of those that are evaluating and those that are being evaluated:

1. healthcare providers, including physicians and accountable care organizations;
2. intermediaries, including health plans and employers;
3. consumers or patients; and
4. society and policy makers.

All of these entities have varying control over resources and often distinct objectives for resource use measurement. Thus, when selecting a resource use measure, or a combination of them, it is critical that the measurement's purpose is well understood and the selection of measures is

related to this purpose. Further, depending on the objective for measurement, evaluators may become the evaluated entity and vice versa.

Exhibit 3. Resource Use Measurement Perspectives[10]



The table below lists and defines each of the types of resource use measure described in this paper and a potential example of its use, framed around perspective.

Exhibit 4. Resource Use Measure Examples and Definitions

Resource Use Measure Type	What Is It?	Example of Use—Perspective
Per capita-population based	All services provided to a person within a specific population, regardless of condition or encounters with system (e.g., health services provided per person 2 years and older residing in California)	Policy decision maker evaluates Medicare for the purpose of reducing unwarranted variation in resource use or cost or examining the effect of a policy change on resource use
Per capita-patient based	All services provided to a specified type of patient (e.g., health services provided for patients 18 years and older with a diagnoses of diabetes type 2)	An employer evaluates health plans for the purpose of contractual negotiations and agreements
Per episode	Bundles of services that are part of a distinctive event provided by one or multiple entities (e.g.,	A physician network evaluates physicians for the purpose of

	health services provided associated with an event or series of events for acute myocardial infarction)	payment for performance or other payment incentives
Per admission (e.g., hospitalization)	Bundles of services (including days) associated with an admission or stay (e.g., the length of stay for acute care hospital admissions)	An employer assesses hospitals with the purpose of reducing unwarranted variation in inpatient days, which affect resource use or cost
Per procedure	Bundles of services associated with a specific procedures (e.g., health service related to knee replacement surgery)	An ACO evaluates physicians for the purposes of reducing unwarranted resource use or cost associated with a procedure

Per Capita-population and Per Capita-patient

The phrase *per capita measurement* refers to measures of healthcare spending for populations in an area, regardless of any one person's exposure to the healthcare system. *Per patient measures* evaluate healthcare spending for an identified patient population, such as children of a certain age with asthma, and may be limited further (e.g., within an area or health plan). Depending on who is measuring what and for what purpose, both types of measures are useful and appropriate. For example, per capita measures that consider an entire population may be the optimal choice when purchasers or policy makers are evaluating large providers, such as an accountable care organization, where they are interested in the health services and outcomes for all persons for whom the entity is responsible, regardless of whether all persons received services during the measurement time period. Disadvantages to these types of measures include the need for a robust risk adjustment to account for the more heterogeneous nature of the measure's target population and the ability of end users to develop and implement actions to change the results when using these measures alone. The best-known example of per capita population-based measurement is the Dartmouth Atlas of Healthcare, which documents geographic variation in healthcare spending per capita using Medicare data to provide information and analysis about national, regional, and local markets, as well as hospitals and their affiliated physicians.[11] Alternatively, per patient measures may be an optimal choice for measuring physician network or group performance for patients treated during a 12-month period.

Per Episode

Episode-based measures use clinical logic to create units for measurement and assign claims to clinically distinct episodes of care. Specifically, the measures include a series of clinically related healthcare services over a defined time period, such as all claims related to a patient's diabetes. Episode-based measures use all types of healthcare claims (e.g., inpatient, physician professional services, outpatient services, and prescription drug services). Episode-based measures are by construction generally more homogenous than per capita or patient measures and thus do not require as powerful a risk adjustment. Further, because they limit their measurement area of interest to a specific episode of illness, they often provide more granular results, allowing for more apparent decisions based on their findings. Despite this advantage, they have some limitations. [12] The NQF Patient-Focused Episodes of Care Steering Committee concluded that episode-based measures do not necessarily distinguish the appropriateness of clinical services and patient preferences for the clinical services rendered; therefore, resource use measurement based purely on episodes should be balanced or accompanied by population-based, per capita resource use measures.[13]

Per Admission

Per admission measures (e.g., hospital admission measures) generally examine the resources used during a hospitalization and some period of time following the stay (e.g., 30 days). These types of measures may resemble episode-based measures but are typically more limited in the services and health settings captured. They may or may not include clinical logic to determine whether the services in the follow-on period are clinically related to the hospitalization.

Per Procedure

Procedure measures examine the resources used for surgeries and other procedures. These kinds of measures often include related pre- and post-procedure services, such as bandage removal and physical therapy, but are more limited in their scope compared an episode-based measure. For patients undergoing a knee replacement surgery, for example, pre-operative services might

include an EKG and physical to determine a patient's risk associated with the procedure. Postoperative services for these patients might include ambulatory physician visits or bandage removal. Similar to per admission measures, these measures might or might not include clinical logic to determine whether the services are clinically related to the procedure.

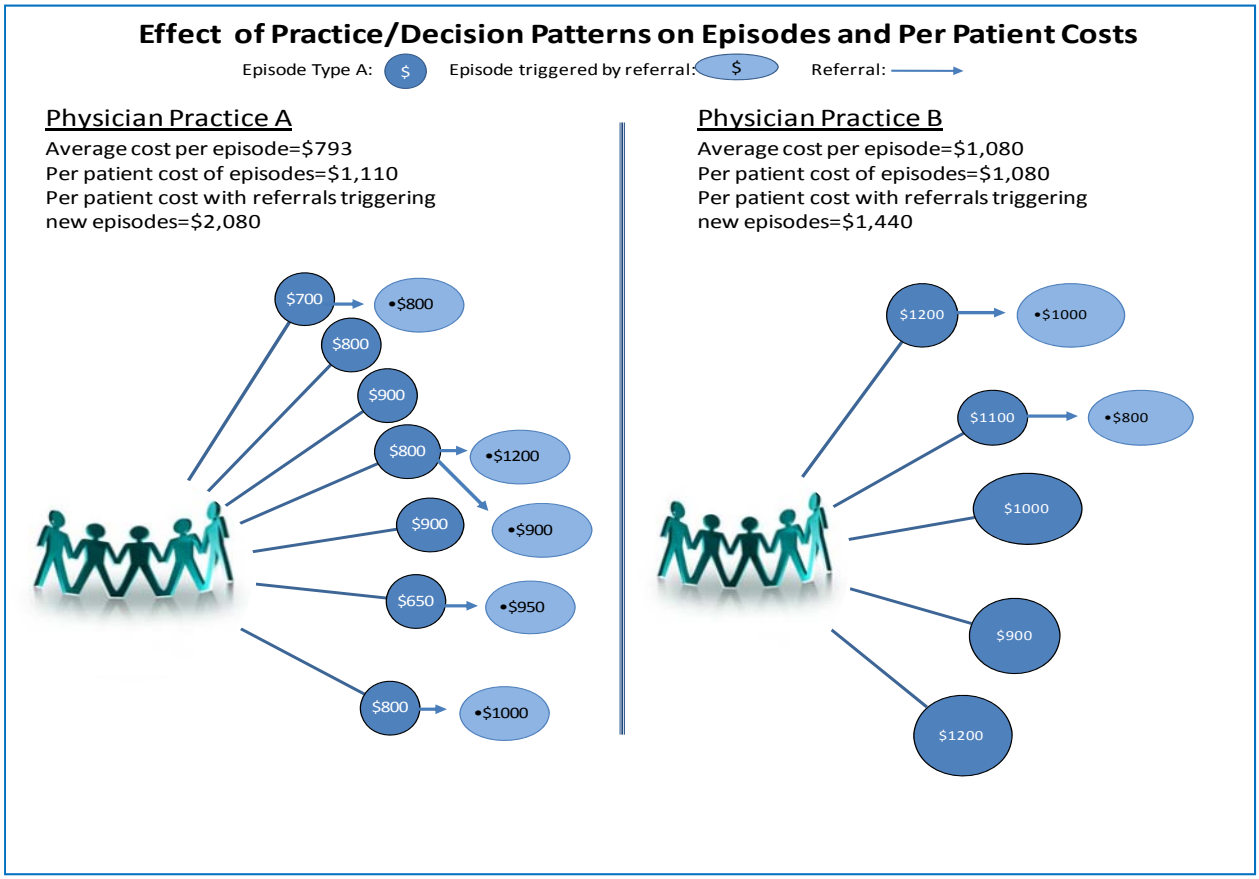
Using Resource Use Measures

In a 2009 report, MedPAC stated that physician level measurement efforts should be flexible enough to measure resource use on both a per episode and a per capita basis.[14] MedPac stated that these measurement types reported together capture more fully the relevant characteristics of physicians' practice patterns by revealing the resources they use in a given episode and the number of episodes they encounter per patient. [15] Further, the differences in the way physicians practice may influence how they compare to other physicians with similar patients. In the relatively straightforward example illustrated in Exhibit 5, for the same five-patient panel, Physician A has lower average episode costs for a particular episode of care; however, this physician's practice pattern results in a higher frequency of the episodes and a higher referral rate. Therefore, while Physician A has lower per episode or average episode costs, she has higher per patient costs when compared to Physician B. It is important to note this difference does not indicate if either physician employs standard practices of care or is associated with higher or lower outcomes. Instead, this scenario illustrates only that different slices and levels of resource use measures are necessary to develop sound policy and decisions to influence resource use. Additional measures—such as rate of prescribing generic drugs and use of basic versus advanced imaging— also should be included when warranted to produce a more complete picture of resource use.

While the NQF evaluation framework for resource measures follows NQF's standard evaluation criteria—against which all submitted resource use measures are individually evaluated in terms of importance to measure and report, scientific acceptability of measure properties, usability, and feasibility users of measures will need to account for perspective and provide a complete resource use picture of those being evaluated. When evaluating resource use measures, the NQF Resource Use Steering Committee has identified major analytic functions or modules that should

be explicitly included in the evaluation criteria for resource use measures. Specifically, the measures' data protocol, measure or episode clinical or construction logic, adjustments for comparability, profiling system, and assigning and reporting will need to be addressed. These considerations will be discussed in detail in the following section.

Exhibit 5: Effect of Practice/Decisions Patterns on Episodes and Per Patient Costs



Section 4: Resource Use Measure Modules

Estimating the resource use amount is only part of the resource use measurement process. A substantial number of decisions also must be made about input data, including their completeness, managing, or cleaning; certain claims, mapping, and grouping diagnostic codes or claims; and how to generate comparative information. Administrative data are the primary source for calculating resource use measures, and the analytic functions necessary to create valid and reliable measures for the purposes of comparability and public reporting are critical to standardized measurement. Specifically, measure users must gather and prepare the administrative data, create units for measurement, and make decisions about how the standard will be estimated, assigned, and compared. Resource use measure specifications must include the analytic functions and decisions for users to produce this type of measure based on the specified data. All analytic functions and decisions must be transparent and explicitly part of the specifications when applicable. When developers submit measure specifications to NQF for endorsement consideration, they must demonstrate a clear rationale and justification for any flexibility or decision by the user.

Resource use measurement approaches can be viewed as having five main analytic functions or modules: 1) data protocol, 2) measure clinical logic, 3) measure construction logic, 4) adjustments for comparability, and 5) measure reporting.[16] The data protocol module includes analytic steps like cleaning or aggregating the relevant data. The clinical logic module may include steps identifying which condition or event is of interest, including the specific diagnoses or procedure codes; any clustering or grouping of diagnoses or procedures into clinical categories; comorbid or disease interactions; as well as other clinically related algorithms. Once the clinical logic is identified, steps on which claims to cluster or group, and how, must be specified. These analytic steps are part of the construction logic. The construction logic includes temporal parameters and other decisions or parameters around the clinical logic (e.g., the trigger and termination rules for a specified episode of care), as well as identification of the resources to be measured. Adjustments for comparability are critical analytic functions for comparative and public reporting and include risk assessment and adjustment, approaches to stratification, and decisions about the costing method to be used. Many of the analytic functions in one way or another are attempting to make adjustments for comparative purposes; the manner in which they

are implemented may likely also vary depending on the perspective of the measurement effort. For example, for the purposes of feedback and confidential reporting to physicians, or when measuring large populations or entities, it may be acceptable to use a less powerful risk-adjustment approach; whereas, for the same measure when the purpose is public reporting, a more complete and vigorous risk adjustment may be necessary. The last module, reporting, or the analytic functions necessary to report resource use measures reliably and validly, includes steps to calculate a benchmark, attribute results to providers or eligible entity, and provide statistical information necessary to interpret findings when reported.

