TO: NQF Members and Public

FR: NQF Staff

RE: Comment on the Proposed Recommendations for Evaluating Episode Groupers

DA: April 14, 2014

In recent years, there has been a drive toward episode-based performance measurement to better understand the utilization and costs associated with certain conditions. Even with growing interest in expanding performance measurement approaches to include episode-based measures, there remains a great deal to learn about these approaches and in understanding the challenges to measuring costs through this lens. Among the various approaches to measuring episodes of care, episode grouper software tools have been evolving, namely in the commercial sector, as a widely accepted method for aggregating claims data into episodes to assess condition-specific utilization and costs. The growing interest in the use of these tools has generated further interest in exploring the need for and implications of a multi-stakeholder consensus-based review of episode groupers. This project seeks to explore and understand the key considerations for and challenges in constructing an episode grouper and defining its key characteristics in order to inform recommendations for evaluating groupers.

To guide this effort, NQF convened a 21-member Expert Panel comprised of stakeholders representing purchasers, health plans, providers and clinicians with expertise in performance measurement, measurement methodologies, clinical quality improvement, and the development of episode groupers. The focus of the panel was not specific to a particular grouper or product, but broad in nature such that the recommended criteria could be applied to any episode grouper that may be submitted for evaluation.

This draft document is being provided to you at this time for purposes of review and comment only and is not intended to be used for voting purposes. You may post your comments and view the comments of others on the NQF website. Thank you for your interest in NQF's work. We look forward to your review and comments.

NQF Member and Public comments must be submitted no later than 6:00 pm ET, May 13, 2014.

Table of Contents

Introduction

In recent years, there has been a drive toward episode-based performance measurement to better understand the utilization and costs associated with certain conditions. Episode-based measurement facilitates this by attributing care to condition-specific or procedure-specific episodes based on the relationship of the healthcare service to the care of a specific condition (i.e., all diabetes-related care is attributed to the diabetes episode of care). Even with growing interest in expanding performance measurement approaches to include episode-based measures, there remains a great deal to learn about these approaches and in understanding the challenges to measuring costs through this lens. Both the public and private sectors have begun using episode-based measurement as a basis for understanding utilization and costs for specific episodes through the implementation and testing of physician profiling and payment programs.[1] To meet the growing demand for this type of information, various measurement approaches have been developed for applications such as bundled payments, gain sharing, and other types of episode-based payment.

Among the various approaches to measuring episodes of care, episode grouper software tools have been evolving, namely in the commercial sector, as a widely accepted method for aggregating claims data into episodes to assess condition-specific utilization and costs. Using a grouper, healthcare services provided over a defined period of time can be analyzed and grouped by specific clinical conditions to generate an overall picture of the services utilized to manage that condition. Among the various vendors offering groupers in the market, the methods by which claims are grouped and attributed to episodes varies significantly. The growing interest in the use of these tools to better understand healthcare costs, the lack of transparency and inherent complexity around the methodologies employed and the recent investment by the Centers for Medicare and Medicaid Services (CMS) to develop a publicly available episode grouper for Medicare beneficiaries has generated further interest in exploring the need for and implications of a multi-stakeholder consensus-based review of episode groupers.

The National Quality Forum (NQF) has undertaken several projects focused on cost and resource use measurement beginning in 2009 with the Patient-Focused Episodes of Care Framework which resulted in a framework which provided a conceptual model for measuring costs across a patient-centered episode of care. Building on that foundational framework, NQF embarked on its first effort to evaluate and recommend cost and resource use measures for endorsement as national consensus standards in 2010, resulting in eight endorsed measures. Lessons from these efforts, including the evaluation and endorsement of two episode-based measures derived from a grouper, have laid the foundation for this project which seeks to explore and understand the key considerations for and challenges in constructing an episode grouper and defining its key characteristics in order to inform recommendations for evaluating groupers.

Specifically, the purpose of this project is to:

- Define the characteristics and challenges of constructing episode groupers;

- Determine the key elements of episode groupers that should be submitted to NQF for evaluation;
- Establish an initial set of criteria by which episode groupers should be evaluated for NQF endorsement; and
- Identify implications and considerations for NQF-endorsement of episode groupers.

Episode Grouper Expert Panel

To guide this effort, NQF convened a 21-member Expert Panel comprised of stakeholders representing purchasers, health plans, providers and clinicians with expertise in performance measurement, measurement methodologies, clinical quality improvement, and the development of episode groupers. The Expert Panel gathered for a two-day in-person meeting in Washington, DC on February 5th and 6th, 2014 to discuss the key issues identified above and provide recommendations on the evaluation of episode groupers. The focus of the panel was not specific to a particular grouper or product, but broad in nature such that the recommended criteria could be applied to any episode grouper that may be submitted for evaluation.

The Policy Landscape

Maintaining program viability in a climate of rising health care costs and increasing demand has been the focus of the last three decades of legislation to amend the 1965 Medicare provisions of the Social Security Act. Physicians are currently reimbursed on a Resource Based Relative Value Scale (RBRVS) established in 1992, a fee-for-service design that links physician payment to the volume of services performed. While the movement has been towards payment models that reward both efficient and effective care through standardized payments for services, reforming physician reimbursement has been challenging.

In 2008 Congress passed the Medicare Improvements for Patients and Providers Act (MIPPA), legislation that expanded coverage for Medicare beneficiaries and enacted provisions to better align quality and value by providing feedback to physicians on comparative resource use. The MIPPA legislation amended the Social Security Act and established the Physician Resource Use Measurement and Reporting Program with the intent to control costs by informing physicians on resource use by patients in their care on an episode, per capita, or both episode and per capita basis.[2]

The Patient Protection and Affordable Care Act (PPACA) passed in 2010 further amended the Social Security Act to broaden the scope of reporting and analysis on resource use directly to physicians, requiring CMS to develop an endorsed, publicly available episode grouper with specific functional requirements by January 1, 2012.

