Evaluating Episode Groupers: A Report from the National Quality Forum

September 5, 2014

This report is funded by the Department of Health and Human Services under contract HHSM-500-2012-00009I Task Order 8.

NATIONAL QUALITY FORUM

Contents

Executive Summary

The United States spends more money on healthcare than any other country in the world.[1] This high rate of spending, however, has not resulted in better health for Americans. Episode-based performance measurement is one approach to better understanding of the utilization and costs associated with certain conditions by grouping care into condition-specific or procedure-specific episodes. For example, all diabetes-related care is grouped into a diabetes episode of care. Episode grouper software tools are a generally accepted method for aggregating claims data into episodes to assess condition-specific utilization and costs. Using an episode 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 used to manage that condition.

Episode grouper software products developed by different vendors use significantly different methods to group and attribute claims to episodes. The growing interest in the use of these tools to better understand healthcare costs, the limited transparency and inherent complexity of the methodologies employed, and the recent investment by the Centers for Medicare & Medicaid Services (CMS) to develop a publicly available episode grouper for the Medicare program have generated further interest in exploring the need for and implications of a multistakeholder consensus-based review of episode groupers.

With funding from the Department of Health and Human Services (HHS), the National Quality Forum (NQF), convened an Expert Panel to define the characteristics and challenges of constructing episode groupers; determine an initial set of criteria by which episode groupers should be evaluated; and identify implications and considerations for NQF endorsement of episode groupers. The Panel did not focus on a particular grouper or product. It instead recommended criteria that can be applied to any episode grouper that may be submitted for evaluation.

The Expert Panel recommended the following submission items for evaluation: descriptive information on the intent and planned use of the grouper; the clinical logic and data required for grouping claims; and reliability and validity testing. The Panel emphasized the importance of understanding the intent and planned use for evaluating potential threats to validity and possible unintended consequences of using the grouper.

The recommended evaluation criteria for episode groupers are based on the standard NQF Measure Evaluation Criteria, and include scientific acceptability (reliability and validity), feasibility, and usability and use. The Panel did not recommend the application of the importance to measure and report or related and competing criteria.

Further input from NQF's Consensus Standards Approval Committee (CSAC) affirmed the complexity of issues regarding the submission and evaluation of episode groupers. CSAC recommended that, while a "yes or no" vote for endorsement for episode groupers would be premature given the current state of the industry, there is a need for a qualitative, peer review process to evaluate them and facilitate transparency for stakeholders. The key elements of a qualitative, peer review process are outlined as a foundation for further work to shape the actual process that would be used in NQF's initial effort to

evaluate episode groupers. The evaluation criteria used in this review process would be based on the evaluation criteria recommended by the Expert Panel, and allow for a pathway toward full endorsement as the field matures. This effort has highlighted the many challenges of expanding evaluation beyond individual measures to episode groupers and the need for future work to explore the evaluation and endorsement of measures developed using episodes generated from groupers.

Introduction

In recent years, there has been a drive toward performance measurement based on the patient's episode of care has to better understand the utilization and costs associated with certain conditions. Measurement based on an episode of care 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.[2] To meet the growing demand for cost performance information, various measurement approaches have been developed for applications such as bundled payments, gain sharing, and other types of episode-based payment.

Episode grouping is one approach among many that has been used to measure costs across an episode of care. Namely in the commercial sector, episode grouper software tools have been evolving as a widely accepted method for aggregating claims data into episodes to assess condition-specific utilization and costs. Using a grouper allows healthcare services provided over a defined period of time to be analyzed and grouped by specific clinical conditions to generate an overall picture of the services utilized to manage that condition. Episode grouper software products developed by different vendors use significantly different methods to group and attribute claims to episodes. The growing interest in the use of these tools to better understand healthcare costs, the limited transparency and inherent complexity around the methodologies employed, and the recent investment by the Centers for Medicare & Medicaid Services (CMS) to develop a publicly available episode grouper for the Medicare program has generated further interest in exploring the need for and implications of a multistakeholder consensus-based review of episode groupers.

The Policy Landscape

Maintaining Medicare viability in a climate of rising healthcare 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. Although there has been a movement toward alternative payment models that shift the system away from fee-for-service, 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.[3]

The Physician Resource Use Measurement and Reporting Program was extended and enhanced by the Patient Protection and Affordable Care Act (ACA) passed in 2010, and was renamed the Physician Feedback Program. The Physician Feedback Program includes two primary components; first, the Physician Quality and Resource Use Reports (QRURs), and second, development and implementation of a value-based modifier (VBPM). The ACA broadened the scope of QRURs by requiring CMS to develop a publicly available episode grouper with specific functional requirements by January 1, 2012, with the intent that it would be submitted for multistakeholder review and endorsement, as appropriate.[4]

It is anticipated that the episode-based measurement produced from the grouper will also be used to support the Physician Feedback Program's development and implementation of a VBPM.[5] Beginning this year, the VPBM, administered by CMS, will require groups with 10 or more eligible professionals to register and report through the Physician Quality Reporting System to avoid a payment penalty in 2016 under the value modifier. Practices that do not meet this requirement can still avoid penalties if at least 50 percent of individual physicians participate within the group. Physicians can review the results of the Physician Feedback Program in their Quality Resource Use Reports (QRURs), allowing for comparisons among physicians and groups on quality and resource use.[6] Based on the legislative mandate, CMS will begin phasing in VPBM reimbursement on January 1, 2015, with the goal of applying the modifier to all physicians who bill Medicare using the fee schedule by 2017. Payment in 2017 will be determined based on the performance of registered groups during 2015.[7]

Project Overview

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 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. 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.

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 5 and 6, 2014, to discuss the key issues identified above and provide recommendations on the evaluation of episode groupers. The Panel did not focus on a particular grouper or product. It instead recommended criteria that can be applied to any episode grouper that may be submitted for evaluation.

Current Landscape of Episode Groupers

Public Grouper

In response to the legislative mandate to create a publicly available grouper for Medicare, CMS began 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-convened 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.[8]

Commercial Groupers

Several commercial episode groupers 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 used these tools to examine prevalence rates for various conditions, incidence rates for various treatments, and complication rates to support internal quality improvement. Purchasers also have used groupers to understand provider utilization and cost variation.