Measure Specification Steps by Module

Module 1. Data Protocol

Analytic steps that occur before the resource use measure identifies the populations, diagnoses, or procedures are designed to determine which data are necessary and adequate, which claims should be grouped, and whether any changes must be made to items on the claims. Preventing data errors in the first place is far superior to detecting errors and attempting to clean the data; however, errors do occur. Analytic functions designed to validate data and further address or account for data issues are critical to a measure's reliability and validity. While some decisions in the data protocol module are presented as options to the user, they are a critical part of implementing reliable and valid measures. Input data issues that affect the reliability and validity of a healthcare services measure are not necessarily captured in a claims edit system, where the primary concern is issues associated with billing. Some common issues that should be addressed include missing data due to capitated environments or because of carved-out or outsourced care relationships. For example, mental health services are often carved out, and the resulting claims data may not be available to those who are measuring resource use. Further, different types of administrative data have different types of data problems, including decisions about the number of diagnosis codes per record it will capture, which may affect the comparability among entities.

Input Data

An important step is to identify explicitly the types of data and aggregate or link these data so that the measure can be calculated reliably and validly. Examples include: enrollment data,

provider data, physician data (including physician specialty information), and claims or encounter data. Further, there are many types of claims data, which are not always collected and stored in the same database, such as pharmacy data feeds from a pharmacy benefit manager and physician professional claim data. The merging of two or more databases may create new errors (i.e., duplicate records).[17] However, caution is warranted—while information about the same event or service may appear in different data sets and treated as duplicates, in many cases the records in the different databases may include additional information that is unique and needs to be integrated into the measurement database.[18] This has implications for resource use measures that capture information across providers or care settings. An additional issue that may arise with merging databases is the mixing of data that are based on different criteria, different assumptions or units of measurements, and different quality control mechanisms.[19]

Data Cleaning

Before applying the clinical or construction logic to produce the measure, users generally conduct additional steps to clean the administrative data files, especially claims data. Prior to implementing a measure, data should be checked to identify inaccurate, incomplete, or unreasonable data, followed by steps to correct data errors or omissions. Steps may include format checks, completeness checks, reasonableness checks, limit checks, or review of the data to identify outliers (e.g., geographic, statistical, temporal) or other errors. Validation checks also may involve checking for compliance against policies and procedures. Typically this cleaning includes removing or truncating very high- and low-dollar-amount claims, unpaid claims, claims with missing information, and claims with questionable information.[20] These data cleaning steps are not always required; however, they do increase the reliability and validity of a measure's outcome by removing inaccurate information, reducing skewed results, and accounting for missing information (some software applications account for these issues) and are designed to ensure fair comparisons of physicians or other entities.

Inclusion and Exclusion

For each measure, decisions are made about which claims and patients to include or exclude in the analysis, regardless of any clinical or procedural event. For example, enrollment criteria may be established to include claims only for patients who were enrolled in a health plan for a full

year or who saw the physician at least once during the measurement period. This step helps ensure that the patient has had some minimal amount of exposure to the healthcare system (e.g., through a plan or physician). The length of enrollment or number of visits the patient to be included in a measure varies depending on the measure construction and what is being measured. For example, a chronic condition may require an entire calendar year, but an acute one might span only a few days. Further, it helps to ensure that the patient did not receive relevant care while not enrolled with the plan or that the patient was not part of the physician's panel of patients, possibly resulting in incomplete data and misleading information about overall resource use. Medicare, Medicaid, and private payers are all affected by enrollees moving in and out of plans during the year. Medicaid beneficiaries tend to gain and lose eligibility, thus moving in and out of the program. Private plan enrollees tend to change plans as they change jobs or during open enrollment periods. Medicare beneficiaries usually maintain their eligibility, but a significant share of Medicare beneficiaries are enrolled in Medicare Advantage plans and often move among such plans or between Medicare Advantage plans and the traditional Medicare benefit. CMS has claims data only for beneficiaries enrolled in traditional Medicare and does not obtain them for beneficiaries enrolled in Medicare Advantage plans. Further, unlike private plans and Medicaid, Medicare has incomplete prescription drug claim information for its beneficiaries. As a result, Medicare has full-year claims data only for some of its beneficiaries and is missing drug claims for a subset of this group.

Exclusion from measurement is not the only option when missing claims data or enrollment gaps exist. Measure developers and users rely on other methods, such as using statistical techniques like imputation, to assign values to missing data based on the available data. Measure specifications also can exclude patients from specific measures based on demographic characteristics (for example, excluding women for prostate cancer measures). Exclusions based on clinical (e.g., diagnostic or procedural) reasons often occur as part of the application of clinical logic analytic functions. Exclusions generally will be applied before claims data are used and grouped into units for measurement.

Module 2. Measure Clinical Logic

Diagnoses, procedures, and events do not always fit neatly into a measure leading to variation in how measure developers define the clinical logic for seemingly the same condition or event. Further, a patient may have two or more conditions that worsen his or her overall illness and increase the need for services exponentially. A measure's clinical logic includes the analytic functions to identify the conditions or events related to the measure's concept and intent. The clinical logic relies on identifying a clinical concept and deciding which diagnoses, events, or services are related to this concept. Measure developers may make different decisions about what comprises or is related to the clinical concept of interest based on input that includes clinical expert consensus or opinion, evidence-based guidelines, or empirical data. As part of this, measures are usually identified as resource use measures for *acute conditions, chronic conditions,* or *preventive services*, which often affects the clinical logic. Chronic and acute diseases can intersect or overlap, as in the case of a patient with CHF, a chronic condition, who has an acute myocardial infarction (AMI), an acute condition. Using this example, two measures may differ on whether the AMI is measured as a standalone acute measure or included in a chronic cardiac condition resource use measure. Measures of chronic disease either ignore or provide an analytic solution to account for how long an individual has lived with the chronic condition based on the assumption that as medical conditions progress, the clinical logic also may need to change. The analytic steps are designed to create appropriately homogeneous units for measurement (e.g., an episode of malignant neoplasm or patients with chronic obstructive pulmonary disease [COPD]).

Other analytic functions often executed as part of the clinical logic module include a hierarchy of conditions, which for any given patient maps diagnoses or events into discrete clinical categories. A broad clinical area may have more than one clinical category—for example, diabetes may have as many as four separate clinical categories to which diagnoses or events are mapped. Based on relative cost, resource use, or severity, these clinical categories are ranked among the related clinical conditions into hierarchies. Severity levels also can be assigned based on the patient's underlying health status. Both hierarchical and severity level rules are meant to increase the validity and comparability of results by addressing the variation in underlying health status among persons.

Module 3. Measure Construction Logic

The measure construction logic includes taking the analytic steps or making decisions that are based on the clinical logic and associated with temporal logic; assigning (or triaging) claims to the correct or best homogenous unit identified in the clinical logic, especially when similar or related units are present for the same patient; and appropriately assigning the health services to each measure. These decisions vary by measure and measure developer, even for the same clinical area, and thus have comparative measurement implications because varying time periods, claim, and health service assignment for similar resource use measures make it difficult to compare providers or health plans among approaches. Further, the perspective or purpose of the measure may influence which set of analytic decisions is best suited for the users of the measure. For example, a health system with a continuity of care objective may be interested in capturing health services related to COPD across many health settings and for longer periods of time, whereas a hospital measurement effort that does not include activities outside the hospital environment may be more interested in a COPD measure that is limited in its care setting inclusion and temporal criteria.

Temporal

Decisions about when to start or end a measurement period must be specified for each measure. Even when measure developers make the same or similar decisions about a measurement's clinical logic, they may not agree on the length of time specified for the unit for measurement (e.g., a 30-day versus 60-day episode of care for knee replacement surgery), which can result in a greater or lesser number of services being grouped in an otherwise similarly defined measure. Often, these temporal parameters are identified through clinical or evidence-based guidelines, expert opinion, or empirical data. For example, a measure may specify a diagnosis of low back pain with no evidence of a preceding diagnosis of low back pain for at least 12 months as the trigger of an acute low back pain resource use measure. During the measure's development, the developers may examine, along with experts in the treatment of low back pain, the frequency of related and unrelated services of low back pain. Based on expert input and the data, the developer may determine that for a commercial population (e.g., patients between the ages of 18 and 65 years), the measure's end date should be 45 days after the diagnosis of back pain that triggered the acute back pain episode. For chronic conditions, an approach some measure

developers take is to break chronic condition periods into year-long (often calendar year) segments. This allows for annual performance comparisons but introduces some distinct disadvantages. Specifically, this approach on its own does not account for the phase, or the point where a particular patient lies on the chronic disease continuum. For example, we can think of a chronic condition as having three large segments on the continuum: 1) onset, 2) treatment, and 3) resolution or end-of-life services. Each patient with a chronic condition has the onset of the disease, when they may be encountering the healthcare system but have not yet been diagnosed with the condition under measurement. The treatment phase includes secondary preventive services or the treatment of complications or flare-ups; and the resolution or end-of life phase includes services rendered at the end of the condition continuum, whether by resolution or death. In addition to the possible service-time truncation in the first- and last-year segments, it is reasonable to assume that chronic treatment periods are likely qualitatively different in the first, middle, and final years of the condition. Further, it is plausible that some physicians or providers will have proportionally more patients in any one of these phases. Therefore, resource measure users who specify chronic care measures that treat the first-, middle-, and final-year segments as homogenous raise methodological questions. Many resource use measures, including episode-based measures, include risk adjustment or stratification methodologies that may address this question. As a result of this measurement limitation, the approach (or lack thereof) for overcoming this issue needs to specified, transparent, and subject to evaluation.

Assigning and Triaging Claims

An important component of any measure specification is making decisions about which services to include in the measure's calculation. Once the clinical logic is determined, which identifies diagnostic or procedural events and groups services around them, decisions about how services or claims are assigned to the defined clinical logic must be made. In addition to the temporal rules established in the measure's construction, decisions about the assigning and triaging of services to the measure or measures must be determined, including how to manage different claims that provide information for the same event (especially those that result in an inflation of resource use amounts), when and how to map or feed claims from different sources into the same measure, or even when and which services trump other services. Some measure applications will assign one service to only one measure for each patient. While this may appear straightforward,

it requires complicated analytic functions, as many conditions overlap and no two patients are alike. Thus, the measurement approach must essentially triage each claim into the best measure for any given patient, with the flexibility that the best measure for one patient may not be the best for another based on that patient's underlying clinical condition profile. Other measurement approaches allow claims to be assigned to more than one service but then place limitations on any global or total resource use estimation. These decisions have implications for the validity of the measure and may be influenced by the type of resource use measure and the measurement effort perspective.

Identifying Units of Resource Use

As part of the measure construction, the units of health services or resource use units, must be identified and defined. The resource units of interest may vary depending on the type of resource use measure, the setting of care, or attribution and other decisions. For example, it may be of interest to measure emergency department (ED) visits for episodes of asthma care along with other units of resources; but for knee-replacement surgery, ED visits may not be of interest. Further, merely stating which units of service are of interest (e.g., pharmacy services) is insufficient; measures must define and provide clear and detailed instructions on how to identify a single health-service unit, including the relevant codes, modifiers, or approaches to identify the amount. For example, Current Procedural Terminology (CPT®) codes often are accompanied by modifier codes. These codes provide additional information and may signal an additional unit of service (e.g., the presence of two surgeons for one procedure). Unlike traditional quality measurement, one diagnosis or event in a single claim is often insufficient for resource use measures. Thus, while billing and payment systems may automatically track, account for, and often require the presence of all the necessary claim line information, measurement efforts that do not have the benefit of this experience or automated applications require this degree of specificity in the measure specification itself.

Module 4. Adjustments for Comparability

Whenever a measure is estimated, external factors can mingle and affect or confound the end result. Confounding occurs if an extraneous factor causes or influences the outcome (e.g., higher

resource use) *and* is associated with the exposure of interest (e.g., episode of diabetes). Administrative data sets may not contain or may have incomplete data on confounders, such as socioeconomic status, but measure developers often include steps to adjust the measure to increase comparability of results among providers, employers, and health plans. Risk adjustment is designed to reduce any negative or positive consequences associated with caring for patients of higher or lower health risk or propensity to require health services. Another type of *adjustment* is stratification, which is important where known disparities exist or where there is a need to expose differences in results so that stakeholders can take appropriate action. It is well known that prices vary substantially across the United States, within regions, and even within local markets. [21] As previously discussed, the perspective is critical in making decisions about the "who," "what," and "why." Thus, measure users may find more utility from one costing method than another.