It is anticipated that the episode-based measurement produced from the grouper can be used in other federal programs aimed at value-based physician measurement. The Physician Feedback Program supports value-based purchasing reforms through the development of a value-based payment modifier (VBPM).[3] The VPBM program, administered by CMS, rewards physicians who have satisfied the requirements of performance measures through the Physician Quality Reporting System (PQRS) by an upward adjustment of Medicare physician fee schedule payments. Physicians can review the results of the Physician Feedback Program in Quality Resource Use Reports (QRURs), allowing for comparisons

between providers on cost and resource use.[4] CMS will begin phasing in VPBM reimbursement on January 1, 2015, with the goal of complete conversion to reimbursement for Medicare services by VPBM by 2017.[5]

Current Landscape of Episode Groupers

Public Grouper

In response to the legislative mandate to create a publicly available grouper for Medicare, CMS began the process to solicit proposals for episode grouping approaches from public and private entities to be considered for adoption. In 2012, CMS awarded the contract to develop a public domain episode grouper for Medicare to Brandeis University. The Medicare grouper was scoped for development over a four year period as a joint effort between the American Board of Medical Specialties Research and Education Foundation, the American Medical Association Physician Consortium for Performance Improvement, the Health Care Incentives Improvement Institute, Inc. (HCI3), the Medicare Quality Improvement Organization for New York State, and Booz Allen Hamilton.[6]

Commercial Groupers

There are several commercial episode groupers that have been in use in the private sector for many years, including the OptumInsight Symmetry Episode Treatment Groups product, the 3M Patient Focused Episode grouper, the Truven Medstat Medical Episode Grouper (MEG), HCI3 Prometheus, and the Cave grouper. These episode grouping products are used by various stakeholders in various applications. For example, commercial insurers and managed care organizations have used episode groupers to facilitate bundled payment and value-based performance programs. Health systems have also used these tools to examine prevalence rates for various conditions, incidence rates for various treatments, and complication rates to support internal quality improvement; and purchasers have used groupers to understand provider utilization and cost variation.

Public and Private Sector Alignment

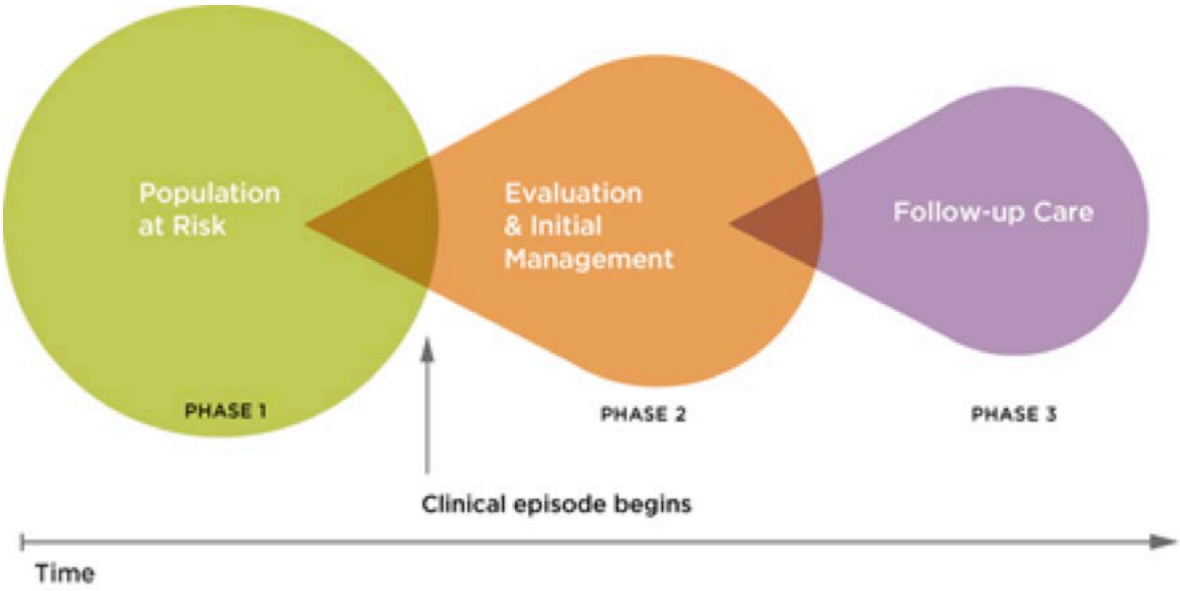
The use of commercial groupers products often varies by region; even within a region, stakeholders may have invested in different products based on their specific needs and preferences. While the groupers perform similar functions, their approach to grouping claims varies and thereby limits the comparability of results among the users of the various systems. Also, the data driving the analysis within the commercial groupers have largely been for commercial populations (<65 years old). The Medicare grouper, inherent to its purpose, is designed to group Medicare claims (generally population ≥65 years old), adding yet another layer of complexity and misalignment of the existing tools. While the groupers are not necessarily limited to grouping claims for a particular age range, further testing would be required in order to determine the appropriateness of the groupers when used with data from across the lifespan, which may be beyond the primary scope and intended use.

Defining Episodes

The underlying concept on which episode groupers rely is the episode of care. Recognizing there are varying definitions of an episode of care, the NQF Episodes of Care Measurement Framework defines an episode as “a series of temporally contiguous healthcare services related to the treatment of a given

spell of illness or provided in response to a specific request by the patient or other relevant entity."[7] These healthcare services can be administered by one or more providers over the course of the episode.[8] Figure 1, developed as a product of the NQF Episodes of Care Measurement Framework report, illustrates the three phases of an episode of care including the population at risk, evaluation and initial management and follow- up care phases, through which a patient would flow over the course of an illness.