Public- and Private-Sector Alignment

The use of commercial grouper products often varies by market; even within a market, stakeholders may have invested in different products based on their specific needs and preferences. Although the groupers perform similar functions, their approach to grouping claims varies and thereby limits the comparability of results by users of various systems. Also, the data used 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

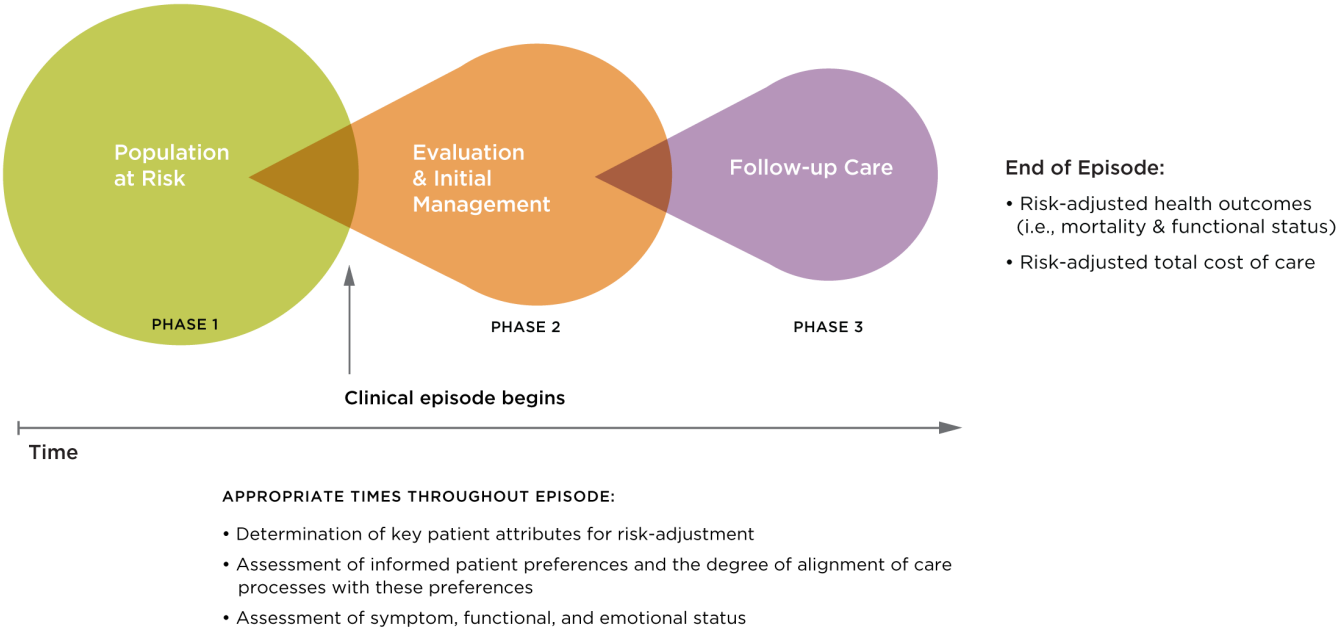
layer of complexity and misalignment of the existing tools. Although the groupers are not necessarily limited to grouping claims for a particular age range, further testing would be required to determine the appropriateness of the groupers for use with data from across the lifespan, which may be beyond the primary scope and intended use.

Defining Episodes

The concept underlying most episode groupers is the episode of care. Recognizing that 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."[9] These healthcare services can be administered by one or more providers over the course of the episode.[10] Using an episode-based approach to performance measurement can 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).

Figure 1, developed as a product of the NQF Episodes of Care Measurement Framework report, illustrates the framework by providing a conceptual model for how an individual moves through the various stages of an illness. The model outlines three phases of an episode of care starting with a population at risk for a given disease. The next phase, evaluation and initial management, is characterized as the phase in which a clinical episode begins and the treatment and management of a newly diagnosed acute or chronic disease is provided. It is in the follow-up phase when long term disease management begins. Although this model is illustrated in a linear fashion, depending on the condition to which it is applied and the clinical course of that condition, an individual may move bi-directionally between phases, for example from follow-up care back into the evaluation and treatment phase. Particularly relevant for Medicare beneficiaries, this model could also be applied to an individual with multiple conditions who may be in a different phase for each condition.

Figure 1. Episode of Care Conceptual Model



Using this model to understand acute episodes, an acute event, 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.

In order to designate a time period for an underlying chronic condition (e.g., CAD) which clinically does not end, a specified period (e.g., a 12-month window) is generally defined to capture the healthcare services related to the treatment of that condition, but this specified period may be unrelated to the clinical course of the disease. For an acute episode (e.g., broken arm), the start and end dates are generally distinct points in time starting with the event when the broken arm occurred until the arm is healed. This type of discrete time period for defining episodes can also be applied for procedures, where episodes are focused on a precisely definable scope of intervention and the follow-up care related to only that procedure.