Risk-Adjustment Approach

Risk adjustment is a corrective approach designed to reduce any negative or positive consequences associated with caring for patients of higher or lower health risk or propensity to require health services. If results are not risk adjusted, providers and health plans may have an incentive to attract healthier patients and avoid those who are sicker or require more complicated and extensive health services.[22,23] Risk-adjustment approaches often are defined as the process of adjusting payments to healthcare providers or health plans to account for the health status of the patients or members.[24] Thus, for comparative measurement purposes, applying a risk-adjustment method to a provider's or other entity's (e.g., health plan's) estimated resource use is meant to equalize or account for any differences in the composition of their panel or enrollees that would affect their resource use amounts. The use of diagnosis and pharmacy-based methods of health-risk assessment for profiling reflects the desire to provide equitable and appropriate comparisons. This is necessary because the health status of enrollees can vary significantly across health plans and healthcare providers. [25]

Medical diagnosis codes in administrative claims data often are used to assess health risk. Users of resource use measures often assess the extent to which a physician's or entity's total claims costs of services provided are greater or less than costs expected for those patients, given the

patients' demographic characteristics and health conditions.[26] The federal government and state agencies use medical diagnosis codes to adjust payments to the Medicare and Medicaid health plans, and even employers use diagnosis-based methods of risk assessment to analyze how employee contributions should vary by choice of provider or health plan.[27] Resource use measures, including episode-based measures, generally risk adjust as part of the steps to address differences in patients' characteristics and disease severity or stage. The module or phase of measure production at which the risk adjustment occurs may vary depending on the approach the measure developer selects as most appropriate for the construct of its measures. Risk adjustment within episode-based measures is different than per capita or population-based risk adjustment, which adjusts total spending per person for the person's overall risk. For example, when GAO used per capita measurement to explore differences in physicians' practice patterns, it adjusted risk using Diagnostic Cost Group (DCGs).[28] DCGs use beneficiary characteristics—age, sex, and Medicaid status—as well as diagnosis codes to assign each beneficiary a single health-risk score. Many episode-based measures build risk adjustment into the definition of the episode unit of measurement, which they accomplish by subsetting, or splitting out, condition groups into multiple categories so that initial comparisons can be made at a more granular level.

Risk adjustment approaches used in resource use or cost measures often are based on administrative and claims data only. The reliability and validity of such risk-adjustment approaches is influenced by the accuracy and completeness of the administrative and claims data. As discussed in the protocol section, steps must be taken to ensure the completeness and reasonableness of the data. Even after these steps are taken, there are concerns about the lack of clinical detail, which include important pathophysiological information that distinguish between conditions and complications.[29] Consequently, the validity of risk-adjustment systems that solely rely on administrative data has been challenged.[30] In the limitations section of this paper, we discuss more broadly the limitations of claims data that may lead to misclassification[31] and the need for measures to address and users to understand these limitations.

Stratification Approach

Arranging or separating resource use results by certain confounding patient or other relevant characteristics may be helpful to decision makers when important disparities exist. Stratification

of results can be used to aid decision makers' ability to take action on the results. In addition to exposing disparities, a measure may specify stratification of results within in a major clinical category (e.g., diabetes) by severity or other clinical differences. Balancing stratification and risk adjustment, which accounts for differences prior to the final estimation rather than separating results, is an important consideration that involves the perspective of the measurement effort.

Costing Methodology

Depending on the perspective, users of resource use measures may be interested in the count of services, the actual amount paid, or an approach that allows them to compare the use and intensity of health services while holding actual paid amounts constant (e.g., standardized prices).

Prices that purchasers pay for the same service vary substantially and for numerous reasons. Insurance plans negotiate different rate structures with the providers in their network and with purchasers. Plans that cover out-of-network services usually pay different rates to these providers. Even traditional Medicare's administrative pricing includes payment policies that introduce variation. For example, for the same discharge diagnosis, Medicare pays a rural community hospital less than it pays a major teaching hospital in an urban area for reasons such as differences in the local wage index, disproportionate share hospital classification, and indirect and direct graduate medical education payments.

Known measurement efforts, such as the CMS and MedPAC physician resource use measurement analyses using episode-based measures, use standardized payments, which remove variation in resource costs due to price variation. Approaches to determine the standard price for any given unit of service typically attempt to account for differences in the intensity among services. For example, an outpatient office visit with a surgical procedure service is a more intense service than an outpatient office visit with an evaluation and management service and would have a higher standard price attached to it, though both represent one outpatient office visit. Thus, applying standardized prices to the resource units compares variation in the amount *and* intensity of health services used and holds constant differences in local or negotiated prices.

Private insurance plans often use both standardized prices and the prices paid, depending on the question that is being asked. Since their overall costs are a result of negotiated prices with providers or of the benefit design, private insurers often include their prices paid in the total resource use measure so they can examine the impact of these negotiated rates and the benefit designs. Providers that negotiate high payment rates, therefore, may not look as efficient as providers that negotiate lower rates, unless they keep their resource use units low enough to offset the higher prices. Differences in coverage policies, i.e., benefit designs, also may influence the delivery of services and should be considered in the context of the measure results and comparative efforts.

Both standardized prices and actual prices paid provide valuable information. Comparing physicians' performance using standardized prices makes sense when the reasons for price differences among physicians are known and desired. For example, Congress decided that Medicare should pay a higher price for the same services to physicians who choose to practice in rural areas. The higher price is designed to improve access to physician services in those areas. If resource use measurement used actual prices in this instance—and did not standardize prices to neutralize the increased rural price offered as an incentive—then the exact same treatment pattern for an episode for a rural physician would be higher in cost than for an urban physician, making the rural physician appear less desirable based on a policy decision rather than on differences in the services delivered. Alternatively, comparing physicians' performance using actual prices paid makes sense when the reasons for price differences among physicians are not fully known or understood and may not be desired. For example, if a health plan pays one physician group differently than others in its network because of price negotiations, these price differences are not transparent to consumers and employers. These differences may be desirable if the physician groups differ on quality, geographic access, or similar characteristics, but they also may be based on other characteristics, such as market share. If the health plan were to measure resource use in this instance using standardized prices, then the results would obscure price differences and allow them to interpret the resource use results based on the type, frequency, and intensity of indicated services delivered. Whichever method is applied must be transparent to such a degree that decision makers can make relevant and appropriate inferences.

Module 5. Measure Reporting

Once the resource use measures have been estimated, users must consider and identify options concerning the reporting of measure results. This includes decisions about assigning or attributing the results to providers or entities, identifying the relevant peer group, estimating the benchmark or comparative values, setting and managing thresholds values, considering statistical matters, and sharing or reporting the results.

Attributing Resource Use Measures

One of the main goals of resource use measurement is to attribute the care provided as part of an episode of illness, the care of a population or event to a provider (e.g., physician, physician groups) or other entity (e.g., health plan) and, in combination with quality or health outcome performance, quantify how efficient their use of resources was for their patients. The breadth of a measure may influence the level of attribution that is valid. For narrower measures, such as those that are procedure specific, responsibility for the resources used for the procedure generally can be assigned to an individual physician—the physician who performed this procedure. For broader measures, such as per capita and per episode, more services—and therefore more physicians—are involved in each unit of measure, making attribution more of a challenge. Further, the type of delivery system the patient is exposed to may influence rules of attribution. In one extreme, with plans that assign patients to a primary care physician and explicitly hold the primary care physician accountable for the care the patient receives—such as HMOs that use gatekeepers—the attribution of a patient's resource use is relatively straightforward. In this case, attribution of the resource use measure is dictated by a policy decision. However, in other delivery systems in which patients may not have a gatekeeper or an assigned primary care physician and can refer themselves to specialists (e.g., in an open-access preferred provider organization), attribution is less straightforward and requires resource use measure users to make qualitative decisions about who they think *should* be responsible. Often these decisions may be supported by empirical data, or patterns in claims data may be used to attribute the resource use. For example, attribution may be assigned to a provider who contributed the most to the overall cost or to the provider who had the most evaluation and management visits during the measurement period.

A study conducted for CMS by Acumen, LLC, found that even for many broad, per episode measures, attribution can be straightforward.[32] The study reported that "generally speaking, care for a patient's episode is primarily influenced by just one provider, as indicated by a majority of episodes constructed from [Medicare Part B] claims submitted by a single provider."[33]

A key decision about how to attribute resource use units to a responsible entity is whether to attribute them to individual physicians, physician groups, larger entities (such as health systems and accountable care organizations), or multiple entities. Ideally, resource use measures should be flexible enough to allow attribution to these different types of entities. Rather than different resource use measures for different entities, measures should be harmonized so the same measure can be used across the continuum of entities. For example, individual physicians could be measured for the patients they see; these results could be aggregated for the groups to which each physician belongs and further rolled up for larger entities. Further, to ensure worthwhile public reporting, the level of entities to which responsibility is attributed should correlate with the different levels at which patients make choices. For example, measures aggregated to the health plan level would help patients or employers make plan enrollment decisions. Measures at the physician and other provider group level would help patients select providers for routine and unexpected care, and measures at the individual physician level would allow patients to opt for the provider best aligned with their preferences and needs. Concerns have been raised about the appropriateness of attributing responsibility for episodes to individual physicians. (See Section 5 for further discussion.) In a study for the Assistant Secretary for Planning and Evaluation (ASPE), RAND summarized the entities that have been used or proposed for attribution as follows:[34]

- **Individual physicians.** Commonly proposed criteria for assigning responsibility to an individual physician include a count of evaluation and management (E&M) visits or costs, physician specialty type, or some combination thereof.[35]
- **Individual physician—hospital care only.** One approach that has been tested is to attribute acute inpatient episodes to the attending physician for the hospitalization.
- **Hospitals.** Another strategy is to hold hospitals accountable for episodes of care that include a hospitalization in addition to physician services or services from other providers, such as skilled nursing facilities.[36,37]

- **Integrated delivery systems and physician group practices.** Existing integrated provider organizations are likely to have the greatest ability to assume responsibility for episodes of care because of the defined relationships between providers.[38,39,40,41]
- **Hospital medical staff.** This model would assign accountability for acute care episodes to the entire medical staff of a hospital (holding the hospital accountable as well).
- **Virtual groups.** Some have suggested the possibility of using virtual groups, that is, groups defined by geographic areas or other characteristics primarily for the purposes of episode-based performance measurement or payment.[42]

Another key decision about how to attribute resource use measures to physicians is whether to use single attribution (holding a single physician or entity responsible for the care provided) or multiple attribution (holding more than one physician or entity responsible for the care provided). Single attribution is designed to identify the decision maker, perhaps the primary care physician, and hold this individual responsible for all care rendered. Multiple attribution acknowledges that the decision maker, if there is one, has incomplete control over treatment by other physicians or specialists, even if the decision maker referred the patient to those other physicians, and acknowledges the truth that often f professional teams are responsible for the delivery of care to a patient.

MedPAC found that the choice of attribution method selected did not significantly affect physicians' resource use or efficiency scores. Physicians who appear to be efficient (or inefficient) under one attribution method generally appear to be efficient (or inefficient) under others. MedPAC concluded, therefore, that the choice among attribution methods probably comes down to a qualitative decision based on the program's policy goals.[43] For example, episode-based measure users who would like physicians to focus more on the effects of their referrals might select a single attribution method. Alternatively, users who wanted to trigger conversations among physicians caring for the same patient might select a multiple attribution method.

On the other hand, other researchers have found that the choice of attribution method did affect which physicians were assigned responsibility for episodes. RAND found significant variation in both the share of episodes that could be assigned to a physician and the level of agreement to

which a physician was held responsible.[44] For example, comparing the results of two different rules found that 50 percent of the episodes were assigned to different physicians. The study examined 13 attribution methods that differed on characteristics such as the basis of attribution (e.g., costs versus visits) and whether only one or multiple physicians were assigned to an episode. The Acumen study described above also found significant variation in the share of episodes that could be assigned to a physician using different attribution rules.[45]

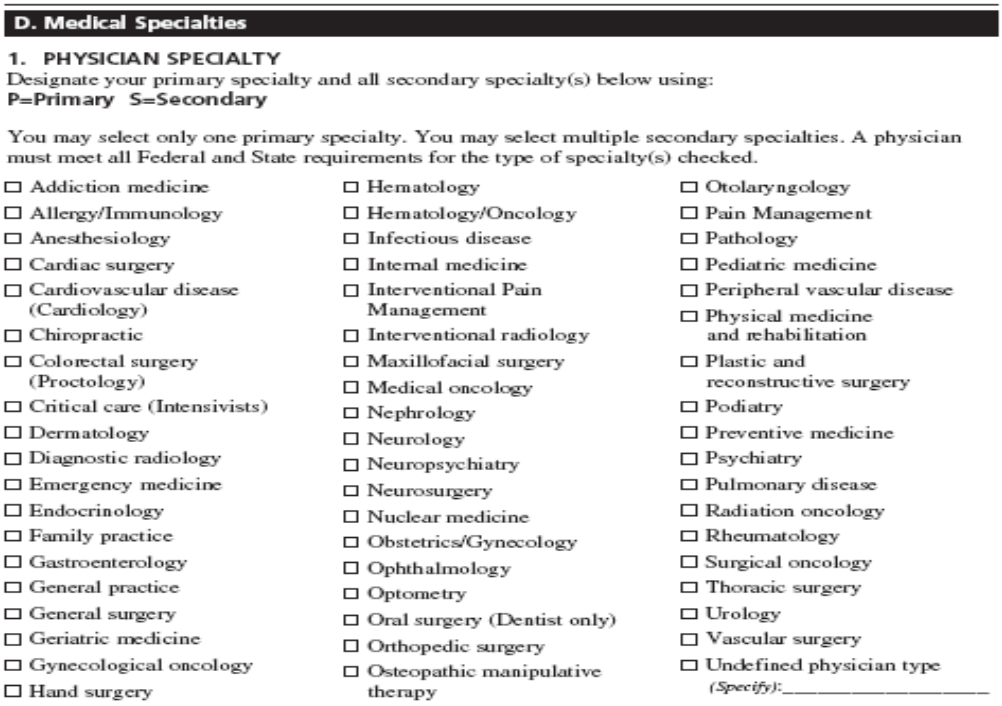
Like other resource use measures, per capita results are attributed to physicians or other entities and opt for either single or multiple attributions. By design, per capita measurement includes healthcare for individuals who may have none or multiple conditions and episodes. It also likely involves more physicians or entities per person than per episode measurement. Therefore, it may be preferable to use multiple rather than single attribution.