Figure 1. Episode of Care Conceptual Model



Using this model to understand acute episodes, an acute illness such as acute myocardial infarction (AMI) generally begins with an event for which treatment in Phase 2 (e.g., surgery or stent placement) and follow up care in Phase 3 could encompass cardiac rehabilitation as well as the management of the underlying coronary artery disease (CAD) over the lifespan. Appendix B includes a detailed illustration of an AMI episode using this model. For measurement purposes, this does not enable the identification of an end date to the episode. In these instances, a time period for the chronic episode (e.g., CAD) is generally defined to capture the healthcare services related to the treatment of the condition for a specified period (e.g. a 12-month window) that may be unrelated to the specific condition. For an acute episode (e.g., broken arm), the start and end date are generally discrete periods in time starting with the event where the broken arm occurred until the arm is completely healed, allowing for a more definitive timeframe for the episode.

One of the major challenges in defining episodes is determining when and how to attribute claims for the treatment of conditions that occur as complications of the underlying condition or procedure. During an episode for a given clinical condition, any series of complications could develop; these complications may be considered as individual episodes (e.g., separate CAD and AMI episodes) or be attributed to another related episode (e.g., AMI claims included in the CAD episode). In these cases, the challenge is in fairly and appropriately attributing the utilization for the complication(s) to the

underlying condition. Using an episode-based approach can also highlight the linkage of services provided in different settings and by different providers into an episode that otherwise may not have been considered together (e.g., diabetic podiatry visit and acute admission for diabetes complications are linked to the diabetes episode).

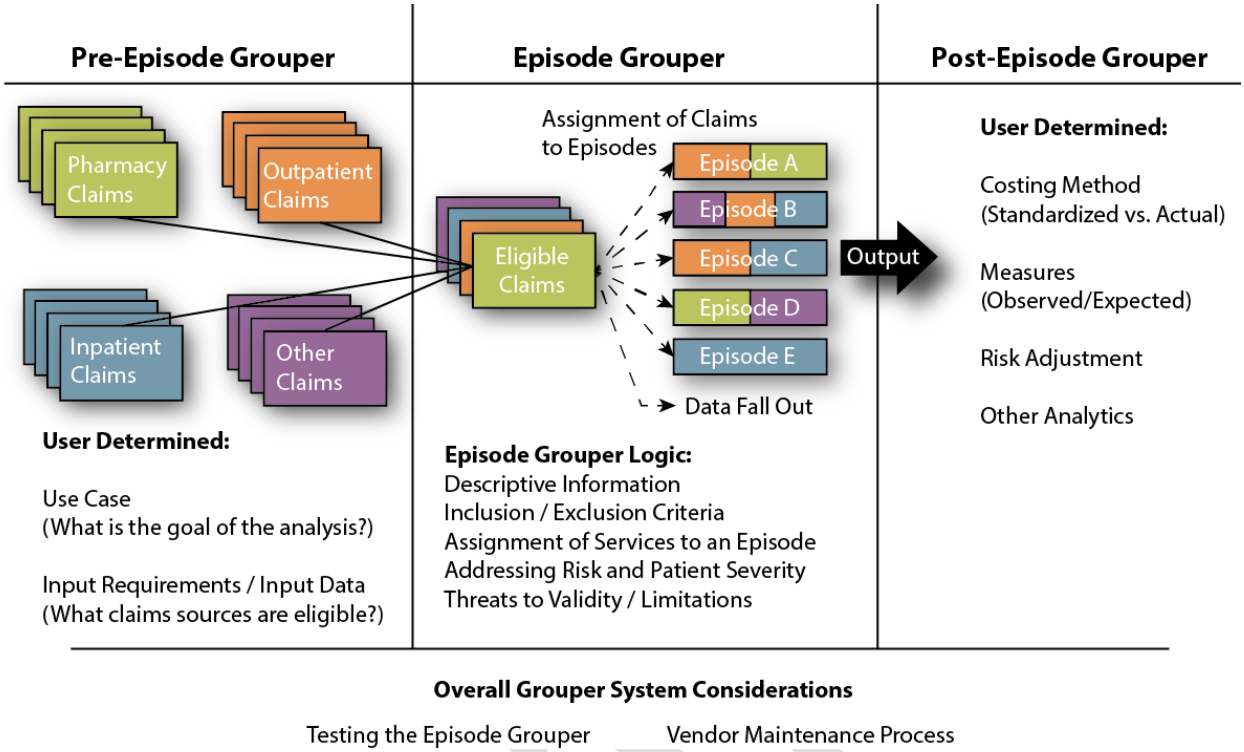
Understanding Episode Groupers

Episode groupers can be defined as the software and logic that assigns patient utilization to clinically relevant episodes of care. Episode grouping is operationalized using software tools to provide a picture of healthcare utilization for relevant conditions over a defined period of time. Currently most episode grouper software is developed to parse administrative claims data into episodes of care; however, development efforts are underway to use electronic health record data. By understanding the utilization patterns for a condition using administrative claims, the dollar amount assigned to each claim in an episode can be aggregated to understand total cost for an episode of care. Many groupers have the ability to create hundreds of condition-specific episodes. The creation of these episodes is dependent on the intricate decision logic that determines to which episode a claim should be assigned. A single patient with multiple co-occurring conditions may trigger multiple condition-specific episodes during a time period. In a simple example, a patient who has been diagnosed with heart failure and diabetes visits his primary care provider. During the visit, the provider checks his blood sugar level and orders a heart imaging study. Although both services were initiated in a single visit, the blood sugar check would be assigned to the diabetes episode of care and the heart imaging study would be assigned to the heart failure episode.