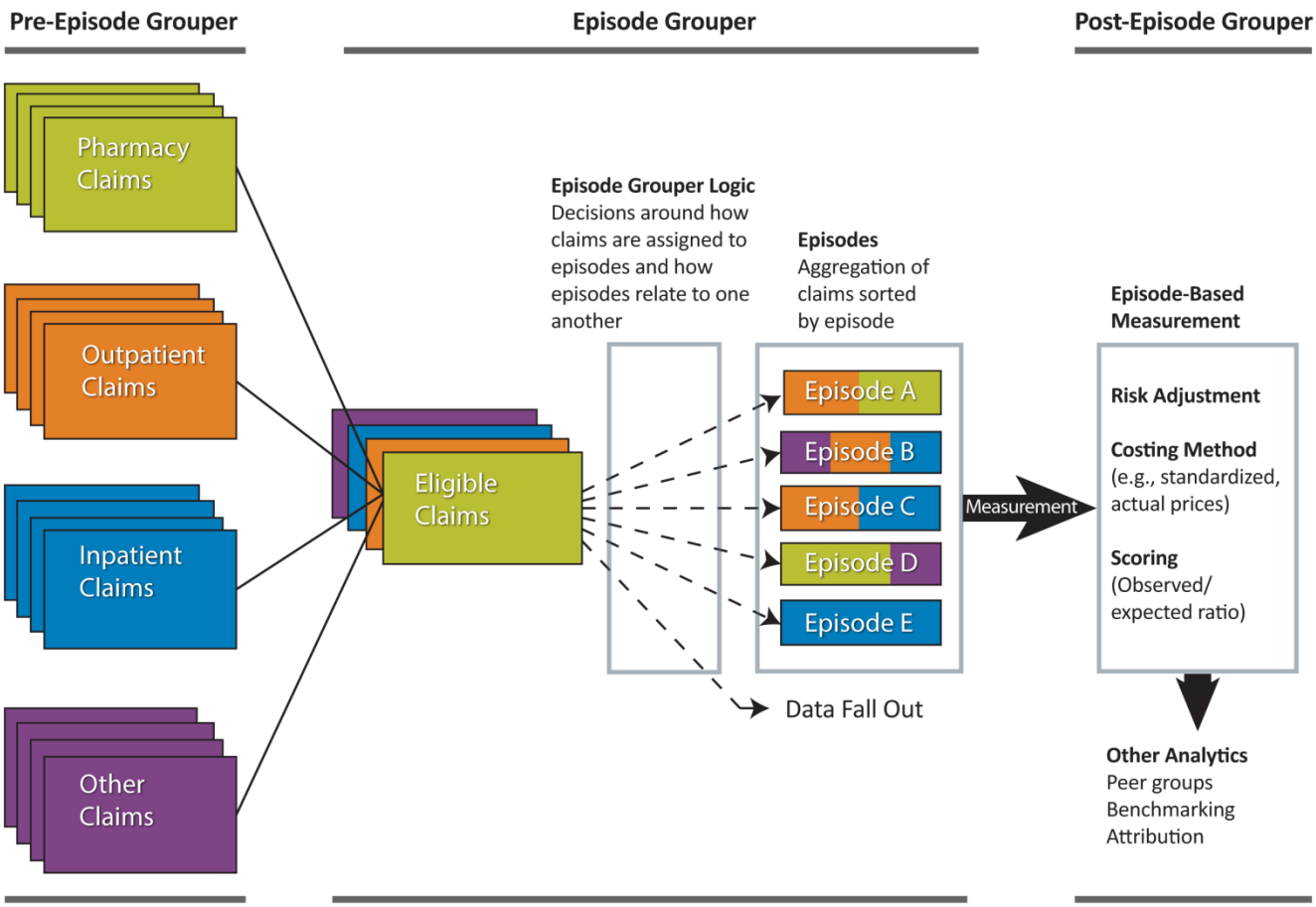
One of the major challenges in defining episodes is determining when and how to attribute services for the treatment of conditions that occur as complications of the underlying condition or procedure. During an episode for a given clinical condition, any series of complications could develop; these complications 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 resources used for the complication(s) to the underlying condition.

Understanding Episode Groupers

Episode groupers can be defined as the software and logic that assign patient claims representing their utilization of healthcare services to clinically relevant episodes of care. Episode grouping is operationalized using these 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 explore the use of electronic health record data and clinical registries to understand and capture healthcare utilization that would not be captured in administrative claims data. By examining the utilization patterns for a condition, the dollar amount assigned to each claim in an episode can be aggregated to understand total cost for an episode of care for a condition or procedure. Many groupers have the ability to create hundreds of condition-specific episodes. The creation of these episodes depends on the intricate decision logic that determines to which episode a claim should be assigned.

Figure 2 illustrates the basic function of an 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. During grouping (assignment of claims to clinical episodes), logic can be applied for addressing risk and severity, determining inclusion and exclusions at both the patient and service levels, and addressing threats to validity. Once the claims are aggregated into clinical groupings, or episodes (e.g., Episode A, Episode B, etc.), analysis of the episodes post-grouper may occur. Post-grouper analysis may include analysis of resource utilization, profiling, identification of cost drivers and opportunities for improvement, and highlighting variability of services and examining patient care pathways.

Figure 2. Illustrating Episode Grouping



Using the grouper, a user is able to capture costs for multiple conditions at once, enabling the analysis of multiple concurrent clinical episodes during the same time period. Using 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 grouper would assign the blood sugar check to the diabetes episode of care and the heart imaging study to the heart failure episode.

While there is a desire to ensure that the cost for "appropriate" care can be captured, most groupers do not solely count claims or resources used based on "appropriate" care that would be outlined in guidelines, but rather, sort all eligible claims and attribute them based on their relevance to the clinical condition. The use of clinical guidelines is likely most relevant in the post-grouper phase of grouping claims when the costs for delivering care can be attributed to various providers based on user-defined rules and additional analysis is allowed, which could include the identification of those claims that represent clinically appropriate care based on guidelines.

Accounting for the variation in risk and severity of clinical condition of the patients within the data is an important step in episode grouping that may be done within the grouper logic or in the post-grouper

analysis where it is user-defined. The purpose of risk or severity adjustment is to account for the patient-related clinical factors that exist prior to the patient's encounter with the provider being measured. This adjustment ensures that providers are accurately being measured on outcomes or processes that they can reasonably influence, rather than underlying differences in patient severity. There are also some inherent limitations in the handling of risk in the development of episodes. Given that many episode groupers currently use administrative claims data, 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).

There may be multiple strategies for handling the issue of risk. First, the grouper can stratify patient risk through the grouping mechanism by creating new episodes for increased risk. However, this approach presents many of the same challenges and trade-offs regarding narrowly defined episodes. Secondly, groupers may offer supplementary risk modules that can be applied after the grouping function is completed. Again, these different approaches may be appropriate depending on the intended use of the grouper given that the developer is transparent about the design and rationale.

Approaches to Episode Grouping

Most episode groupers employ a patient-centric approach to grouping episodes using the patient's experience as the framework for triggering a clinical episode and assigning claims to clinical groupings. This approach enables the analysis of patient care for a specified condition across all providers, settings, and interventions throughout the episode to better understand gaps in care coordination and care integration. Attribution of costs associated with utilization to specific providers often occurs post-grouper and is designed around the user's needs, specific application, and intended use.