Peer Group Identification and Assignment

Once responsibility for the resource use measures has been attributed to physicians or other entities, the next steps are to assign a physician or entity to an appropriate peer group (e.g., cardiologists, thoracic surgeons, or Medicare Advantage plans) and compare them to a standard within their peer group. Unlike quality measures, which normally compare performance to an agreed-upon standard (e.g., providing flu vaccinations to a percentage of eligible patients) and direction for improvement (higher or lower performance is better), preferred resource use amounts often are not standardized, and it is not always clear if higher or lower resource use is preferable. Instead, resource use measure users often compare a physician's or entity's performance to the average performance of their peers. The two key characteristics of the physician peer group are most often medical specialty and geographic location. For example, a user could compare a cardiologist's resource use to the average resource use of other cardiologists in the same metropolitan statistical area (MSA). Alternatively, a user could compare a family medicine physician to all other primary care physicians in the state. While narrow peer groups may provide for fair comparisons, it may yield fewer observations and providers for comparison.

In practice, identifying a physician's specialty is difficult. Physicians often have more than one specialty, and discerning which specialist "hat" is most relevant to their encounter with any given patient is not always possible. In addition, payers and purchasers often have incomplete or imperfect data about their physicians' specialties. For example, Medicare requires physicians to indicate their primary and secondary specialties when they apply to become a participating Medicare physician (or at other specified times, such as when renewing participation).[46,47] However, Medicare does not use specialty designation for payment purposes, so it is not subject to audit, does not require physicians to update their specialty designation over time, and does not require physicians with multiple specialty designation to indicate which specialty "hat" they are wearing when providing services (see Exhibit 6).[48] When the specialty information is believed to not fully represent physician practicing specialty, measure users may opt to use claims data to examine the patterns of claims associated with a physician.

Exhibit 6. Physician Specialty Information Collected by Medicare



D. Medical Specialties

1. PHYSICIAN SPECIALTY

Designate your primary specialty and all secondary specialty(s) below using:
P=Primary S=Secondary

You may select only one primary specialty. You may select multiple secondary specialties. A physician must meet all Federal and State requirements for the type of specialty(s) checked.

- ☐ Addiction medicine
- ☐ Allergy/Immunology
- ☐ Anesthesiology
- ☐ Cardiac surgery
- ☐ Cardiovascular disease (Cardiology)
- ☐ Chiropractic
- ☐ Colorectal surgery (Proctology)
- ☐ Critical care (Intensivists)
- ☐ Dermatology
- ☐ Diagnostic radiology
- ☐ Emergency medicine
- ☐ Endocrinology
- ☐ Family practice
- ☐ Gastroenterology
- ☐ General practice
- ☐ General surgery
- ☐ Geriatric medicine
- ☐ Gynecological oncology
- ☐ Hand surgery
- ☐ Hematology
- ☐ Hematology/Oncology
- ☐ Infectious disease
- ☐ Internal medicine
- ☐ Interventional Pain Management
- ☐ Interventional radiology
- ☐ Maxillofacial surgery
- ☐ Medical oncology
- ☐ Nephrology
- ☐ Neurology
- ☐ Neuropsychiatry
- ☐ Neurosurgery
- ☐ Nuclear medicine
- ☐ Obstetrics/Gynecology
- ☐ Ophthalmology
- ☐ Optometry
- ☐ Oral surgery (Dentist only)
- ☐ Orthopedic surgery
- ☐ Osteopathic manipulative therapy
- ☐ Otolaryngology
- ☐ Pain Management
- ☐ Pathology
- ☐ Pediatric medicine
- ☐ Peripheral vascular disease
- ☐ Physical medicine and rehabilitation
- ☐ Plastic and reconstructive surgery
- ☐ Podiatry
- ☐ Preventive medicine
- ☐ Psychiatry
- ☐ Pulmonary disease
- ☐ Radiation oncology
- ☐ Rheumatology
- ☐ Surgical oncology
- ☐ Thoracic surgery
- ☐ Urology
- ☐ Vascular surgery
- ☐ Undefined physician type *(Specify):*_____

Source: Medicare Enrollment Application: Physicians and Non-Physician Practitioners. Available at http://www.cms.gov/cmsforms/downloads/cms855i.pdf. Last accessed August 2010.

The key characteristics of a provider or entity peer group may include specialty (e.g., oncologist), type of care setting (e.g., hospitals), product or product line (e.g., commercial HMO), and geographic location. For example, a user could compare commercial HMO in a specific metropolitan statistical area (MSA) to the average resource use of other commercial HMO in the same MSA.

Calculating Comparisons

After the comparison peer groups are selected, a user of resource use measures can use these groupings to estimate resource use values for each peer group. The estimations, typically the mean amount, are used to compare performance within the relevant peer group. These comparisons are a key difference between resource use and quality measurement. Quality measures generally use a specified benchmark, such as blood pressure control for patients with hypertension based on clinical evidence. Given the lack of evidence of the appropriate mix of resources, resource use measurement usually compares performance among peers. While there are different approaches among measure developers, one approach is to capture the resource use value for each resource use measure attributed to a physician or entity (typically termed the "observed" amount) and divide it by the average resource use within the identified peer group (typically termed the "expected" amount, i.e., the amount of resource use expected if the physician were performing at the mean). This ratio is called an observed-to-expected (O/E) ratio, where values above 1.00 indicate more resource used than expected and below than 1.00 indicate less resources used than expected. More sophisticated comparisons, such as multilevel regression and Monte Carlo simulation, also are used.[49]

A typical and straightforward approach to estimate O/E results for a physician or entity among multiple resource use measures is to summarize each measure's observed resource use amounts and expected amounts attributed to the provider or entity and calculate a total observed and total expected amount for that provider —this allows for the estimation of a global O/E result for each provider or entity (see Exhibit 7). This method essentially weights each measure result by their total observed and expected costs. It is critical to consider and understand the implications of any approach. For example, using the same data, the average of each measure's O/E ratio provides a

strikingly different picture of this provider's performance—from using more resource than expected in the first approach (1.20>1.00) to using less than expected in this second approach (0.94<1.00).

Exhibit 7: Estimating Global O/E Results-different approaches yield different results

RU Measure	Observed $	Expected $	O/E
A	$120	$180	0.67
B	$45	$110	0.41
C	$6,000	$4,523	1.33
D	$389	$354	1.1
E	$258	$267	0.97
F	$7,890	$6,782	1.16
Total	$14,702	$12,216	

Global O/Es with different results:

Total O/Total E = 1.20

Mean O/E = 0.94

These comparisons are usually performed only for providers or entities that have a minimum number of resource use measures attributed to them (e.g., 20, 30, or more). Some users also require that providers have a minimum number of a certain type of resource use measure rather than just a minimum of all resource use measures (e.g., at least 10 or 30 episodes for a given condition). Alternatively, some users rely on statistical tests rather than rely on a minimum threshold of observations.[50]

In estimating a physician's or entity's global O/E, it is important to consider whether a service is assigned to only one measure or to multiple measures. The answer has implications for physicians (or any entity, for that matter) because if a service's resource use or cost is being assigned to multiple resource use measures, a global result that does not account for this will be

inflated. Approaches to deal with this situation include algorithms that determine to which measure any one service is assigned based on patient experiences, or by prorating individual services among the measures it is assigned, so that in the end the global estimate does not exceed the true total cost. Other developers may not provide an approach to estimate a global resource use amount and will instruct users to examine and compare resource use within the specified measures.

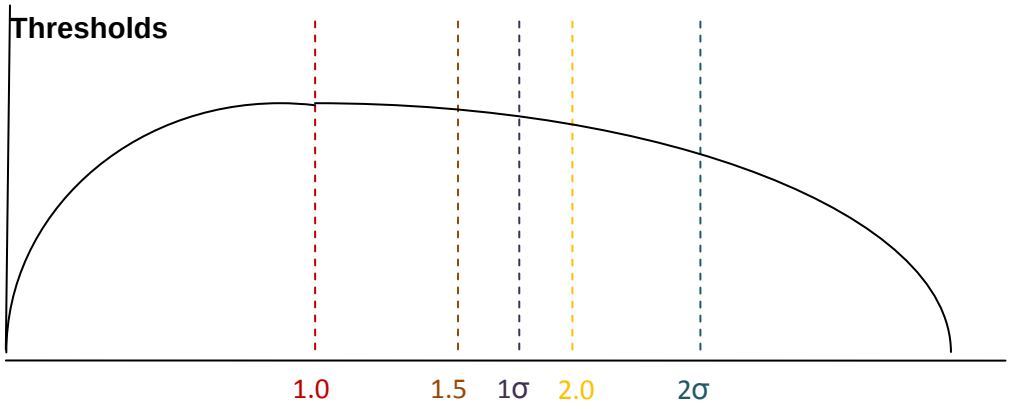
For per capita resource use measures, once the spending per person is attributed, these values need to be rolled up to an average or composite for each entity. This allows comparisons of physician to physician, health plan to health plan, etc. However, one cannot simply compare each entity's total average spending per person to the peer group entities' total average spending per person because the patients seen by each will differ. To compare physicians appropriately on a per capita resource use measurement basis, some form of case mix adjustment is required. An option discussed by the General Accountability Office (GAO) sorts patients into risk categories and compares each physician's share of patients with high resource use, compared to other patients in the same risk category.[51]

Setting Thresholds

Following the estimation of a resource use measure's value, users must determine whether to apply thresholds or remove outliers. Threshold determinations can include discarding or "Windsorizing" (truncating) and can be applied at the claim-line level, measure estimate level, or physician or entity level; applying thresholds or removing outliers provides more context for the values. Outliers can be the result of inappropriate treatment, rare or extremely complicated cases, or coding error. Users often do not completely discard outliers, but rather examine them separately. Claim-level thresholds typically are executed during the data protocol phase. Once the resource use values are estimated, these thresholds typically are determined either by examining the results empirically or for some policy reason. For example, a user may opt to flag and examine separately as outlier all physicians with O/E ratios greater than 1.5 or 2, or those physicians who are 1 or 2 standard deviations (σ) outside the mean (see Exhibit 8). Other users may opt to report on all physicians and choose to throw out or truncate individual resource use measure estimates (e.g., an episode) if it is above or below a determined threshold—for example,

an individual measure estimate that is 1 or 2 standard deviations outside the mean of all the related resource use measures within a given peer group.

Exhibit 8: Illustrative Distribution of Observed-to-Expected Ratios and Possible Thresholds



Providing Detailed Feedback

In a 2002 Society of Actuaries report, results were analyzed using three truncating scenarios: 1) truncate claims at $50,000; 2) truncate claims at $100,000; and 3) do not truncate. The purpose of truncation is to provide more stability in the results when analyzing predictive accuracy. Further, the report stated that large claims for a given person generally are not predictable. Accordingly, some researchers argue that they should be removed or limited when the analysis is performed.[52]

After all of the analytic steps are completed, users of resource use measures must decide which analytic results to include in any feedback or public reports. Often episode-based measures provide much more detailed analytic results than just total resource use by episode. They break down those values by type of service, setting, and other characteristics. For example, a user could show total emergency department usage, rate of generic drug prescribing, and number of physician office visits. For a report example, see CMS's **Prototype Medicare Resource Utilization Report Based on Episode Groupers**.