The functionality of these tools becomes complex since creating condition or procedure episodes may pose challenges for maintaining adequate sample size and accounting for patient severity and risk; each grouper tool that exists in the market has its own rules and logic for addressing these common challenges.

Given the complexity and array of approaches and methodologies used in currently developed episode groupers, the Panel characterized the core components common to all groupers that would need to be transparent in an evaluation. Figure 2 illustrates the basic function of a patient-centered episode grouper showing the flow of patient-level administrative claims data into the grouper, the grouper functions, and the resulting output. The pre-grouper functionality is primarily user-driven; the intended use of the grouper, or "use case," drives the decision logic for the grouper and the potential for calculating measures to support the use case once the grouping is complete.

Figure 2. Illustrating Patient-Centered Episode Grouping



During grouping (assignment of claims to clinical episodes), logic for addressing risk and severity, determining inclusion and exclusions at both the patient and service level, and addressing threats to validity is applied. Once the claims are aggregated into clinical groupings, or episodes (e.g., Episode A, Episode B, etc.), the grouper enables the analysis of these claims for various measurement purposes including resource utilization, profiling, identification of cost drivers and opportunities for improvement, and to highlight variability of services and examining patient care pathways.

As an alternative to the patient-centered episode of care approach based on the NQF-endorsed Patient Centered Episode of Care Framework, some members of the Panel represented a minority opinion that a provider-centered approach for assigning claims should also be considered. In a provider-centered approach, claims would be grouped into various units of analysis (e.g., outpatient care, inpatient care, ED care or care for only a specific clinical circumstance). Each of these units can be examined for each provider in their individual setting. It was further described that aggregation can be for an entire condition (patient-specific) or its components (provider-specific). It is the ability to discriminate the units that enables provider-centricity and facilitates reporting resource use in a manner that clinicians understand.[9] Conceptually, the provider-centric approach isolates claims based on the services directly provided by that clinician, facilitating accountability.

Submitting Episode Groupers for Evaluation

Throughout the discussion with the Panel, several elements emerged and were identified as key to understanding the grouping approach. These elements set the framework for the depth and breadth of information that should be required for submission for evaluating an episode grouper. Measure

developers submitting episode groupers for endorsement by NQF would be required to provide the following information for evaluation:

Descriptive Information

Episode groupers should be transparent about their specific purpose and their intended use. Developers should express upfront what the grouper is able to accomplish and its core capabilities. Given the array of episode grouper methodologies in the field, the experts agreed that the evaluation should recognize that it will be difficult to determine upfront what type of outputs should be expected from any particular episode grouper.

The Panel agreed that the descriptive section should include a discussion by the developers of the limitations of their grouper, either because design decisions have been made in grouping or limitations in the underlying data. For example, the grouper may not be used to profile cancer patients due to the heterogeneity of the patient population and limited staging information in administrative claims data.

Input Requirements / Input Data

The Panel discussed that data loss or data fallout when using an episode grouper can be a challenge for users. Due to the impact of this data loss on future analytic capabilities, it was noted that it would be important to have transparency around the beginning-to-end data flow, the input requirements, or input data required to run the grouper appropriately, and understand the proportion of data that is lost at each processing step. The developer should note any specific requirements for the completeness of diagnostic coding that would impact the anticipated output of the groupings.

The loss of data may be due to at least two different issues. First, this data loss may be due to ungrouped claims or records. The ability of the user to evaluate these ungrouped claims would help to understand if they should have been grouped. Second, developers should be transparent about the impact of incomplete episodes. For example, episodes may be dropped due to a lack of continuous member enrollment, missing data elements required to create complete episodes, requirements for the number of episodes per clinician, or outlier considerations.

Clinical Logic

The experts agreed that developers submitting episode groupers for evaluation should include a detailed description of the clinical logic that supports the purpose and conceptual framework for the episode grouper. The developer should provide a discussion and a list of codes that trigger the start of an episode and what parameters (e.g., clean period) determine the end of each individual episode within the grouper. The Panel agreed that these episode definitions should be developed alongside clinical experts and should be reviewed and updated regularly.

In addition to being transparent about episode definitions, the Panel discussed the approaches for assigning individual services or claims to an episode. Given that many patients, particularly Medicare beneficiaries, have multiple co-occurring conditions, the Panel agreed that while there is no best practice for this process, developers should be transparent about how an individual claim might be assigned to a particular episode or divided into multiple episodes. This process may use predefined clinical logic, statistical inferences, or decision rules also known as tie-breaker logic. The Panel discussed

the challenges of assigning claims to an episode when the patients being profiled by the episode grouper are being treated for several conditions concurrently.[10]

There was consensus that there may not be a uniquely appropriate approach to assigning a service to an episode. There is an inherent trade-off between tightly defining an episode so that there is homogeneity among the patients within an episode, and generating sufficient sample sizes within each episode to enable reliable and valid inferences of resource use. Given this trade-off, developers may select different approaches to assigning related services to a given episode. The assignment of claims to an episode will likely be different between groupers, based on use, and each may be appropriate. For this reason, it is imperative that developers are transparent about the logic and rationale for claim assignment; however, reconciling the rationale for claim assignment with the intended use of the grouper for a patient seeking care from various providers in multiple settings is a challenging endeavor.[11,12]

Depending on their use, episodes could be defined broadly or more narrowly. If episodes are designed to be broad, related services for a given episode may be included. For example, an AMI episode may be defined broadly including the costs of related percutaneous coronary interventions (PCI) or coronary artery bypass grafting (CABG) procedures. Conversely, episodes may also be designed to be narrowly defined, where related services may be grouped to their own episode. In the example above, AMI may be evaluated without the cost of any related to procedures, and PCI and CABG costs are examined independently.