An alternative to the patient-centered episode of care approach, identified by some members of the Panel, is a provider-centric approach. The primary purpose of provider-centric episodes is to profile the resource use of individual providers. In contrast to the patient-centric approach, attribution to providers is a central focus in the assignment of claims and construction of provider-centric episodes. The underlying goal of provider-centric episodes is to group a set of services and outcomes for individual patients that an identifiable category of providers can credibly influence and be held accountable for, and to facilitate reporting resource use in a manner that clinicians understand.[11]

Evaluating Episode Groupers

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 transparent 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 three phases: 1) the grouper logic itself, 2) the episodes, or groups of claims, resulting from the application of the grouper decision logic, and 3) the individual measures that are developed based on the episodes resulting from the grouper using similar 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 grouper decision logic should be designed with the intent of creating episodes that reflect the patient experience.
5. Episode grouper output should be actionable and usable for performance improvement and resource use transparency.
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 a grouper's capabilities or decision logic should be; the methodologies underlying the various episode groupers may have distinct approaches that may each be valid.

One of the key goals of this effort is to identify criteria by which episode groupers can be evaluated. The Expert Panel began the process of considering evaluation criteria for episode groupers by reviewing the existing NQF resource use measure evaluation criteria, and other literature, such as the 2008 AHRQ report to identify, categorize, and evaluate healthcare efficiency measures.[12] The Panel 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, Feasibility, and Usability and Use*. A fifth criterion, *Related and Competing Measures*, is applied as needed to measures that have been identified with similar measure specifications.

The Expert Panel recommended three major criteria that should be used to evaluate episode groupers by future multistakeholder panels: Scientific Acceptability, Feasibility, and Usability and Use. In order to determine whether the aforementioned criteria have been met, the Panel also identified the key elements of an episode grouper that would be require review in order for evaluators to gain a full understanding of the requirements, rationale, approach, testing, and maintenance of the tool. These evaluation criteria would be applied both at the grouper and episode levels. Future efforts should explore the applicability of the current Resource Use Measure Evaluation Criteria to measures developed using episode groupers. Recognizing the multiple uses and methodologies that exist for grouping, the three major criteria identified for episode groupers should be applied 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. Given the array of episode grouper methodologies in the field, the experts agreed that it would be difficult to determine up front what type of outputs should be expected from any particular episode grouper. Evaluators and users of the grouper should be able to ascertain how the design of the grouper aligns with its intended use. As with other

NQF-facilitated multistakeholder evaluation processes, this evaluation process should emphasize full transparency during the submission and evaluation process.

The information that would be required for submission of a grouper for evaluation could closely mirror those elements currently requested in the cost and resource use measure evaluation process. While the current process focuses on individual measures, the evaluation of an episode grouper would need to take into account several, if not hundreds of, possible clinical episode groupings to assess whether these groups appropriately reflect the underlying clinical construct they intend to measure.

Recommended Evaluation Criteria and Associated Submission Elements for Consideration

In order to understand the context for which the grouper has been developed and the key underlying assumptions driving its decision logic, a set of submission elements reflecting descriptive information should be required. First, episode grouper developers should be clear about the grouper's specific purpose and intended use; this includes a description of the core capabilities (e.g., number and type of episodes produced, risk, and severity adjustments). Because there are important implications for the application of episode groupers for payment and provider profiling purposes, information on the intended use is critical for evaluation.

As part of the pre-grouper functionality, it is vital for the user to understand the data requirements to optimally run the grouper. Specifically, data loss or data fallout can be a challenge for users when implementing an episode grouper. The loss of data may be due to at least two different issues. First, ungrouped claims or records identified as unrelated to the episodes being captured. The ability of users to evaluate these ungrouped claims would help them to better prepare the data before entering into the grouper and provide clarity for expected output. Second, missing charges associated with the claims or missing data elements required to create complete episodes may result in data fall out. Due to the impact of this data fall out on potential analytic capabilities, it is important to have transparency on the beginning-to-end data flow. This data flow should include the input requirements or input data required to run the grouper appropriately, and the proportion of data that are lost at each processing step. As such, developers should be required to describe any specific data completeness requirements that would impact the anticipated output of the groupings.

1. Scientific Acceptability of the Episode Grouper

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

Reliability of the episode grouper is a key criterion against which groupers should be evaluated. In order to demonstrate reliability, developers should be required to present testing results that demonstrate that the episode grouping results are repeatable. Specifically, 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. The Expert Panel recognized that the concept of reliability is challenging for episode groupers because the use case for different users may significantly

impact the output or the grouping decisions. The Panel preferred to be less prescriptive in recommending testing approaches that could be used to demonstrate reliability; however, some examples of possible testing approaches were discussed.

As one example of reliability testing, the developers could demonstrate how the episode grouper performs across multiple data sets as applicable, with a focus on using different types of data sets of varying sizes. Another possible approach focuses on 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. Because 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 evaluation of the validity of an 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 (i.e., relevance of the assigned claims to the clinical episode group), and an examination of the grouper specifications. This examination of grouper specifications should include core components of the grouper, including the clinical episode construction, approach to addressing risk and patient severity, and their testing approach.

CLINICAL EPISODE CONSTRUCTION

The construction of the clinical episodes is driven by the clinical logic that supports the purpose and conceptual framework for the episode grouper; it presents many challenges that should be weighed by developers and ultimately described and supported with a rationale in their submission for multistakeholder evaluation. The key elements of the clinical episode construction include a discussion of, and rationale for, the 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 and the rules for how claims are assigned to episodes.

Identifying the codes that trigger the start of an episode is one of the first decisions related to the clinical episode grouping. These trigger codes determine when the episode begins and what type of clinical episode should be started. The sensitivity of the triggers used to open an episode is an important consideration for the evaluation of an episode grouper, as it enables the assessment of whether adequate consideration to the opening of significant numbers of "phantom episodes" (i.e. episodes that are erroneously created with limited claim assignment) have been triggered. The creation of phantom episodes may bias the cost observed, both within the phantom episode itself, and in other episodes to which those claims could have been assigned.

Another key element of creating clinical episode logic is defining and describing the rules for how claims are assigned to clinical episodes.[13] These episode definitions should be developed in consultation with clinical experts, and be reviewed and updated regularly. The approach for assignment of claims to an episode is likely going to be different between groupers, based on use case 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 across multiple settings is a challenging endeavor.[14,15] Given that many patients, particularly Medicare beneficiaries, have multiple co-occurring conditions, 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," in which the logic is designed to force decisions on which episode a claim is assigned based on the relevance of other claims or data on the claim line.