Reporting with Descriptive Statistics

Depending on the perspective and whether the measure will be used for internal improvement or public reporting, decisions about which statistics must accompany the resource use measure results are critical. For example, confidence intervals used around a resource use estimate provide certainty of the estimate itself. Other statistics may be used, but they should be selected with a strong consideration for their interpretability by all relevant stakeholders and audiences. Similarly, decisions need to be made about which estimates and the degree of detail that results should be presented in feedback reports or public reports. In general, more detailed, actionable feedback requires that measures capture necessary information, such as spending by type or service.

These types of analytic results can provide the detailed information necessary to make feedback actionable for all stakeholders. However, a number of options will need to be considered provide reports with maximum actionability without information overload.

Section 5: Limitations to Resource Use Measurement

As previously noted, NQF's evaluation criteria require that measures demonstrate importance to measure and report, scientific acceptability of measure properties, usability, and feasibility. To meet the criterion of scientific acceptability, for example, a standard must reliably and validly measure what it is intended to evaluate. If the standard is not measuring what it is intended to measure, it cannot facilitate improvements in healthcare systems, and already limited resources for measuring and reporting are potentially wasted. During the NQF submission and review process, the measure developer must provide evidence demonstrating reliability and validity of the measure. The analysis must demonstrate that methods for scoring and analyzing the specified measure allow for identifying statistically significant and practically or clinically meaningful differences in performance. While not all sources of measurement bias can be eliminated, an attempt should be made to provide details necessary to minimize common sources of bias for resource use measurement.

Claims and Other Administrative Data Limitations

Most resource use measures rely primarily on claims and other administrative data (e.g., enrollment data), and the limitations of these data sources can have an impact on the measure. Administrative data are a product of healthcare service delivery and reimbursement and provide a minimum amount of patient and provider information. Administrative data are often used because they are readily available, inexpensive to acquire, computer readable, and typically encompass large populations.[53] However, gaps and incomplete clinical information compromise the ability to use administrative data for measurement; [54] the content of administrative data is often limited and may lack clinical details. The concordance between the medical record and administrative data varies[55] and may vary depending on the condition or setting of care.[56,57] Further, even when the administrative data are highly concordant with the medical record, some systems do not maintain all the diagnostic information submitted on the claim—thus providing a less-than-complete picture that may bias measurement results.[58] This complicates the ability of resource use measures to assign (or group) claims into homogenous groupings or clinical episodes of care by diagnosis or to assess patient severity or risk levels.

To complicate the use of claims data further, different provider types' claims offer different opportunities to provide granular, complete, or disaggregated services. Physician professional claims, for instance, provide line-item detail on specific services, whereas facility-based claims often bundle or miss services. Acumen found that among institutional claims there was substantial variation in the amount of detail provided and captured.[59] To address this type of variation, some resource use measures split up claims or services and assign facility-based services to different episodes of care, while others will require them to be assigned entirely to one episode. Measure specifications also may include instructions on how to manage incomplete claims, zero-dollar claims, and claims from ancillary settings. Algorithms also may include approaches to ensure the diagnosis under consideration is valid by requiring two instances of the same diagnosis within a 12-month period. Thus, strategies to address some of these issues must be provided to users of resource use measures with the rationale and implications of steps taken to address issues with claims data.

Many of these measurement limitations reflect challenges associated with using administrative and claims data that initially were primarily constructed to inform payment. They are more limitations of the claims data themselves than of the measurement methodology. Approaches to assign or split claims into homogenous clusters or episodes ideally *should* be included in the resource use measure methodology. While the future of electronic clinical information is promising, failing to understand or address the current limitations of the administrative data may lead to misclassification.[60] Also, claims data could be refined to be more consistent across provider types and to include more clinical information useful for measurement, such as lab values. More complete, granular, and consistent claims and administrative data are an essential foundation for payers to become more sophisticated, value-based purchasers of healthcare services in emerging payment reform models, such as ACOs and medical homes.

Small Sample Sizes

Having an adequate sample size for any type of measurement is critical—the goal is to have a sample size that is large enough to minimize the effect of chance and that supports adequately precise results. Determining how large a sample should be is not easy. The answer depends on the tolerance for inaccurate results and the expected confidence in the results. Users of resource

use measures, including those that are episode based, often note potential small sample sizes, which mean there may be too few observations to produce statistically valid measurement for comparisons. When this problem occurs it is often ascribed to the availability of small or limited datasets, measures designed to have high specificity (i.e., false positives have been removed), or measures assess outcomes in areas with few occurrences in the population. This issue is exacerbated when users divvy up limited observations among individual physicians, rather than large physician groups or larger entities that benefit from a larger population from which to measure, in an attempt to hold those physicians accountable for the services they deliver. Typically, as sample size increases, the confidence in the measurement result increases, as does the ability to detect statistical differences. Assessing the practical or clinical meaningfulness of these differences is critical, however.

It is important to note that small sample size is only one characteristic that determines the level of confidence in a physician's or entity's score being non-random. The range of the results within the physician's or entity's peer group also determines how confidently one can determine whether a physician or entity differs from his or her peers. Further, recent studies examined not just the effects of sample size, but also the mix of episodes and risk adjustment and found they all contribute to the reliability of results.[61]

Some argue that the most expedient way to address concerns about measurement precision stemming from small sample sizes is to measure not at the individual physician level but at levels with more patients, such as physicians' groups or ACOs. Often a priori analyses can estimate the likely size of a sample for a measure from a given population, e.g., from a panel of patients or health plan when the prevalence or incidence of occurrence in the population is known. For example, colorectal cancer screening is conducted at a rate of 168.2 per 1,000 member years, which within a 12-month period would yield 1,682 observations for measurement for a health plan with 10,000 members. For a physician panel of 1,000, however, the same period would yield only 168 events. Conversely, heart failure, a very serious and costly condition, has a low prevalence of 0.6 per 1,000 member years, yielding 6 patients for measurement in the health plan and not even 1 full patient (0.6) for the physician panel of 1,000 patients. Recently, one pay-for-performance program reported that not only are claims data often incomplete or poorly coded,

but even large physician groups often have too few patients experiencing most types of episodes to permit statistically valid measurement for public reporting and incentive payment.[62]

However, resource use measurement at the individual physician level should not be ruled out because many physicians are in solo or small practices, and because treatment and economic decisions still occur at the physician level. Ideally, measures should use individual physicians as the basic building block of resource use measurement but be capable of aggregating these measures in multiple ways, such as by physician group practice and by accountable care entities. This flexibility is critical to allowing users to assess different levels of the health system and to adjust who and for what purpose they are measuring based on their perspective. It also permits users to measure the nearly 40 percent of physicians who continue to practice as solo practitioners[63] and will help to avoid problems in markets where group practices are so large and command so much market share that there are too few peers for comparison.

Furthermore, because NQF-endorsed measures are intended to be useful for both public reporting and quality improvement, measure developers and users should strive to produce results at a level that decision-makers (e.g., individuals, beneficiaries, providers, or health plans) can use and offer flexibility for tailored use. For example, a beneficiary who receives his primary care at a small family medical practice where his appointments might be with any of the physicians in the practice would most likely want to consider the performance of the group as a whole. On the other hand, the same beneficiary could seek cardiology care at a large multispecialty group practice with numerous satellite offices. If the beneficiary planned to visit only one of those offices and use only cardiology care, more aggregated performance measures would not be as helpful.

"Black Box" Methodology

Critics of commercially available episode-based resource use measures have long argued that they have relied on "black box" methodology that is proprietary and therefore not transparent. This criticism, in part, has motivated the creation of grant-funded, episode-based measures such as Prometheus and ABMS. However, even commercially available episode-based resource use measures have become much more transparent. In March 2009, Ingenix, Inc., released its ETG

measurement methodology for public review and comment.[64] In June 2009, Thomson Reuters also released its MEG methodology.[65]

The hallmark of NQF's endorsement process is transparency. Even if developers maintain charges to users for publicly reporting their performance measures, the review committees must have full and complete access to all measure logic and coding. The cost associated with the use of the measure for improvement or public reporting is considered under NQF's evaluation criteria of feasibility.

Section 6: Summary of NQF Evaluation Criteria for Measures of Resource Use

A critical component of this project is to inform the review and adaptation of the NQF evaluation criteria for evaluating resource use measures. Appendix B, *Proposed Resource Use Evaluation Criteria Comparison Table*, was developed based on this paper and the NQF Resource Use Steering Committee's guidance. This section focuses on the description of resource use measures, lays out principles for evaluating resource use measures, and offers the rationale for the proposed subcriteria for evaluating resource use measures.

Resource Use Measure Description

As with quality measures, the careful design and evaluation of resource use measures is imperative. Resource use measures introduce unique issues, including how to describe the measures, the reliability and validity of the measure, the rules of attribution, and the methods used to estimate the resource use measure values, including risk adjustment. A general description of a resource use measure listed below should allow evaluators and users to assess quickly what is being measured. Acknowledging NQF's approach to describing quality measures as having a denominator and numerator, the following is proposed:

- **Description of Measure:** the measurement focus, target population, and type of final score (e.g., the observed-to-expected ratio of outpatient services for an episode of asthma for children between 5 and 18 years of age among primary care physicians). The measure reports the observed value and expected value, along with the ratio result. The description must specify the type of measure (e.g., per patient, per episode), clinical or target area of measurement (e.g., asthma or all women), the metric result, final score, and comparison peer groups.
- **Resource Units:** the resource utilization of interest, including the service categories, and its measurement value. This includes details about which resources are being measured, how it is being estimated (e.g., the costing method), and comparison estimates (e.g., the mean performance among the peer group).
- **Measurement Standard:** This portion of the resource use measure is analogous to the denominator of a quality measure. It is the standard to which the resource units will be

applied (e.g. pharmacy costs (resource units)/ hip surgery patient (measurement standard)). It can also be considered the target population, event or measure of analysis (e.g., an episode of asthma) that is defined and specified.

Note: The descriptions are not the measure specifications, but rather they describe in words the purpose of the specifications. Specifications include temporal criteria as well as diagnostic, procedure, place of setting, and other relevant codes that allow for the application of the measure algorithms necessary to calculate the resource use measure in full.

There is no specific classification of resource use measures that parallels those used for the three types of individual quality measures (i.e., structure, process, and outcome). Rather, there is a spectrum of resource use measurement types, spanning from per capita (population based), to episode based to procedure specific. The proposed resource use evaluation criteria were created with this spectrum in mind and are intended to include the appropriate evaluation components for all types of resource use measures.

Resource Use Measure Evaluation Principles

Before identifying the specific evaluation criteria for resource use measures, the Steering Committee articulated some general principles that underlay the evaluation of resource use measures and the goals of this project. While resource use measures present with fundamental differences, these principles should apply across all types and approaches.

Principles for Resource Use Measure Evaluation

1. Efficiency is one of the IOM five quality aims; it is a function of resource use and health outcomes:

   Efficiency = fx(resource use, health outcomes)

2. Resource use measures are the amount of resources used per population, episode, or procedure.

3. Resource use measures are an important building block to measures of efficiency of care; future measurement efforts should integrate and explicitly incorporate measures of quality, health outcomes or appropriateness.
4. The justification for and intended purpose of resource use measures is to examine, understand, and ultimately reduce unnecessary costs in care.
5. There is a continuum of resource use measures; all types under consideration for endorsement must meet NQF evaluation criteria for such measures.
6. The resource use measure specification and calculation must be explicitly stated and transparent so the approach can be deconstructed and implemented in a standard manner.
7. Comprehensive measures are preferable, even if combining multiple service categories into one resource use estimate increases complexity; using methodologically sound methods is of paramount importance.
8. The final resource use measure or result should be simple and readily interpretable by all stakeholders.
9. Methods for combining the component scores influence the interpretation of the measure results and must be justified (e.g., averaging across all component scores may obscure low or high scores of individual components).
10. While resource use measure developers may have fundamental differences in approach, these principles should apply across all types and approaches.
11. NQF considers transparency as key to ensuring the intended audiences understand the results and can use them for decision making. Resource use measures are often highly complex, with lengthy algorithm decision trees that can make clarity difficult when some approaches may be only partially transparent to the user.