Finally, the Panel also agreed that the sensitivity of the triggers used to open an episode should be considered during evaluation of an episode grouper to ensure there are not significant numbers of phantom episodes created. The creation of these types of phantom episodes may bias the cost observed within the particular clinical episode.

Addressing Risk and Patient Severity

There may be multiple strategies to handling the issue of risk. First, the grouper can stratify patient risk through the grouping mechanism by creating new episodes for increased risk. The Panel acknowledged the relationship between creating new episodes, and the inherent tradeoff of tightly defining an episode discussed above. Second, groupers may offer supplementary risk modules that can be applied after the grouping function is completed. The Panel reiterated that there are various approaches that would be appropriate depending on the intended use of the grouper; however, the developer should be transparent about their design and their rationale.

There are some inherent limitations in the handling of risk in the development of episodes. Given that many episode groupers currently use administrative claims data, the Panel agreed that there may not be sufficient granularity in the data to capture clinical characteristics or severity for certain episode types (e.g., community- vs. hospital-acquired pneumonia, or staging information for cancer patients).

Testing the Episode Grouper

There was broad agreement that testing should be completed by the developer in order to demonstrate reliability of the output, the validity of the clinical logic in the episodes, and the grouper as a whole.

Further, developers should be transparent about the testing methods and the results derived from their approach. Given the various methodologies and trade-offs required in grouping claims to episodes, the Panel agreed that NQF should not be prescriptive about the submission requirements for testing. Developers should be transparent about how they handled trade-offs, potential threats to validity and how they were addressed.

Description of Vendor Maintenance Process

The Panel noted that the maintenance process for an episode grouper system needs to be extensive and developers should provide detailed information on their process for keeping their system current and their endorsement current. Developers should provide a plan and costs for upgrading new versions of the grouper, recognizing that rapid upgrading may be challenging for users to keep up with. Further, such upgrades may introduce additional costs for users.

Summary of Proposed Submission Items

- Descriptive Information
  - Description of intent of grouper (i.e., use cases such as provider profiling)
  - Planned use of the grouper (e.g., specific programs for public reporting or payment)
  - List of the clinical episodes the grouper is capable of generating
- Input Requirements / Input Data
  - Description of the input and data requirements that enable the grouper to group claims as intended
  - Description of rules for identifying claims that are ineligible for assignment (i.e., fall out claims)
- Clinical Logic
  - Description (including codes) and rationale for clinical inclusions and exclusions
- Description of general rules for assigning claims to each episode, and hierarchies
- Addressing Risk and Patient Severity
  - If the grouper adjusts for risk using a risk adjuster, description of the model, including the factors included, and data demonstrating performance of the model (adequate calibration)
  - If the grouper accounts for patient severity in the assignment of claims to episodes, description of the method for assigning risk, including hierarchies, and logic for assigning these claims
- Testing the Episode Grouper
  - Reliability Testing
    - Description of the testing method/approach
    - Description of the data sample used
    - Description/discussion of results
  - Validity Testing
    - Description of the testing method/approach
    - Description of the data sample used
    - Description/discussion of results
    - Description of threats to validity and limitations of the grouper
    - Discussion of how those threats and limitations were addressed
- Description of the vendor maintenance process (frequency, scope, process, implementation)

- Description of any actual or anticipated unintended consequences identified through the use of or implementation of the grouper, and how the benefits of the use of the grouper might outweigh these unintended consequences
- Fee schedule for purchase and/or implementation of the grouper

Evaluating Episode Groupers for Endorsement

Throughout the Panel discussion, a number of core principles emerged to guide the evaluation of episode groupers. These principles are not intended to limit innovation in the design and methods used in episode groupers; rather they represent a baseline agreement on the critical issues that should be considered when evaluating episode groupers in the future.

1. The episode grouper output should be readily understood and reviewed by affected stakeholders to understand the process of how results were derived and to explain the results to those being measured.
2. The evaluation of the grouper should be done in two-phases: first, evaluation should be focused on the grouper logic itself and the episodes using the submission elements and criteria discussed; second, evaluation of the individual measures that result from the grouper should be evaluated using separate but related criteria.
3. The evaluation of a grouper should be done in the context of the stated intended use. Further, the grouper logic and maintenance processes (e.g., updating the codes, upgrading versions, and routine maintenance) will vary based on the intended use.
4. The episodes should be patient-centered and developed based on the patient experience.
5. Output should be actionable and usable for performance and quality improvement.
6. There are challenges inherent in episode grouping which should be addressed by each developer to provide transparency as to how these challenges are handled (or not) in the tool.
7. The evaluation process should not predetermine what the grouper capabilities or decision logic should be; the methodologies underlying the various episode groupers may have distinct approaches that may all be valid.

Proposed Evaluation Criteria for Episode Groupers

The Expert Panel began the process of considering evaluation criteria for episode groupers by reviewing the existing NQF resource use measure evaluation criteria. The group assessed which criteria may be relevant to the evaluation of episode groupers, and whether additional criteria should be considered. Candidate resource use consensus standards are evaluated by NQF Steering Committees for their suitability for endorsement based on four major criteria in the following hierarchical order: *Importance to Measure and Report, Scientific Acceptability of Measure Properties, Usability and Use, and Feasibility*. A fifth criterion, *Related and Competing Measures*, is applied as needed to measures that have been identified with similar measure specifications.

The Expert Panel proposed three major criteria that should be used to evaluate episode groupers by future multi-stakeholder panels.

1. Scientific Acceptability of the Episode Grouper

The Panel discussed that the goal of this criterion should be to determine the extent to which the episode grouper produces consistent (reliable) and credible (valid) results about the cost or resources used to deliver care.