Because multiple approaches to grouping can be employed, developers must also weigh the challenges and consequences of defining episodes narrowly or broadly. 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 be narrowly defined, where related services or procedures are grouped to their own episode. In the example above, AMI may be evaluated without the cost of any related procedures, and PCI and CABG costs are examined independently in their respective episodes. There is an inherent trade-off between tightly defining an episode so that there is greater clinical homogeneity among the patients within an episode, and generating sufficient sample sizes within each episode to enable reliable and valid inferences of resource use.

ADDRESSING RISK AND PATIENT SEVERITY

The criterion for evaluating risk assessment is intended to assess the approach and rationale for risk and severity adjustments made within the grouper logic. This requires that developers provide a description of their method for assigning risk, including any hierarchies and the underlying logic. If the grouper adjusts for risk using a risk adjuster, a description of the model, including the risk factors and data demonstrating performance of the model (i.e., adequate calibration), is required. The issue of addressing sociodemographic factors, particularly when measuring resource use among certain populations is a vital one. Any adjustments built into the grouper intended to address sociodemographic factors should be fully described with a rationale supporting its conceptual and empirical relationship to resource use. NQF does not require that developers use specific methods and approaches; however, what they develop and submit should be evaluated to determine if it is adequate and aligns with the intended use of the grouper.

TESTING

In order to demonstrate the properties of a valid grouper, developers should have tested the validity of the clinical logic in the episodes and the grouper as a whole. To evaluate this testing, transparency about the testing methods, the results derived from their approach, and the rationale for the clinical decision logic are vital. Given the various methodologies and trade-offs required in grouping claims to episodes, the testing approach submission requirements and methods for testing should not be prescriptive. Rather, developers should be transparent about how they handled trade-offs, potential threats to validity, and how those potential threats were addressed.

At minimum, a systematic assessment of the face validity of the grouper could suffice, although empirical demonstration of validity would be preferred. Approaches to demonstrating clinical face validity could include the examination of the performance of the grouper through the implementation of clinical use cases and validation by clinicians. An example would be identifying a sample set of

patients and asking clinicians to review how claims are assigned to individual episodes in order to validate that episodes contained clinically appropriate patients and claims and that claims were assigned in alignment with the actual clinical course. This method would enable clinicians to recreate the medical history to examine treatment patterns, understand when conditions are resolved, identify when there is an exacerbation of a condition, and evaluate the rate of complications.

Validity could also be demonstrated through an assessment of construct validity. This type of validity testing is intended to demonstrate that the grouper correctly reflects the cost of care or resources provided. One approach to construct validity testing could test the consistency of the clinical appropriateness of the claims assigned to the episodes. Another approach could test the construct validity of a clinical episode, like pneumonia for example, by assessing whether a case of pneumonia that is identified by the episode grouper is in fact a case of pneumonia. This could be validated by examining the prevalence of the condition that the episode grouper expresses in the test population compared to the rates noted in the literature.

Given the variety of methods for grouping that exist, it is key for developers to disclose both real and perceived threats to validity that have been identified in the development and testing the grouper and how those threats have been addressed. Similarly, any limitations of the grouper as designed should also be transparent to users and evaluators such that they are explained and there is a rationale for how those limitations can be mitigated or addressed. 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 because the origin (i.e., community or hospital acquired) may not be captured systematically in the administrative claims data used to create the episode. Additionally, the developer may note whether all pneumonias are considered in a single episode or if there is discrimination based on the site of care (inpatient versus outpatient). The submission elements should include a discussion of the limitations of the grouper related to both the design decisions made in grouping and limitations in the underlying data. For example, due to the limited staging information for cancer patients in administrative claims data, a grouper may not optimal for profiling episodes of cancer patients.

2. Feasibility

The Panel determined that the feasibility criterion is relevant for episode groupers. 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. In particular, given the commercial nature of many episode groupers, an evaluation of the financial burden due to the costs associated with the use of the grouper needs to be assessed. This assessment should include cost-license fees and the cost of proprietary components required to implement and run the grouper.

3. Usability and Use

The goal of this criterion is to assess the extent to which potential audiences (e.g., consumers, purchasers, providers, policymakers) are using or could use 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. Given the burden of implementation the complexity will be an important consideration to weigh in conjunction with the other criteria.

Additionally, the maintenance of an episode grouper was identified as a key element for consideration in determining its usability. Maintenance of an episode grouper requires diligence and a comprehensive understanding of the relationship of the various elements of the software and decision logic. Thus, developers should provide detailed information on their process for keeping their system current, including a plan and costs for upgrading new versions of the grouper. There was, however, a recognition that rapid upgrades and ongoing maintenance may be challenging to keep up with and introduce additional costs for users.

Criteria Not Recommended

1. Importance to Measure and Report

The Panel discussed the relevance of the *Importance to Measure and Report* criterion but ultimately agreed that given the multiple uses and broad scope of episode groupers. An evaluation of this criterion would not be informative or useful.

2. Evaluation of Related or Competing Groupers

The Panel agreed that episode groupers are substantially different in method, design, and intended use making it challenging to compare them. The Panel ultimately agreed that this criterion should not be applied for the purposes of evaluating episode groupers.

A summary of proposed criteria and the submission elements that will be requested for evaluation are included in Appendix C.

Considerations for NQF Endorsement of Episode Groupers

Considering the types of episode groupers that could be brought forward, the Expert Panel agreed that not every use case or type of episode grouper may warrant endorsement. Many episode groupers are used for internal business purposes and are extensively customized to that end. Such groupers would not benefit from NQF endorsement.

Members of the Panel were concerned that this proposed endorsement process would discourage developers from participating with NQF if there was not a compelling reason to do so. This could 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 may not be able to gain entry because they lack the resources to bring their product through the endorsement process. On the other hand, the Panel acknowledged that other stakeholders (e.g., health plans, employers, clinicians) could have a significant role in reinforcing the value of NQF endorsement for episode groupers if they strongly support the notion that the episode grouper product they are measured by or use be NQF-endorsed.