Importance to Measure and Report

The importance criterion is focused on evaluating the extent to which to the measure focus is important to making significant gains in healthcare quality and improving health outcomes for high-impact aspects of healthcare where there is variation in or overall poor performance.[66] Rather than gains in quality or health outcomes, in the context of resource use measures, importance will be judged on a measure's significant contributions toward understanding healthcare costs for a high-impact aspect of healthcare where there is *unexplained* variation in or

a demonstrated high-impact aspect of healthcare. In addition to the existing criterion (1a), measurement areas should focus on the evaluation and alignment with the National Priorities Partnership Goal and demonstrate high-impact aspects of healthcare. The importance of resource use measures will be further evaluated for evidence of variation in costs and provider performance associated with the condition or episode. In refining the criteria for resource use measures, language was expanded to indicate that the opportunity for improvement in the context of resource use measures can be demonstrated largely with data showing considerable unexplained variation in costs. Further, broad comprehensive measures of resource use are preferable, and the health services (or units of resource use) selected for measurement should be conceptually coherent. Omitting key resources indicated by the population, condition, episode, or event could lead to an incomplete measure of resource use and have implications for interpretation, attribution, and implementation.

Scientific Acceptability of Measure Properties

Evaluating scientific acceptability includes evaluating the specifications, which must be precise and complete, as well as the reliability and validity of the measure, demonstrated by testing these properties. Thus, resource use measures will be evaluated based on the extent to which the measures, as specified, produce consistent (reliable) and accurate (valid) results about the cost or resources used to deliver care. While most of the subcriteria for quality measures also apply to resource use measures, the evaluation of scientific acceptability for resource use measures requires reviewing the measure specifications and testing requirements specific to these types of measures. Like all measures submitted to NQF for endorsement consideration, well-defined and precise specifications for resource use measures must be complete. Missing or incomplete specifications or testing results must be clearly justified, with a rationale and implications provided by the measure developer, at the time of submission. For example, a resource use measure may not include a separate risk adjustment approach because it is imbedded in the clinical and construction logic—the submission must clearly explain this rationale and any implications.

In addition to the basic measure descriptors, the developer will be expected to describe in detail the steps and decisions made during the development and specification of the measure within each of the five modules of resource use measure: 1) data protocol, 2) measure clinical logic, 3) measure construction logic, 4) adjustments for comparability, and 5) measure reporting. For the fifth module, the committee is considering requesting guiding principles, rather than specifications, to meet this module requirement, demonstrating well-thought-out and tested methods for reporting out and using resource use measure results that are made available to users of the resource use measure under review.

The second component to evaluating a measure's scientific acceptability is determining whether it is reliable and valid. This is demonstrated through testing results. Measure testing findings proving the measure's reliability (i.e., the demonstrated ability that the measure results are repeatable and produce the same results for the same population in the same time period) will be requested for each of the five modules. Developers will be tasked with selecting the testing method that best fits their measures and submitting the results.

Validity testing findings, which establish the credibility of the measure, will be required for the clinical logic, construction logic, adjustment for comparability, and reporting modules. This criterion will be evaluated in conjunction with the stated purpose and intended use of the measure to determine if it is accurately measuring what it should. Validity of resource use measures can be assessed using face, criterion, content, or construct validity methods. While each of these approaches may be acceptable, it is the developer's decision which method will be used to demonstrate the submitted measure's validity. Validity testing demonstrates that the measure reflects the resources used for a particular condition, event, or population and adequately distinguishes high and low resource use. If face validity is the only validity addressed, it is systematically assessed. Examples of validity testing include, but are not limited to: 1) determining if measure scores adequately distinguish between providers known to have high or low resource use assessed by another valid method; 2) correlation of measure scores with another valid indicator of resource use for the specific topic; 3) ability of measure scores to predict scores on some other related valid measure; and 4) content validity for multiple-item resource use measures.

The final testing category is for measure exclusions. Because exclusions occur at various steps in the process of the measure construction and specification, each of these steps should be tested for sensitivity and demonstrated with empirical data supporting the decisions made for exclusions within the steps for data preparation, clinical logic and construction, and profiling (e.g., determination of thresholds and outliers).

Usability

As with quality measures, a resource use measure's usability is based on whether the intended audiences find the information the measure produces to be meaningful, understandable, and useful both for public reporting and internal improvement.[67] Because a resource use measure's output provides little information about whether it is the *right* amount, the results of a measure must be put into context with benchmarks and are most useful when presented relative to quality. The link to quality is key to determining an input's value. For this reason, the Steering Committee agreed that resource use measures that are used alongside quality or health outcome measures would be given preference over those that are not. Resource use measures that are used this way are one step closer to the goal of understanding efficiency and the value of care provided. As part of these criteria, measure developers or stewards will be asked to provide a list of NQF-endorsed measures known to be reported along with the submitted resource use measure.

Feasibility

The feasibility criterion requires that the developer demonstrate the extent to which the required data are accessible, retrievable without undue burden, and able to be implemented for internal improvement and public reporting. While many resource use measures use administrative data to determine inputs, making data accessible and feasible to collect, they may be very complex and require programming and risk-adjustment methods to estimate. Further, resource use measures often have detailed algorithms used to describe the clinical logic and grouping of clinical conditions or events. For users of resource use measures with limited resources, this presents a challenge to implementing the measures. The cost associated

with the use of measures for public reporting or quality improvement is considered as part of the criteria.

1578

1. McGlynn EA, *Identifying, Categorizing, and Evaluating Health Care Efficiency Measures*, Rockville, MD: Agency for Healthcare research and Quality; 2008.

2. Institute of Medicine (IOM), *Crossing the Quality Chasm: A New Health System for the 21st Century*, Washington, DC: National Academies Press; 2001.

3. McGlynn EA.

4. IOM.

5. These terms are adopted from AQA Principles of Efficiency Measures. Available at http://www.aqaalliance.org/files/PrinciplesofEfficiencyMeasurement.pdf Last accessed August 2010.

6. These definitions do not adequately capture the concept of health outcomes--efficiency examine the cost of care for a given set of health outcomes; however, given the challenges of associating outcomes with healthcare interventions, assessing the quality of care is a way to operationalize efficiency.

7. Thomas JW, Grazier KL, Ward K, Comparing accuracy of risk-adjustment methodologies used in economic profiling of physicians, *Inquiry* 2004;41(2):218-231. Thomas and Grazier compared the predictive accuracy and consistency of methods used for provider profiling, finding that while there was much consistency overall, different software identified different providers as relatively high cost or low cost.

8. MaCurdy T, Theobald N, Kerwin J, et al., *Prototype Medicare Resource Utilization Report Based on Episode Groupers*, Burlingame, CA: Acumen; 2008. Available at www.cms.gov/reports/downloads/MaCurdy2.pdf . Last accessed August 2010.

9. McGlynn EA.

10. Figure adapted from McGlynn, EA, pg. 16.

11. The Dartmouth Atlas of Health Care. Available at www.dartmouthatlas.org/. Last accessed August 2010.

12. National Quality Forum (NQF), *Measurement Framework: Evaluating Efficiency Across Patient-Focused Episodes of Care*, Washington, DC: NQF; 2010. Available at www.qualityforum.org/Publications/2010/01/Measurement_Framework__Evaluating_Efficiency_Across_Patient-Focused_Episodes_of_Care.aspx. Last accessed September 2010.

13. NQF, *Measurement Framework: Evaluating Efficiency Across Patient-Focused Episodes of Care*.

14. Medicare Payment Advisory Commission (MedPAC), *Report to the Congress: Improving Incentives in the Medicare Program*, Washington, DC: MedPAC; 2009.

15. Ibid.

16. These modules and the analytic steps within them are generally, though not strictly, in sequential order. They are grouped by their purpose, where different measures may specify an analytic function within another module. For example, a measure may use its clinical logic to specify the risk adjustment.

17. Chapman AD, *Principles and Methods of Data Cleanin Primary Species and Species-Occurrence Data*, version 1.0, Report for the Global Biodiversity Information Facility, Copenhagen; 2005. Available at www2.gbif.org/DataCleaning.pdf. Last accessed August 2010.

18. Ibid.

19. Ibid.

20. Thomas JW, Ward K, Economic profiling of physician specialists: use of outlier treatment and episode attribution rules, *Inquiry,* 2006;43(3):271-282.

21. Government Accountability Office (GAO), *Federal Employees Health Benefits program: Competition and Other Factors Linked to Wide Variation in Health Care Prices,* GAO-05-856, Washington, DC: GAO; 2005

22. Newhouse JP, Reimbursing health plans and health providers: selection versus efficiency in production, *Journal of Economic Literature*, 1996;34:1236-1263.

23. Altman D, Cutler DM, Zeckhauser RJ, Adverse selection and adverse retention, *American Economic Review, Papers and Proceedings*, 1998;88:122-126.

24. Cummings RB, Knutson D, Cameron BA, et al., A Comparative Analysis of Claims-Based Methods of Health Plan Risk Assessment for Commercial Populations, Society of Actuaries; 2002. Millman USA, Inc., Minneapolis. Available at www.soa.org/files/pdf/_asset_id=2583046.pdf. Last accessed August 2010.

25. Ibid.

26. Thomas JW, Grazier KL, Ward K, Economic profiling of primary care physicians: consistency among risk-adjusted measures, *Health Services Research,* 2004;39(4):985-1003.

27. Cummings RB, Knutson D, Cameron BA, et al.

28. GAO, *Medicare: Focus on Physician Practice Patterns Can Lead to Greater Program Efficiency*, GAO-07-307, Washington, DC: GAO; 2007.

29. Pine M, Jordan HS, et al., Enhancement of claims data to improve risk adjustment of hospital mortality, *JA MA*, 2007;297(1):71–76.

30. Ibid.

31. Shahian DM, Silverstein T, Lovett AF, et al.

32. MaCurdy T, Theobald N, Kerwin J, et al.

33. Ibid.

34. Damberg CL, Sorbero ME, Hussey PS, et al., *Exploring Episode-Based Approaches for Medicare Performance Measurement, Accountability, and Payment*, Washington, DC: RAND; 2009. Available at http://aspe.hhs.gov/health/reports/09/mcperform/report.pdf. Last accessed August 2010.

35. California Cooperative Healthcare Reporting Initiative (CCHPI), *California Physician Performance Initiative: Methodology for Physician Performance Scoring*, San Francisco: CCHPI; 2008. Available at www.cchri.org/programs/documents/CPPI_Methods_Oct2008.pdf. Last accessed August 2010.

36. Jencks SF, Dobson A, Strategies for reforming Medicare's physician payments: physician diagnosis-related groups and other approaches, *N Engl J Med*, 1985;312(23):1492–1499.

37. Welch WP, Prospective payment to medical staffs: a proposal, *Health Affairs* 1989;8(1):34–49.

38. Davis K, Guterman S, Rewarding excellence and efficiency in Medicare payments, *Milbank Quarterly* 2007;85(3):449-468.

39. MedPAC, *Bundled Payment for Services Around a Hospitalization*, Medicare Payment Advisory Commission Public Meeting, Oct. 3, 2007. Available at http://medpac.gov/transcripts/1003-04MedPAC.final.pdf. Last accessed February 2008.

40. MedPAC, *Bundling Payments in the IPPS*, MedPAC Public Meeting, Jan. 9, 2007. Available at http://medpac.gov/meeting_search.cfm?SelectedDate=2007-01-09%2000:00:00.0. Last accessed February 2008.

41. MedPAC, *Moving Toward Bundling Payment Around Hospitalizations*, MedPAC Public Meeting, November 8, 2007. Available at http://medpac.gov/transcripts/1108-09Medpac%20final.pdf . Last accessed February 2008.

42. Davis and Guterman.

43. MedPAC, *Report to the Congress: Improving Incentives in the Medicare Program*, Washington, DC: MedPAC; 2009.

44. Mehrotra A, et al., *Methodological Issues in Measuring Physician-Level Quality and Efficiency*, Academy Health Conference, Orlando, FL, June 5, 2007. Available at www.academyhealth.org/files/2007/tuesday/southernhemisphere1/mehrotraa.pdf. Last accessed September 2010.

45. MaCurdy T, Theobald N, Kerwin J, et al.

46. Centers for Medicare & Medicaid Services (CMS), Medicare Enrollment Application: Physicians and Non-Physician Practitioners. Available at www.cms.gov/cmsforms/downloads/cms855i.pdf. Last accessed August 2010.

47. CMS 1500 Health Insurance Claim Form. Available at www.cms.gov/. Last accessed August 2010.

48. MedPAC, *Assessing Alternatives to the Sustainable Growth Rate System,* Washington, DC: MedPAC; 2007, p. 79.

49. Houchens RL, McCracken S, Marder W, et al., *The Use of an Episode Grouper for Physician Profiling in Medicare: A Preliminary Investigation*, Washington, DC: MedPAC; 2009. Available at www.medpac.gov/documents/Jun09_EpisodeGrouperStability_CONTRACTOR_JP.pdf. Last accessed August 2010.

50. Adams JL, Mehrotra J, Williams T, et al., Physician cost profiling–reliability and risk of misclassification, *N Eng J Med*, 2010;362:11.