Reliability

The Panel agreed that reliability of the episode grouper should be assessed. Developers of episode groupers should provide reliability testing results that demonstrate that the grouping results are repeatable. Secondly, the developers should demonstrate how the episode grouper performs across data sets of variable size and multiple data sets (if applicable).

There was broad agreement that reliability in the context of episode groupers should demonstrate that the grouper produces consistent results when the input requirements are met and the use case is constant.

Members of the panel expressed concern that the concept of reliability is challenging for episode groupers since the use case for different users may significantly impact the output or the grouping decisions. Similar to testing guidance for performance measures, the Panel did not want to be prescriptive in terms of testing approaches that could be used to demonstrate reliability. Some members of the Panel discussed testing options similar to NQF's recommendations for eMeasure testing. Specifically, the episode grouper could be applied to a simulated data set that includes sample patient data with the data and input requirements for the episode grouper. Since the simulated dataset is constructed, the patient's clinical experience is known. When the episode grouper is applied to the simulated data set, it should return consistent episode groups.

Validity

The Panel also agreed that that validity of the episode grouper should be assessed. There was general agreement that the evaluation of the episode grouper should include an examination of the known limitations of the grouper compared to its intended use, an evaluation of the clinical face validity, and an examination of the construct validity of the episode grouper.

Developers should discuss the limitations of their grouper methodology and provide adequate explanation of how they addressed the known limitations. Given the variety of methods that exist, it would be important for developers to disclose the real and perceived threats to validity that exist in the use of their product and how they have chosen to address these threats. For example, a developer may note that the episode grouper should be used with caution when it is used to discern utilization based on the type of pneumonia episode since the origin, community- or hospital-acquired, may not be captured systematically in the administrative claims data used to create the episode.

Developers should provide validity testing demonstrating that the grouper correctly reflects the cost of care or resources provided. The Panel discussed that validity could be demonstrated through an examination of clinical face validity and construct validity. Clinical face validity could be demonstrated by giving use cases and examine the performance of the grouper. For example, developers could examine a sample of patients and use clinicians to review how claims are assigned to individual episodes to ensure

episodes were clinically appropriate and that claims were assigned in alignment with the actual clinical course. This method would allow clinicians to recreate the medical history through the chronology of services from the administrative claims to evaluate treatment patterns, when conditions are resolved, when there is an exacerbation of a condition, and examine complications.

Further, developers could examine the construct validity of the episode grouper. There are many approaches to testing the construct validity depending on the methodology of the episode grouper. One approach may be to test the homogeneity of the episodes themselves, or whether the method accounts for risk in other forms. Another approach may test the construct validity of a grouper by assessing whether a case of pneumonia that is identified by the episode grouper, is in fact a case of pneumonia. This could be tested by examining the prevalence rate that the episode grouper expresses in the population and comparing those prevalence rates to prevalence rates noted in the literature.

2. Feasibility

The Panel agreed that a feasibility criterion should assess the extent to which the required data are readily available or could be captured without undue burden, and can be implemented for performance measurement. The Panel agreed that the subcriteria that may be used to understand the feasibility of an episode grouper would include an assessment of whether the required data elements are routinely generated during care delivery and an assessment of whether the required data elements are available in electronic sources. Finally, the Panel agreed that an evaluation should assess the financial burden due to the costs associated with the use of the grouper. This assessment should include cost-license fees, and the cost of propriety components required to run the grouper.

3. Usability and Use

The Panel discussed the goal of this criterion is to assess the extent to which potential audiences (e.g., consumers, purchasers, providers, policymakers) are using or could the episode grouper for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations. The Panel agreed that this criterion should include an assessment of the current and future or planned use of the grouper, in addition to an evaluation of the benefits of the grouper compared to the unintended consequences of the grouper.

Summary of Proposed Criteria

Principles for application of the criteria:

1. Episode grouper developers are required to provide the appropriate information in order to determine the extent to which each of the criteria has been met.
2. The application of these criteria requires both evidence and expert judgment.
3. Subcriteria delineate how to demonstrate that the major criteria are met.
4. The assessment of each criterion is a matter of degree (rather than all or nothing), generally rated on a scale of high, moderate, low, and insufficient.

Scientific Acceptability

The extent to which the grouper produces consistent (reliable), and clinically relevant (valid) episodes.

Reliability

- Reliability testing demonstrates that the episode groupings are repeatable, consistent results a high proportion of the time when assessed with the same data in the same time period (with input requirements met and use case constant).

Validity

- The intended use of the episode grouper aligns with the logic for grouping claims.
- Validity testing demonstrates that the episodes are clinically relevant and appropriate.
- Severity and risk adjustment strategy is clearly specified and is based on patient factors that influence the clinical course and assignment of claims.
- Threats to validity (i.e. limitations) are adequately described including how threats have been addressed.

Feasibility

The extent to which the required data are readily available or could be captured without undue burden, and can be implemented for performance measurement.

- Required data elements are routinely generated during care delivery.
- Required data elements are available in electronic sources.
- Demonstration that the data collection strategy can be implemented (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures).

Usability and Use

The extent to which potential implementers and potential audiences (e.g., consumers, purchasers, providers, policymakers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

- The intended use(s) of the episode grouper are clearly described.
- The planned use of the episode grouper is clearly described.
- The benefits of the use of the episode grouper outweigh any unintended consequences.
- The maintenance plan demonstrates adequate maintenance of the grouper to enable ongoing meaningful use of the output by users and implementers.

Criteria Not Recommended

1. Importance to measure and report

The Panel discussed the relevance of the importance to measure and report criteria but ultimately agreed that given the multiple uses and broad scope of episode groupers, it would be less relevant to evaluate this criterion. The Panel agreed that the developer should express their intended use of the grouper in the usability section of the evaluation.