Another concern raised was the juxtaposition of the fluid nature of a grouper's decision logic and definitions compared to the stagnant nature of endorsement. Currently endorsement is limited to performance measures which are generally associated with the version of a measure that was reviewed and endorsed. 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 to 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 would be needed.

Episode groupers are unique in that they are built to allow for user inputs and enable flexibility. Given the variability of data inputs, user specifications, and use cases, a single episode grouper can have significant variability in its outputs (i.e., episodes and related measures) depending on the user. NQF endorsement has traditionally represented the consensus-based, multistakeholder approval of a standard that can be implemented consistently and that is comparable across measured entities. Based on this characterization, the endorsement of episode groupers would challenge this notion, as it would be difficult to capture and endorse a grouper based on a single user's preferences.

The Panel explored the difference between endorsing software and endorsing a methodology or logic and cautioned against NQF endorsing software, noting that it is often difficult to extract certain pieces of logic from the overall grouper software application.

Fostering Public- and Private-Sector Alignment

Efforts to align the public and private grouping methodologies to obtain a single endorsed grouper present tremendous challenges with potentially 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, potentially 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, it may not be feasible to require that a grouper allow for national comparisons.

Linking with a 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 quality outcomes to drive costs down. Given the various approaches to episode grouping, the link to quality measures may not always be done within the grouper system. However, many groupers generate some 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; however, the ability to supplement these data with other electronic sources such as EHRs could eventually produce substantial quality signals along with the cost information. The Panel strongly supported the notion that linking episode grouper-based measures to quality measures should continue to be a goal.

Implications for Measure Applications Partnership (MAP) Input

MAP is charged by HHS with making recommendations for the inclusion and application of specific measures in various CMS programs during the pre-rulemaking cycle. 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 cost/resource measures from an episode grouper for application without considering how costs were assigned to other co-occurring episodes or 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 multistakeholder review at the grouper level, episode level, and at the individual measure level prior to selection of measures by MAP for use in federal programs.

Implementing NQF Evaluation of Episode Groupers

Using guidance from the Expert Panel and NQF's Consensus Standards Approval Committee (CSAC), an NQF governing body charged with providing guidance and oversight of NQF's processes and policy, the following guidance was developed to provide a framework for further defining NQF's future process for the evaluation of episode groupers. Based on directives from CSAC, NQF's first foray into the evaluation of episode groupers will not be for the purposes of endorsement, but rather to facilitate a qualitative peer review process focused on providing transparency to all stakeholders of the construction of submitted groupers through the application of the criteria previously discussed.

Many of the challenges with evaluating episode groupers have been discussed, but the implementation of a multistakeholder review process for these complex systems also presents challenges. In particular, the Panel expressed concern that the stakeholders traditionally convened by NQF are volunteers, but given the volumes of data and information that would be required to evaluate an episode grouper, the inherent complexity of the tools, potential time limitations and the limited pool of expertise for this work, the burden on those participating in the process could be overwhelming. In particular, the capacity of both NQF staff and volunteer experts should be a major consideration when implementing this process to ensure there is sufficient infrastructure and support to adequately facilitate the process. In order to mitigate these concerns, the process for evaluating episode groupers may require some additions to NQF's standard process for evaluating performance measures.

In preparation for convening expert and multistakeholder bodies for the evaluation of episode groupers, it will be vital to establish a plan for training and educating each of the volunteers and the developer organization(s) on the evaluation process, application of the criteria, and the complexities of episode groupers. Given the need for clinical, technical experts, and multistakeholder input, there will be varying degrees of training needs to address both the process and grouper evaluation. As the process for evaluation evolves, the multistakeholder panel should be supplemented with a technical expert panel to review of the grouper functionality. The charge of each of these bodies and the flow of inputs to evaluation will need to be determined.

Soliciting Episode Groupers

NQF's standard process for soliciting measure submission is to publicly solicit participation from organizations via a call for measures. This approach promotes transparency and enables any organization that wishes to participate to self-select into the process. This approach, however, will not be conducive to the evaluation of episode groupers given the limitations in capacity for both volunteers and NQF to facilitate the review of multiple groupers simultaneously. Further, as previously discussed, the volume and complexity of the information that will be submitted will require substantially more time and effort than a typical measure evaluation process. Therefore, CSAC recommended that NQF's initial effort to evaluate groupers should be limited to one grouper, such as the CMS public episode grouper. Once a process has been vetted through the first effort, further consideration will be given to improving the process and whether additional groupers should be solicited going forward.

Grouper Evaluation Approach

The evaluation process for episode groupers will require a collaborative effort that includes NQF staff, the developer organization, and the volunteer bodies, each providing relevant inputs. While this process would primarily focus on reviewing and sharing qualitative assessments of the grouper based on the criteria, there remains a major role for NQF staff to facilitate this process to ensure that the flow of information is clear, complete, accurate, and timely. Consideration should also be given to the vehicle (i.e., electronic submission form) and infrastructure required to collect the information from developer organizations in a format that is conducive to multistakeholder review. Additionally, the need for technical assistance or other resources will be required for developer organizations to submit their highest quality application for review into a new process. Given the developers' familiarity with the grouper product, there would be an expectation that they would also play a large role in educating and orienting the various evaluation bodies to the construction and functionality of their grouper.

One approach for consideration would be to first convene technical expert panels comprised of both clinical and technical experts, to review select groups of episodes or conditions (e.g., cardiac conditions, pulmonary conditions) and the associated grouper and clinical decision logic. Clinical and technical experts from existing NQF standing committees could be drawn upon to participate based on expertise and interest in evaluating episode groupers. A second body of stakeholders, including consumers, purchasers, health plans, and others, could be convened to discuss the inputs of the technical and clinical experts' application of the criteria and further contribute a multistakeholder perspective to the analysis. While the process will not result in a final recommendation for endorsement, the current Resource Use Standing Committee, which has had extensive experience evaluating cost and resource use measures over several cycles, could oversee the process and be tasked with reviewing and weighing all inputs, including, developer feedback, and distilling the key issues for a qualitative peer review that would be shared for public and member commenting.

Process Oversight

Recognizing that this would be NQF's first effort to evaluate episode groupers, there will be a deliberate effort to monitor and evaluate the process. In addition to soliciting feedback from those participating in the process, a summary of the process, lessons learned, challenges, and recommendations for a path

forward will be shared with the CSAC who will make final recommendations on the future efforts of NQF to evaluate additional groupers.