51. GAO, *Medicare: Focus on Physician Practice Patterns Can Lead to Greater Program Efficiency*, GAO-07-307, Washington, DC: GAO; 2007. Available at www.gao.gov/new.items/d07307.pdf . Last accessed August 2010.

52. Cummings RB, Knutson D, Cameron BA, et al., *A Comparative Analysis of Claims-Based Methods of Health Plan Risk Assessment for Commercial Populations*, 2002; Society of Actuaries. Available at www.soa.org/files/pdf/_asset_id=2583046.pdf. Last accessed August 2010.

53. Iezzoni LI, Measuring quality, outcomes, and cost of care using large databases: the Sixth Regenstrief Conferences, Assessing quality using administrative data, *Ann Intern Med*, 1997;127:666-674.

54. Ibid.

55. Quam L, Ellis L, Venus P, et al., Using claims data for epidemiologic research: the concordance of claims-based criteria with the medical record and patient survey for identifying a hypertensive population, *Med Care*, 1993;31(6):498-507.

56. Shahian DM, Silverstein T, Lovett AF, et al., Comparison of Clinical and Administrative Data Sources for hospital Coronary Artery Bypass Graft Surgery Report Cards, Boston, MA: Tufts University School of Medicine, *Circulation,* 2007;115:1518-1527.

57. Quam L, Ellis L, Venus P, et al.

58. Iezzoni LI, Foley SM, Daley J, et al., Comorbidities, complications, and coding bias: does the number of diagnosis codes matter in predicting in-hospital mortality? *JAMA*, 1992;267(16):2197-2203.

59. Acumen, *Evaluating the Functionality of the Symmetry ETG and Medstat MEG Software in Forming Episodes of Care Using Medicare Data*; 2008. Available at www.cms.gov/Reports/downloads/MaCurdy.pdf . Last accessed August 2010.

60**.** Shahian DM, Silverstein T, Lovett AF, et al.

61. Adams JL, Mehrotra A, Thomas JW, et al.

62. Robinson JC, Williams T, Yanagihara D, Measurement of and reward for efficiency in California's pay-for-performance program, *Health Affairs*, 2009;28(5):1438-1447.

63. Hing E, Burt CW, Characteristics of office-based physicians and their practices: United States, 2005–2006, *Vital and Health Statistics*, 2008;13(166). Available at http://www.cdc.gov/nchs/data/series/sr_13/sr13_166.pdf. Last accessed August 2010

64. Ingenix, Ingenix makes market-leading Episode Treatment Groups® methodology available for public review; 2009. Available at www.ingenix.com/News/Article/92/. Last accessed August 2010.

65. Thomson Reuters, Thomson Reuters Medical Episode Grouper Transparency Documentation, June 30, 2009. Available at http://thomsonreuters.com/content/healthcare/white_papers/medical_episode_grouper_transpar. Last accessed August 2010.

66. NQF Measure Evaluation Criteria. Available at www.qualityforum.org/Measuring_Performance/Submitting_Standards/Measure_Evaluation_Criteria.aspx. Last accessed September 2010.

67. Resource use measures are measures of costs or inputs and are not directly correlated with quality. They can be used for internal improvement and review but do not independently indicate where improvements in quality can be made.

Doris H. Lotz, MD, MPH (Co-Chair)
New Hampshire Department of Health and Human Services, Concord, NH

Bruce Steinwald, MBA (Co-Chair)
Independent Consultant, Washington, DC

Paul G. Barnett, PhD
VA Palo Alto Health Care System, Menlo Park, CA

Jack Bowhan
Wisconsin Collaborative for Healthcare Quality, Middleton, WI

Jeptha P. Curtis, MD
Yale University School of Medicine, New Haven, CT

Kurtis S. Elward, MD, MPH
Family Medicine of Albemarle, Charlottesville, VA

William E. Golden, MD
Arkansas Medicaid, Little Rock, AR

Lisa M. Grabert, MPH
American Hospital Association, Washington, DC

Ethan A. Halm, MD, MPH
University of Texas Southwestern Medical Center, Dallas, TX

Ann L. Hendrich, RN, MSN, PhD(c)
Ascension Health, St. Louis, MO

Thomas H. Lee, MD
Partners HealthCare System, Inc., Boston, MA

Renée Markus Hodin, JD
Community Catalyst, Boston, MA

Jack Needleman, PhD
University of California, Los Angeles School of Public Health

Mary Kay O'Neill, MD, MBA
CIGNA HealthCare, Seattle, WA

David F. Penson, MD, MPH
Vanderbilt University Medical Center, Nashville, TN

Steve Phillips, MPA
Johnson & Johnson Health Care Systems Inc., Washington, DC

David Redfearn, PhD
WellPoint, Woodland Hills, CA

Jeffrey B. Rich, MD
Mid-Atlantic Cardiothoracic Surgeons Ltd., Norfolk, VA

William L. Rich, III, MD
Northern Virginia Ophthalmology Associates, Falls Church, VA

Tom Rosenthal, MD
UCLA School of Medicine, Los Angeles, CA

Barbara A. Rudolph, PhD, MSSW
The Leapfrog Group, Fitchburg, WI

Joseph Stephansky, PhD
Michigan Health & Hospital Association, Lansing, MI

James N. Weinstein, DO, MS
The Dartmouth Institute for Health Policy and Clinical Practice & The Dartmouth-Hitchcock Clinic, Lebanon, NH

Dolores Yanagihara, MPH
Integrated Healthcare Association, Oakland, CA

NQF Staff

Helen Burstin, MD, MPH
Senior Vice President

Marybeth Farquhar, PhD, MSN, RN
Vice President

Sally Turbyville, MA, MS
Senior Director

Jennifer Podulka, MPAff
Senior Director

Edison Machado, MD, MPH
Senior Director

Ashlie Wilbon, RN, MPH
Project Manager

Sarah Fanta
Research Analyst

The following table provides a side-by-side comparison of the standard NQF-evaluation criteria (left column) and the Proposed Resource Use Measure Evaluation Criteria (right column). The resource use evaluation criteria is grounded in the standard NQF evaluation criteria, keeping the four major criteria (importance, scientific acceptability, usability, and feasibility) in place, but modifying the subcriteria as appropriate to reflect the specific needs of resource use measure evaluation. Each of the standard NQF subcriteria that are applicable to resource use measures is included in the right column; additions and substitutions to the criteria are noted by the bolded text. The notes for the subcriteria have also been updated to provide specific guidance around meeting the criteria for resource use measures, including appropriate data analysis methods and clarification of concepts.

NQF Quality Measure Evaluation Criteria	***Proposed Resource Use Measure Evaluation Criteria***
Conditions for Consideration	
A. The measure steward is a governmental organization or a Measure Steward Agreement is signed. B. The measure owner/steward verifies there is an identified responsible entity and process to maintain and update the measure on a schedule that is commensurate with the rate of clinical innovation, but at least every 3 years. C. The intended use of the measure includes both public reporting and quality improvement. D. The requested measure submission information is complete. Generally, measures should be fully developed and tested so that all the evaluation criteria have been addressed and information needed to evaluate the measure is provided. Measures that have not been tested are only potentially eligible for a time-limited endorsement and in that case, measure owners must verify that testing will be completed within 12 months of endorsement.	A. The measure steward is a governmental organization or a Measure Steward Agreement is signed. B. The measure owner/steward verifies there is an identified responsible entity and process to maintain and update the measure on a schedule that is commensurate with the rate of clinical innovation, but at least every 3 years. C. The intended use of the measure includes both public reporting and quality improvement. D. The requested measure submission information is complete. Generally, measures should be fully developed and tested so that all the evaluation criteria have been addressed and information needed to evaluate the measure is provided. ***Based on existing NQF policy, complex measures are not eligible or time-limited endorsement. Resource use measures are complex in nature and therefore must be fully tested at the time of submission.***

NQF Quality Measure Evaluation Criteria	***Proposed Resource Use Measure Evaluation Criteria***
1. Importance to measure and report	
Extent to which the specific measure focus is important to making significant gains in healthcare quality (safety, timeliness, effectiveness, efficiency, equity, patient-centeredness) and improving health outcomes for a specific high-impact aspect of healthcare where there is variation in or a demonstrated high-impact aspect of healthcare (e.g., affects large numbers, leading cause of morbidity/mortality, high resource use [current and/or future], severity of illness, and patient/societal consequences of poor quality) or overall poor performance. Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria.	***Resource use measures will be evaluated based on*** the extent to which the specific measure focus is important to making ***significant contributions toward understanding healthcare costs*** for a specific high-impact aspect of healthcare where there is ***unexplained*** variation or a demonstrated high-impact aspect of healthcare (e.g., affects large numbers, leading cause of morbidity/mortality, ***high or unexplained variation in resource use [current and/or future]***, severity of illness, and patient/ societal consequences of poor quality) or overall poor performance.

1a. The measure focus addresses: • Specific national health Goal/Priority identified by the Partners of the NQF convened National Priorities Partnership: OR • Demonstrated high-impact aspect of healthcare (e.g., affects large numbers, leading cause of morbidity/mortality, high resource use [current and/or future], severity of illness, and patient/societal consequences of poor quality). 1b. Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating considerable variation, or overall poor performance, in the quality of care across providers and/or population groups (disparities in care). 1c. The measure focus is: an outcome (e.g., morbidity, mortality,function, health-related quality of life) that isrelevant to, or associated with, a nationalhealth goal/priority, the condition, population,and/or care being addressed; OR if an intermediate outcome, process,structure, etc., there is evidence thatsupports the specific measure focus as follows:	1a. The measure focus addresses: • Specific national health Goal/Priority identified by the Partners of the NQF convened National Priorities Partnership: OR • Demonstrated high-impact aspect of healthcare[1] (e.g., affects large numbers, leading cause of morbidity/mortality, high resource use [current and/or future], severity of illness, and patient/societal consequences of poor quality). 1b. Demonstration of resource use or cost problems and opportunity for improvement, i.e., data[2] demonstrating ***unexplained*** variation in the delivery of care across providers and/or population groups (disparities in care). 1c. The measure focus is: an outcome (e.g., morbidity, mortality,function, health-related quality of life) that isrelevant to, or associated with, a nationalhealth goal/priority, the condition, population,and/or care being addressed; OR if an intermediate outcome, process,structure, etc., there is evidence thatsupports the specific measure focus as follows: --Efficiency[3] – demonstration of an association between the measured resource use and level of performance with respect to one or more of the other five IOM aims of quality. IOM Quality Domains:• Effectiveness• Efficiency• Equity• Patient-centered• Safety• Timeliness

--Intermediate outcome – evidence that the measured intermediate outcome (e.g., blood pressure, Hba1c) leads to improved health/avoidance of harm or cost/benefit. --Process – evidence that the measured clinical or administrative process leads to improved health/avoidance of harm and if the measure focus is on one step in a multistep care process, it measures the step that has the greatest effect on improving the specified desired outcome(s). --Structure – evidence that the measured structure supports the consistent delivery of effective processes or access that lead to improved health/avoidance of harm or cost/benefit. --Patient experience – evidence that an association exists between the measure of patient experience of healthcare and the outcomes, values, and preferences of individuals/the public. --Access – evidence that an association exists between access to a health service and the outcomes of, or experience with, care. --Efficiency – demonstration of an association between the measured resource use and level of performance with respect to one or more of the other five IOM aims of quality. IOM Quality Domains:• Effectiveness• Efficiency• Equity• Patient-centered• Safety• Timeliness	
Composite. 1d. The purpose/objective of the composite measure and the construct for quality are clearly described.	1d. The purpose/objective ***of the resource use measure (including its components)*** and the construct for ***resource use/costs*** are clearly described.
Composite. 1e. The component items/ measures (e.g., types, focus) that are included in the composite are consistent with and representative of the conceptual construct for quality represented by the composite measure. Whether the composite measure development begins with a conceptual construct or a set of measures, the measures included must be conceptually coherent and consistent with the purpose.	1e. ***The resource units (e.g., types of resources/costs)*** that are included in ***the resource use measure*** are consistent with and representative of the conceptual construct represented by the measure. Whether the ***resource use*** measure development begins with a conceptual construct or a set of ***resource units,*** the ***units*** included must be conceptually coherent and consistent with the purpose.