2. Evaluation of Related or Competing Groupers

The Panel agreed that episode groupers include significant differences in methods and design making it challenging to compare methods. Further, the intended use of the grouper would have a significant impact on its design making comparisons between them inappropriate. The Panel ultimately agreed that this criterion should not be applied for the purposes of evaluating episode groupers.

Recommendations for the Application of Criteria and the Evaluation Process

The Panel generally agreed that these evaluation criteria should be applied at the grouper and episode level, and separate but similar criteria should be explored for use at the measure level. Recognizing the multiple uses and multiple appropriate methodologies that exist, the Experts agreed that the criteria should be applied to episode groupers by future NQF Steering Committees with an understanding that there is no one gold standard but rather multiple appropriate design options depending on the use case. The Panel encouraged future efforts are undertaken to examine the submission elements and criteria that should be used to evaluate measures that are a result of an episode grouper. Many on the Panel indicated that the current cost and resource use measurement submission elements and criteria may be a starting point. If a grouper is able to output multiple episodes, then each one should be evaluated to ensure that they are constructed as clinically homogenous episodes or the risk variation within an episode is sufficiently handled. Some members of the Panel expressed concern that this approach may overwhelm future Steering Committees, and may not allow smaller episode grouper developers to participate due to the time and resource burden of such a review.

Considerations for NQF Endorsement of Episode Groupers

When considering the types of episode groupers that could be brought forward, the Expert Panel agreed that not every use case of episode groupers is important for endorsement. Many episode groupers are used for internal business purposes and are extensively customized to that end. These uses would generally not rise to the level of requiring a review for endorsement. Potentially, the only use of an episode grouper that needs a national endorsement is the public episode grouper due to the broad use of the grouper and its potential impact.

Concern was raised by members of the Panel that this proposed endorsement process would discourage developers from participating with NQF if there wasn't a compelling reason to do so. This would then create downstream effects for the evaluation of cost and resource use measures based on commercial groupers that have not been through the endorsement process when compared against measures based on an endorsed episode grouper. The Panel was also concerned that an overly prescriptive endorsement process could block innovation in the field. New competitors would not be able to gain entry because they lack the resources to support the endorsement process. On the other hand, the Panel acknowledged that major employers could have a significant bearing on the value of NQF endorsement for episode groupers if they demanded that the episode grouper product they select be NQF-endorsed. In this case, commercial grouper developers would likely go through the endorsement process.

The Panel raised the issue of the fluid nature of the logic and definitions of episode grouper systems as it pertains to endorsement at a particular time. Episode grouper software is perpetually evolving and improved upon by developers as feedback is obtained from the end users. A consistent method for versioning groupers and tracking each version would need be developed and integrated into the review and endorsement-maintenance process. Further, a better understanding of what type of change or update to the grouper would require a version upgrade is unclear.

The Panel explored the difference between endorsing software and endorsing a methodology or logic. The Panel cautioned against NQF endorsing software, however, noting that it is often difficult to extract

certain pieces or logic from the overall grouper software application. In this instance, it would be helpful to parse out the clinical inputs for the grouper that are distinct and separate from the logic used in the software.

Public and Private Sector Alignment

The Panel agreed that any efforts to align the public and private grouping methodologies to obtain a single endorsed grouper present tremendous challenges with many unintended consequences. The field of episode grouping is continually evolving and conforming to a single methodology would stifle innovation. Additionally, the public payment system (Medicare) is quite different from many private payment systems, necessitating differing grouping methodologies. NQF seeks to endorse national standards that allow for comparisons across measured entities. Due to the inherent flexibility of many episode groupers and the ability for end users to customize the product to serve their own business purposes, the Panel agreed that it is not feasible to require that a grouper allow for national comparisons.

Multi-Stakeholder Input

The Panel expressed concern about whether NQF's current multi-stakeholder process for evaluating and endorsing performance measures could be used to evaluate episode groupers and episode grouper methodologies due to their complexities. Members of the Panel suggested that a technical review of the grouper by external experts be performed and provided to a multi-stakeholder group as input. In the case of the Medicare episode grouper, the Panel agreed that multi-stakeholder input was still essential because of the far-reaching impact of the grouper. In order to be acceptable to the various measured entities, they would need to have input during the evaluation and endorsement process. Members of the Panel also suggested that, short of endorsement, NQF could be involved in convening experts to create standard definitions for episodes that would then be used in the development of episode groupers. The common element for all episode groupers is that they have lists of codes that are assigned to any given episode. Variation in the definitions of these episodes and the codes assigned prevents any sort of meaningful comparison and could be rectified by the standardization of the episode definitions.

Quality Signal

The Panel supported the idea that a quality signal could accompany the cost signal in the output of an episode grouper. Evaluating costs independent of outcomes could lead to unintended consequences, such as sacrificing functional and medical outcomes to drive costs down. Members of the Panel suggested that many groupers could already produce several quality signals, including occurrence of post-operative infections, complications, and readmissions, among others. Administrative claims, the data source for many episode groupers, may prevent the development of robust quality measures and poses challenges; however, the ability to supplement this data with information from electronic data could eventually produce substantial quality signals along with the cost information.

Measure Applications Partnership (MAP) Input

The MAP is charged by HHS with providing input to CMS on pre-rulemaking by making recommendations for the inclusion and application of specific measures in various CMS programs. The Panel examined the necessary considerations for making decisions about the application of measures

based on episode grouper methodologies for federal programs. Many expressed concern that selecting individual measures for application without considering how costs were assigned to other co-occurring conditions may be misleading. Given that the process for attributing costs is not always clear, transparent, or understandable when considered in isolation of the entire system, the Panel encouraged the CMS Episode Grouper to undergo NQF endorsement both at the grouper level, episode level, and at the individual measure level prior to selection for use by the MAP.