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 CMS public episode grouper, the Expert Panel generally agreed that CMS grouper would be a palatable starting point to serve as a learning opportunity to understand the 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; however, this would not preclude their participation in future NQF efforts in this arena.

In order to fully implement this process, additional work will need to focus on further refining the criteria and submission elements, and clearly delineating 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 an episode grouper can be evaluated and endorsed.

Endnotes

1. Romley JA, Jena AB, Goldman DP. Hospital spending and inpatient mortality: evidence from California: an observational study. *Ann Intern Med.* 2011;154(3), 160-167.

2. Hussey PS, Sorbero ME, Mehrotra A, et al. Episode-based performance measurement and payment: making it a reality. *Health Aff (Millwood).* 2009;28(5):1406-1417.

3. Medicare Improvements for Patients and Providers Act of 2008. Pub L. No. 110-275, 110th Cong (2008).Available at http://www.gpo.gov/fdsys/pkg/PLAW-110publ275/pdf/PLAW-110publ275.pdf. Last accessed March 2014.

4. Medicare and Medicaid Extenders Act of 2010. Social Security Act, Vol. II, Pub L. No. 111-309, 111th Cong (2010). Payment for Physicians' Services. § Section 1848(n)(9)(A)(B)(C)(D) (2010). Available at http://www.gpo.gov/fdsys/pkg/PLAW-111publ309/pdf/PLAW-111publ309.pdf. Last accessed September 2014.

5. Patient Protection and Affordable Care Act. Pub L. No. 111-148, 111th Cong (2010). Available at http://www.gpo.gov/fdsys/pkg/BILLS-111hr3590enr/pdf/BILLS-111hr3590enr.pdf Last accessed March 2014.

6. American College of Physicians (ACP). *Value-Based Payment Modifier*. Philadelphia, PA: ACP; 2013 (unpublished.) Available at http://www.acponline.org/advocacy/where_we_stand/assets/vii2-value-based-payment-modifier.pdf Last accessed March 2014.

7. Robert Wood Johnson Foundation. *Medicare's Value-Based, Physician Payment Modifier: Improving the Quality and Efficiency of Medical Care*. Washington, DC:AcademyHealth; 202. Available at https://www.academyhealth.org/files/HCFO/HCFOPhysicianValueModifier.pdf. Last accessed March 2014.

8. American Medical Association. Call for Nominations to Participate in the CMS Episode Grouper Project. *Physician Consortium for Performance Improvement Newsletter*. June 11, 2013. .

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

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

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

12. McGlynn, EA. *Identifying, Categorizing, and Evaluating Health Care Efficiency Measures. Final Report*. Rockville, MD: Agency for Healthcare Research and Quality; 2008. AHRQ Publication No. 08-0030. Available at http://www.ahrq.gov/research/findings/final-reports/efficiency/efficiency.pdf. Last accessed July 2014.

13. Hornbrook MC, Hurtado AV, Johnson RE. Health care episodes: definition, measurement and use. *Med Care Rev*. 1985;42(2):163-218.

14. Hussey PS, Sorbero ME, Mehrotra A, et al. Episode-based performance measurement and payment: making it a reality. *Health Aff (Millwood).* 2009;28(5):1406-1417.

15. Hornbrook MC, Hurtado AV, Johnson RE. Health care episodes: definition, measurement and use. *Med Care Rev*. 1985;42(2):163-218.

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
Cheif Scientific Officer, Quality Measurement

Ashlie Wilbon, RN, MPH
Managing Director, Quality Measurement

Taroon Amin, MA, MPH
Senior Director, Quality Measurement

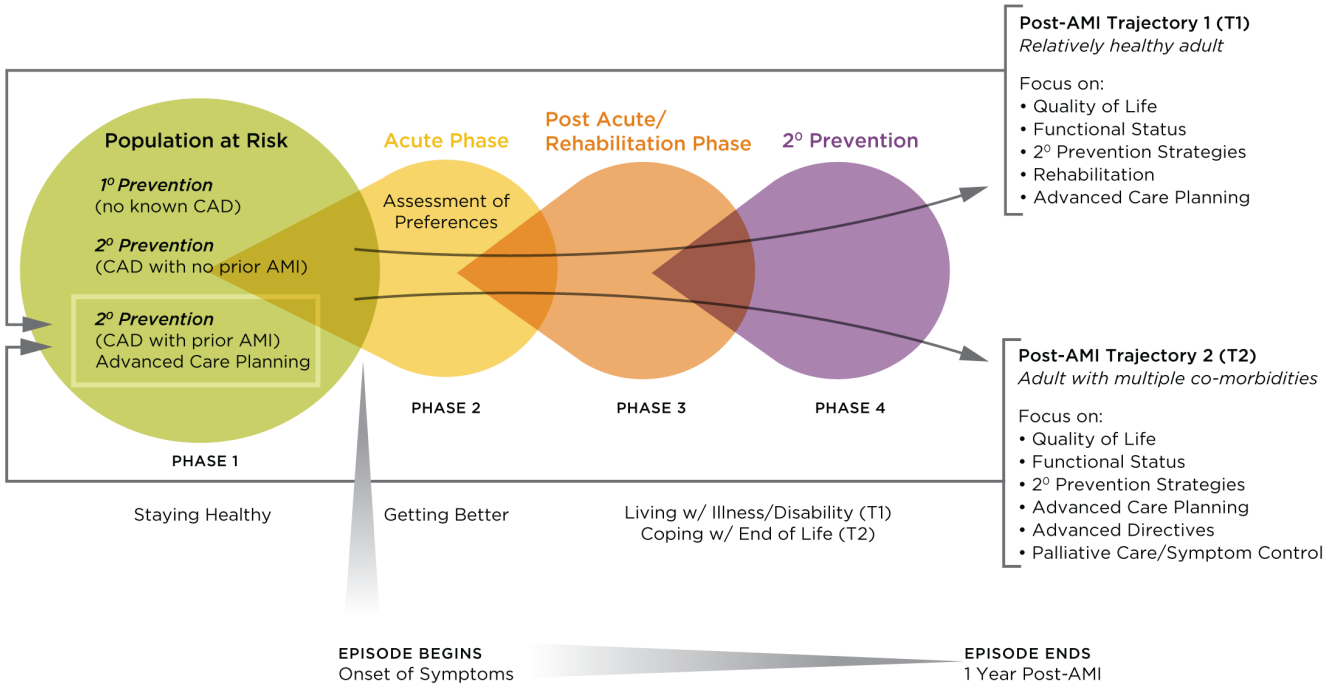
Evan M. Williamson, MPH, MS
Project Manager, Quality Measurement

Elizabeth Carey, MPP
Project Manager, Quality Measurement

Ann Phillips
Project Analyst, Quality Measurement

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.