NQF Quality Measure Evaluation Criteria	***Proposed Resource Use Measure Evaluation Criteria***
2. Scientific acceptability of the measure properties	
Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented.	Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about ***the cost or resources used to deliver care.***
2a. The measure is well defined and precisely specified so that it can be implemented consistently within and across organizations and allow for comparability. The required data elements are of high quality as defined by NQF's Health Information Technology Expert Panel (HITEP). 2b. Reliability testing demonstrates that the measure results are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period. 2c. Validity testing demonstrates that the measure reflects the quality of care provided, adequately distinguishing good and poor quality. If face validity is the only validity addressed, it is systematically assessed. 2d. Clinically necessary measure exclusions are identified and must be: supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; AND Clinically appropriate exception (e.g., contraindication) to eligibility for the measure focus; AND Precisely defined and specified. If there is substantial variability in exclusions across providers, the measure is specified so that exclusions are computable and the effect on the measure is transparent (i.e., impact clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion). If patient preference (e.g., informed decision making) is a basis for exclusion, there must be evidence that it strongly impacts performance on the measure, and the measure must be specified so that the information about patient preference and the effect on the measure is	2a. The measure is well defined and precisely specified[4] so that it can be implemented consistently within and across organizations and allow for comparability. The required data elements are of high quality as defined by NQF's Health Information Technology Expert Panel (HITEP).[5] 2b. Reliability testing[6] demonstrates that the measure results are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period. 2c. Validity testing[7,8] demonstrates that the measure reflects ***the cost of care or resources provided, adequately distinguishing high and low cost or resource use.*** 2d. Clinically necessary measure exclusions are identified and must be: supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion[9,10]; AND Clinically appropriate exception (e.g., contraindication) to eligibility for the measure focus; AND Precisely defined and specified. If there is substantial variability in exclusions across providers, the measure is specified so that exclusions are computable and the effect on the measure is transparent (i.e., impact clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion). If patient preference (e.g., informed decision making) is a basis for exclusion[11], there must be evidence that it strongly impacts performance on the measure, and the measure must be specified so that the information about patient preference and the

transparent (e.g., numerator category computed separately, denominator exclusion category computed separately).	effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately).
2e. For outcome measures and other measures (e.g., resource use) when indicated: an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified and is based on patient clinical factors that influence the measured outcome (but not disparities in care) and are present at start of care OR rationale/data support no risk adjustment.	2e. For outcome measures and other measures (e.g., resource use) when indicated: an evidence-based risk-adjustment strategy[12,13] (e.g., risk models, risk stratification) is specified and is based on patient clinical factors that influence the measured outcome (but not disparities in care) and are present at start of care OR rationale/data support no risk adjustment.
2f. Data analysis demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful differences in performance.	2f. Data analysis demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful differences in performance[14].
2g. If multiple data sources/methods are allowed, there is demonstration that they produce comparable results.	2g. If multiple data sources/methods are allowed, there is demonstration that they produce comparable results.
2h. If disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results (e.g., by race, ethnicity, socioeconomic status, gender) OR rationale/data justifies why stratification is not necessary or not feasible.	2h. If disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results (e.g., by race, ethnicity, socioeconomic status, gender) OR rationale/data justifies why stratification is not necessary or not feasible.

NQF Quality Measure Evaluation Criteria	***Proposed Resource Use Measure Evaluation Criteria***
3. Usability	
Extent to which intended audiences (e.g., consumers, purchasers, providers, policymakers) can understand the results of the measure and are likely to find them useful for decision-making.	Extent to which intended audiences (e.g., consumers, purchasers, providers, policymakers) can understand the results of the measure and are likely to find them useful for decision-making ***Usefulness of resource use measures are in the context of quality.***
3a. Demonstration that information produced by the measure is meaningful, understandable, and useful to the intended audience(s) for both public reporting (e.g., focus group, cognitive testing) and informing quality improvement (e.g., quality improvement). An important outcome that may not have an identified improvement strategy still can be useful for informing quality improvement by identifying the need for and stimulating new approaches to improvement.	3a. Demonstration that information produced by the measure is meaningful, understandable, and useful to the intended audience(s) for both public reporting[15] (e.g., focus group, cognitive testing) and informing quality improvement[16] (e.g., quality improvement). An important outcome that may not have an identified improvement strategy still can be useful for informing quality improvement by identifying the need for and stimulating new approaches to improvement.
3b. The measure specifications are harmonized with other measures and are applicable to multiple levels and settings.	3b. The measure specifications are harmonized with other measures and are applicable to multiple levels and settings.[17]
3c. Review of existing endorsed measures and measure sets demonstrates that the measure provides a distinctive or additive value to existing NQF-endorsed measures (e.g., provides a more complete picture of quality for a particular condition or aspect of healthcare, is a more valid or efficient way to measure).	***3c. List NQF-endorsed quality measures known to have been used alongside the resource use measure.***
Composite. 3d. Data detail is maintained such that the composite measure can be decomposed into its components to facilitate transparency and understanding.	3d. Data ***and result*** detail are maintained such that the ***resource use measure, including the clinical and construction logic for a defined unit for measurement,*** can be decomposed to facilitate transparency and understanding.
Composite. 3e. Demonstration (through pilot testing or operational data) that the composite measure achieves the stated purpose/objective.	3e. Demonstration (through pilot testing or operational data) that the ***resources use*** measure achieves the stated purpose/objective[18].

NQF Quality Measure Evaluation Criteria	***Proposed Resource Use Measure Evaluation Criteria***
4. Feasibility	
Extent to which the required data are readily available, retrievable without undue burden, and can be implemented for performance measurement.	Extent to which the required data are readily available, retrievable without undue burden, and can be implemented for performance measurement[19].
4a. For clinical measures, required data elements are routinely generated concurrent with and as a byproduct of care processes during care delivery. 4b. The required data elements are available in electronic sources. If the required data are not in existing electronic sources, a credible, near-term path to electronic collection by most providers is specified, and clinical data elements are specified for transition to the electronic health record. 4c. Exclusions should not require additional data sources beyond what is required for scoring the measure (e.g., numerator and denominator) unless justified as supporting measure validity. 4d. Susceptibility to inaccuracies, errors, or unintended consequences and the ability to audit the data items to detect such problems are identified. 4e. Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, etc.) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use).	4a. For clinical measures, required data elements are routinely generated concurrent with and as a byproduct of care processes during care delivery. 4b. The required data elements for the resource use measures are available in electronic sources. If the required data are not in existing electronic sources, a credible, near-term path to electronic collection by most providers is specified, and clinical data elements are specified for transition to the electronic health record. 4c. Exclusions should not require additional data sources beyond what is required for scoring the measure (e.g., numerator and denominator) unless justified as supporting measure validity. 4d. Susceptibility to inaccuracies, errors, or unintended consequences and the ability to audit the data items to detect such problems are identified. 4e. Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, etc.) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use).

Notes for Proposed Resource Use Evaluation Criteria

Notes for Importance

1. Examples of data on opportunity for improvement include, but are not limited to: prior studies, epidemiologic data, measure data from pilot testing, or implementation. If data are not available, the measure focus is systematically assessed (e.g., expert panel rating) and judged to be a quality or ***performance*** problem.
2. ***Findings from peer reviewed literature review, empirical data are examples of information that can be used to justify importance and demonstrate unexplained variation. It is the proof of the measure's concept that enables the Committee to determine if the measure is valid in addressing this concept.***
3. ***Efficiency is a multi-dimensional concept that includes inputs and outputs, and specifically the amount of resources used (the inputs) and the degree of quality or health outcomes achieved (output)—resource use measures alone do not capture efficiency but are a building block of efficiency: Efficiency = fx (outcomes, resource use).*** Efficiency of care is a measurement construct of cost of care or resource utilization associated with a specified level of quality of care. It is a measure of the relationship of the cost of care associated with a specific level of performance measured with respect to the other five IOM aims of quality. Efficiency might be thought of as a ratio, with quality as the numerator and cost as the denominator. As such, efficiency is directly proportional to quality, and inversely proportional to cost. (NQF's Measurement Framework: Evaluating Efficiency Across Episodes Of Care; based on AQA Principles of Efficiency Measures.

Notes for Scientific Acceptability

4. Well defined and ***precise specifications for resource use measures include each of the five specification modules (i.e. data protocol, measure clinical logic and method, measure construction logic, adjustments for comparability, and reporting***). For those steps not included in the specifications, justification for and implications of not specifying those steps is required. Specifications should also include the identification of target population, measurement time window, exclusions, risk adjustment, definitions, data elements, data source and instructions, sampling, scoring/computation. Data protocol steps are critical to the reliability and validity of the measure; specifications must be detailed enough such that users can execute necessary.
5. The HITEP criteria for high quality data include: a) data captured from an authoritative/accurate source; b) data are coded using recognized data standards; c) method of capturing data electronically fits the workflow of the authoritative source; d) data are available in EHRs; and e) data are auditable. NQF. Health Information Technology Expert Panel Report: Recommended Common Data Types and Prioritized Performance Measures for Electronic Healthcare Information Systems. Washington, DC: NQF; 2008.
6. Examples of reliability testing include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest, split-half reliability***. Reliability testing may address the data items or final measure score. Reliability for resource use measures should be demonstrated for each of the modules (data protocol methodology, clinical logic and measure construction,***

stratification, risk adjustment, and costing methodology). For those steps not included in the specifications, justification for and implications of not specifying those functions is required.

7. Examples of validity testing include, but are not limited to: determining if measure scores adequately distinguish between providers known to have ***high or low resource use or cost*** assessed by another valid method; correlation of measure scores with another valid indicator of ***resource use or cost*** for the specific topic; ability of measure scores to predict scores on some other related valid measure; content validity for multi-item scales/tests. ***The scoring/aggregation and weighting rules used during measure scoring and construction are consistent with the conceptual construct. If you use differential weighting it should be justified. Differential weights are determined by empirical analyses or a systematic assessment of expert opinion or values-based priorities. This is in addition to weighting the pricing methodology introduces, if any.***
8. Face validity is a subjective assessment by experts of whether the measure reflects the cost or resource use of the care delivered. If face validity is the only validity addressed, it must have been systematically assessed (e.g., ratings by relevant stakeholders), the measure is judged to represent ***cost or resource use*** for the specific topic, and the measure focus is the most important aspect of ***cost or resource use*** for the specific topic. ***Validity testing for resource use measures should demonstrate validity for each module (clinical logic and measure construction, risk adjustment, stratification, costing methodology, and reporting (including attribution, peer groups, threshold and outliers, benchmarking). For those steps not included in the specifications, justification for and implications of not specifying those steps is required.***
9. Examples of evidence that exclusion distorts measure results include, but are not limited to: frequency ***or cost*** of occurrence, sensitivity analyses with and without the exclusion, and variability of exclusions across providers. ***For example, a measure may specify to exclude a patient with active from a COPD resource use measure because cancer is the dominant medical condition with known high costs. Exclusions must be justified and supported with appropriate evidence on the effect of the exclusions.***
10. ***Testing for resource use measure exclusions should address the appropriate specification steps (i.e data protocol, clinical logic, and thresholds and outliers). For those exclusions not addressed, justification for and implications of not addressing them is required.***
11. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions. ***If there is substantial variability in exclusions across providers, the measure is specified so that exclusions are computable and the effect on the measure is transparent (i.e., impact clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion). If patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that it strongly impacts performance on the measure and the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). Patient co-pays or co-amounts should not exclude a service from inclusion or justification to exclude these patients or services should be provided. Specifically, claims for services received by the patient should be included in the measure even when the patient pays a portion of the claims, unless otherwise justified—all approaches should be transparent.***

12. Risk factors that influence quality outcomes ***or resource use/cost*** should not be specified as exclusions, exclusions for resource use or cost that influence results must be justified.
13. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care such as race, socioeconomic status, gender (e.g., poorer treatment outcomes of African American men with prostate cancer, inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than adjusting out differences.
14. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall poor performance may not demonstrate much variability across providers.

Notes on Usability

15. Public reporting and quality improvements ***(including strategies around cost or resource use management)*** are not limited to provider-level measures—community and population measures also are relevant for reporting and improvement.
16. Informing improvement may be facilitated using ***relevant quality improvement initiatives or cost containment strategies***.
17. Measure harmonization refers to the standardization of specifications for similar measures on the same topic (e.g., influenza immunization of patients in hospitals or nursing homes), related measures for the same target population (e.g., eye exam and HbA1c for patients with diabetes), or definitions applicable to many measures (e.g., age designation for children) so that they are uniform or compatible, unless differences are dictated by the evidence. The dimensions of harmonization can include numerator, denominator, exclusions, and data source and collection instructions. The extent of harmonization depends on the relationship of the measures, the evidence for the specific measure focus, and differences in data sources.
18. ***Pilot testing results should address how and who has used the measure practically and in effecting decisions (e.g., concurrent validity testing using correlation analysis).***

Notes on Feasibility

19. All data collection must conform to laws regarding protected health information. Patient confidentiality is of particular concern with measures based on patient surveys and when there are small numbers of patients.