Conclusion/Next Steps

This effort has highlighted the many challenges to expanding evaluation, and potentially endorsement, beyond individual measures to episode groupers. Given the expressed need of an evaluation of the public episode grouper, the Expert Panel agreed that this seems to be a more palatable starting point to serve as a learning opportunity to understand feasibility of applying the approach, criteria and submission requirements to other types of groupers. Commercial sector groupers have been in the market for a number of years, and many in the group did not see an explicit need for endorsement of these products at this time.

Further consideration will also be needed to determine whether endorsement of groupers is the appropriate path, rather than some other type of approval process—particularly given the implication that an endorsed tool would be a national standard. The variation in approaches and perspectives on grouping claims may make it challenging for establishing a standard method for approaching the evaluation of any grouper.

In order to fully implement this process, additional work will need to focus on further refining the criteria, submission elements, and clearly delineate a process for evaluation. With NQF's focus on measurement and performance improvement, subsequent efforts to explore the evaluation and use of groupers should focus on how the measures developed from grouper output, that are used to measure providers, can be evaluated and endorsed.

Notes

1. Hussey PS, Sorbero ME, Mehrotra A, Liu H, Damberg CL. Episode-based performance measurement and payment: making it a reality. *Health Affairs*, October 2009, Vol. 8, No.5. Available at http://content.healthaffairs.org/cgi/reprint/28/5/1406

2. Medicare Improvements for Patients and Providers Act of 2008. Available at http://www.gpo.gov/fdsys/pkg/PLAW-110publ275/pdf/PLAW-110publ275.pdf Last Accessed March 2014.

3. Patient Protection and Affordable Care Act. Available at http://www.gpo.gov/fdsys/pkg/BILLS-111hr3590enr/pdf/BILLS-111hr3590enr.pdf Last Accessed March 2014.

4. American College of Physicians. "Value-Based Payment Modifier" Available at http://www.acponline.org/advocacy/where_we_stand/assets/vii2-value-based-payment-modifier.pdf Last Accessed March 2014.

5. Robert Wood Johnson Foundation. "Medicare's Value-Based, Physician Payment Modifier: Improving the Quality and Efficiency of Medical Care" Available at https://www.academyhealth.org/files/HCFO/HCFOPhysicianValueModifier.pdf Last Accessed March 2014.

6. American Medical Association. "Call for Nominations to Participate in the CMS Episode Grouper Project" Available at http://www.ama-assn.org/resources/doc/cqi/pcpi-newsletter-061113.pdf Last Accessed March 2014.

7. National Quality Forum (NQF). Measurement Framework: Evaluating Efficiency Across Patient-Focused Episodes of Care. Washington, DC: NQF; 2009.

8. Solon JA, Feeney JJ, Jones SH, et al. Delineating episodes of medical care, *AM J Pub Health Nations Health,* 1967;57(3):401-408.

9. Levine M, email communication, March 31, 2014.

10. Hornbrook M, Hurtado A, et al. Health care episodes: definition, measurement and use, Medical Care Research and Review, 1985;42(2):163.

11. Hussey PS, Sorbero ME, et al.

12. Hornbrook M, Hurtado A, et al.

Appendix A: Expert Panel Roster

Kristine Martin Anderson, MBA (Co-Chair)
Booz Allen Hamilton, Rockville, MD

Joseph Cacchione, MD (Co-Chair)
Cleveland Clinic, Cleveland, OH

Stephen Bandeian, MD, JD
Agency for Healthcare Research and Quality (AHRQ), Rockville, MD

David Bodycombe, MSc, ScD
Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD

Francois de Brantes, MS, MBA
Health Care Incentives Improvement Institute, Newtown, CT

Dan Dunn, PhD
Optum, Waltham, MA

Nancy Garrett, PhD
Hennepin County Medical Center, Minneapolis, MN

Jennifer Hobart, MBA, MSc
Blue Shield of California, San Francisco, CA

David Hopkins, PhD
Pacific Business Group on Health, San Francisco, CA

Jim Jones, MBA
AmeriHealth Caritas Family of Companies, Philadelphia, PA

Marjorie L King, MD, FACC, MAACVPR
American Association of Cardiovascular and Pulmonary Rehabilitation - AACVPR, West Haverstraw, NY

Mark Levine, MD, FACP
CMS, Denver, CO

Jim Loiselle
McKesson Corp., Londonderry, NH

Thomas MaCurdy, PhD
Stanford University, Stanford, CA

Jelani McLean, PhD, MPA
Blue Cross Blue Shield Association, Chicago, IL

David Mirkin, MD
Milliman MedInsight, New York, NY

James Naessens, ScD, MPH
Mayo Clinic, Rochester, MN

David Redfearn, PhD
Independent Consultant, Las Vegas, NV

Andrew Ryan, PhD
Weill Cornell Medical College, New York, NY

Tamara Simon, MD, MSPH, FAAP
University of Washington School of Medicine; Seattle Children's Hospital, Seattle, WA

Christopher Tompkins, PhD
Brandeis University, Waltham, MA

NQF Staff

Helen Burstin, MD, MPH
Senior Vice President, Performance Measurement

Ashlie Wilbon, RN, MPH
Managing Director, Performance Measurement

Taroon Amin, MA, MPH
Senior Director, Performance Measurement

Evan M. Williamson, MPH, MS
Project Manager, Performance Measurement

Ann Phillips
Project Analyst, Operations

Elizabeth Carey, MPP
Project Manager, Strategic Partnerships

Appendix B: An AMI Episode

The figure below, developed as a product of the NQF Episodes of Care Measurement Framework report, illustrates the context for considering an AMI episode.