Appendix C: Summary of Recommended Episode Grouper Evaluation Criteria

Principles for Evaluating Episode Groupers

1. The episode grouper output should be transparent 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 three phases: 1) the grouper logic itself, 2) the episodes, or groups of claims, resulting from the application of the grouper decision logic, and 3) the individual measures that are developed based on the episodes resulting from the grouper using similar 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 grouper decision logic should be designed with the intent of creating episodes that reflect the patient experience.
5. Episode grouper output should be actionable and usable for performance improvement and resource use transparency.
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 a grouper's capabilities or decision logic should be; the methodologies underlying the various episode groupers may have distinct approaches that may each be valid.

Recommended Criteria

Scientific Acceptability

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

Reliability

- The grouper specifications are clear, facilitate understanding and enable consistent implementation by the user.
- Reliability testing demonstrates that the episode groupings are repeatable, with 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 with factors that demonstrate a conceptual and empirical relationship to the episode(s) being measured.
- 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 grouper output 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.

Appendix D: Recommended Episode Grouper Evaluation Criteria and Associated Submission Elements

Evaluation Criteria	Submission Elements	Grouper and/or Episode Level Evaluation
Scientific Acceptability The extent to which the grouper produces consistent (reliable), and clinically relevant (valid) episodes.		
Reliability		
The grouper specifications are clear, facilitate understanding and enable consistent implementation by the user.	• General description of the grouper design and construction • Description of intent of grouper (i.e., use cases such as provider profiling) • Description of the level of analysis • Description of the target population(s) category • Description of the data source used to develop the grouper • Description of the data source or collection instrument • Data dictionary and/or code tables • Description of data requirements for users (i.e., data fall out thresholds) • Description of steps to prepare the data associated with missing data and the rationale for this methodology (e.g., any statistical techniques to impute missing data) • List of clinical or procedure episodes measured by the grouper (e.g., AMI, pneumonia)	Grouper and Episode Levels

Evaluation Criteria	Submission Elements	Grouper and/or Episode Level Evaluation
Reliability testing demonstrates that the episode groupings are repeatable, with 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).	• Description of the testing method/approach • Description of the data sample used • Description/discussion of results	Grouper and Episode Levels
Validity		
The intended use of the episode grouper aligns with the logic for grouping claims.	• Description (including codes) and rationale for clinical inclusions and exclusions • Description of general rules for assigning claims to each episode and hierarchies • Rationale and decisions related to concurrent services	Grouper and Episode Levels

Evaluation Criteria	Submission Elements	Grouper and/or Episode Level Evaluation
The intended use of the episode grouper aligns with the logic for grouping claims.	• Description and rationale for trigger and end mechanisms for each clinical episode (e.g., codes, clean periods)	Grouper and Episode Levels
	• Description of the steps used to cluster, group, or assign claims beyond those associated with the episode's clinical logic	Grouper Level
	• Description and rationale of the methods used for identifying concurrent clinical events, episode redundancy and overlap and disease interactions (e.g., disease hierarchies, severity hierarchies)	Grouper Level
	• Description of how complementary services have been linked to the episode and rationale for this methodology	Grouper and Episode Levels
	• Identification, description and rationale for resource use service categories including the method or algorithms to identify resource units, including codes, logic and definitions	Episode Level
	• Costing methodology (if costing is applied within the grouper decision logic)	Grouper and Episode Levels
Validity testing demonstrates that the episodes are clinically relevant and appropriate.	• Description of the testing method/approach • Description of the data sample used • Description/discussion of results	Grouper and Episode Level

Evaluation Criteria	Submission Elements	Grouper and/or Episode Level Evaluation
Severity and risk adjustment strategy is clearly specified and is based on patient factors that influence the clinical course and assignment of claims.	• 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, provide a description of the method for assigning risk, including any hierarchies and logic for assigning these claims	Grouper and Episode Levels
Threats to validity (i.e. limitations) are adequately described including how threats have been addressed.	• Description of threats to validity and limitations of the grouper • Discussion of how those threats and limitations were addressed	Grouper and Episode Levels
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.	• Indication/description of whether data elements are generated as part of care processes	Grouper Level
Required data elements are available in electronic sources.	• Description of the availability of specified data elements that are needed to compute the episode in defined, computer-readable fields.	Grouper Level

Evaluation Criteria	Submission Elements	Grouper and/or Episode Level Evaluation
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).	• Description of what was learned/modified as a result of testing and/or operational use of the episode regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues. • If applicable, a description of the fees/costs associated with the purchase of software, licensing or other costs required to implement the grouper.	Grouper and Episode Levels
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.	• Description of current use • If not currently in use, description of how the grouper could be used to further the goal of high-quality, efficient healthcare for individuals or populations.	Grouper and Episode Levels
The planned use of the episode grouper is clearly described.	• Planned use of the grouper (e.g., specific programs for public reporting or payment)	Grouper and Episode Levels

Evaluation Criteria	Submission Elements	Grouper and/or Episode Level Evaluation
The benefits of the use of the episode grouper outweigh any unintended consequences.	• Description of unintended negative consequences to individuals or populations identified during testing and describe how benefits outweigh them or actions taken to mitigate them • 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	Grouper and Episode Levels
The maintenance plan demonstrates adequate maintenance of the grouper to enable ongoing meaningful use of the output by users and implementers.	• Description of the vendor maintenance process (frequency, scope, process, implementation)	Grouper Level

National Quality Forum
1030 15th St NW, Suite 800
Washington, DC 20005
http://www.qualityforum.org

ISBN 978-1-933875-73-6