Quality Indicator Measure Development, Implementation, Maintenance, and Retirement

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
http://www.qualityindicators.ahrq.gov

Contract No. 290-04-0020 (AHRQ SQI-II)

Prepared by:
Battelle
505 King Avenue
Columbus, OH 43201

HEALTH POLICY STANFORD
117 Encina Commons
Stanford, CA 94305

May 2011

Disclaimer

This document is based on research conducted by the Battelle under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-04-0020). The findings and conclusions in this document are those of the author(s), who are responsible for its content, and do not necessarily represent the views of AHRQ. No statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help clinicians, employers, policymakers, and others make informed decisions about the provision of health care services. This document is intended as a reference and not as a substitute for clinical judgment.

This document may be used, in whole or in part, as the basis for the development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

Suggested Citation:

Center for Health Policy/Center for Primary Care and Outcomes Research & Battelle Memorial Institute. Quality Indicator Measure Development, Implementation, Maintenance, and Retirement (Prepared by Battelle, under Contract No. 290-04-0020). Rockville, MD: Agency for Healthcare Research and Quality. May 2011.

None of the investigators has any affiliations or financial involvement that conflicts with the material presented in this document.

Contents

List of Tables

List of Figures

Executive Summary

Overview

This document describes the Agency for Healthcare Research and Quality (AHRQ) Quality Indicator (QI) measure development, implementation, maintenance, and retirement processes. It describes the overall approach to indicator development and then outlines the steps taken to develop and maintain indicators. Timelines are provided, which are based on indicator development within a field with established measurement concepts. These timelines may change as indicator development moves to alternative areas. An executive summary is presented, followed by a detailed document.

Quality Indicator Development Model

Quality indicators consist of:

- a concept , the specific aspect of quality captured by the measure (e.g., healthcare associated infections)

- a perspective, the point of view from which the measure is taken (e.g., patient, clinical, system)

- a method, how is the actual concept measured (e.g., data source, measure type, observable event, specification and risk adjustment)

- an application, how is the measure actually used (e.g., pay for performance, quality improvement, comparative reporting)

Each of these aspects come to together to inform the implementation of a measure, which include the data collection guidelines, software tools and implementation guidance. It is the full implementation that must be considered when assessing the validity and usefulness of the indicator. Development, validation, and use occur in a continuous cycle, with use informing further development and validation activities.

Phase I: QI Measure Development

Task 1: Identification of Candidate Indicators

Literature Review

Time line: Approximately 2.5 months

The first step in QI measure development is to identify candidate indicators. This includes a semi-systematic review of the peer reviewed literature, grey literature and related databases [(e.g., National Quality Forum Endorsed® Standards[1] and the National Quality Measures Clearinghouse[2]]. Search strings are used to identify articles that address potential indicators, data sources for potential indicators, existing validation efforts and risk adjustment. Abstracted information is used to identify a list of candidate indicators.

Development of Conceptual Model

During the first phase, it is useful to develop a conceptual model of the area of interest. The conceptual model includes: the clinical pathways for the area of interest, the multiple perspectives (when applicable), or observable events.

Expert Engagement

Time line: Throughout task

Experts are important to the QI measure development process, as they enhance the scientific acceptability of the QIs. Expert engagement helps facilitate the development of a conceptual model to inform the entire QI measure development process. As the QI measure development process proceeds (e.g., after the literature review or consultation with current experts), additional experts may be identified to enhance the understanding of the team in specific areas related to the topic of interest.

Task 2: Assessment of Candidate Indicators

The second step focuses on the evaluation of the candidate indicators. The evaluation follows the National Quality Forum Measure Evaluation Criteria[3]: Importance, Scientific Acceptability, Usability, and Feasibility.

Initial Specifications of Candidate Indicators and Existing QIs

Time line: Approximately 1 month

Initial specifications are based first on the specification identified in the literature if available and include numerator, denominator and exclusion criteria. Modifications to specifications are made to adapt the indicator to the available data, improve the indicator based on new evidence, harmonize with other indicators, or incorporate updates to the data source such as coding changes.

Literature Review: Evidence Base for Candidate Indicators

A second literature review focuses on the abstraction of evidence supporting the indicators. This search includes not only the evidence from articles identifying candidate indicators, but also articles that discuss the outcome or event of interest without proposing the event as a quality indicator.

Panel Review

Time line: Approximately 3.5 months

The panel review provides clinical face validity (i.e., the QI measure assess what it "looks like" it will) for the indicators. The structured review uses a Modified Delphi or Nominal Group process, based on the RAND/UCLA Appropriateness method.[9,10] The process uses techniques meant to maximize information exchange while minimizing cognitive biases. Initially, panelists independently rate the indicators, followed by a conference call to exchange opinions. Panelists then again independently rate the indicators. This final rating is used to tier the indicators as to their relative utility. The panel traditionally has assessed validity from a clinician perspective, buy the technique could be used with other stakeholder groups.

Risk Adjustment

Time line: Approximately 3 months

The process of risk adjustment allows the candidate indicators to account for certain relevant factors (e.g., comorbidities) that may otherwise dilute the utility of the information obtained from the candidate indicators. Risk adjustment models are created from information gathered during literature review, team and panel review, and initial indicator testing.

Empirical Analyses

Time line: Approximately 2 months, with existing data

The empirical analyses serve to determine relative bias of the candidate indicators, the precision and reliability of each indicator, rates and variation in rates between providers or areas, and the relatedness of the candidate indicators within providers or areas. The analytic plan is created by information obtained in the literature review, team and panel review, and initial indicator testing. Analyses generally are performed on available data, but chart review may augment that process.

Finalization of Specifications

Time line: Approximately 1 month

Based on all development activities, the specifications are finalized to maximize validity.

Summary of Evidence for each Recommended Candidate Indicator

Time line: Approximately 2 months

Using the information from the finalized specifications, a summary of evidence gathered over the QI develop process is created for each recommended candidate indicator to facilitate review and decisions regarding the indicators.

AHRQ Review and Decision on Indicators

Time line: Approximately 1 month

AHRQ uses the finalized specifications and summary of evidence on the candidate indicators to determine if some or all of the recommended indicators warrant an additional development phase for inclusion in a publicly released module.

Phase II: QI Implementation

If AHRQ endorses the advancement of recommended indicators to Phase II of QI measure development, the focus is on implementation.

Coding Quality Indicators into Software

Time line: Approximately 1 month

The software team codes the indicators into user-friendly SAS and Windows based software for release to users as a QI module.

Testing

Time line: Approximately 2 months

All QI modules are internally and externally tested, including implementation with existing data, to ensure accuracy and consistency. Testing includes identifying and deploying an appropriate test dataset for use with the AHRQ QIs.

User Documentation

Time line: Approximately 1.5 months

User documentation is developed that includes specifications (i.e., the indicator statement, numerator, denominators, exclusions and coefficient tables for risk adjusted measures) for each QI, user guides (i.e., the evidence summaries for each measure), SAS and WinQI software instructions and logs of changes from the prior QI version to the current version.

Phase III: QI Maintenance – Preserving Scientific Acceptability

In order for the QIs to remain scientifically acceptable and useful, they must be maintained and potentially enhanced on a regular cycle. QIs need to be updated based on such factors as: recent evidence published in the literature (particularly as publications are made available using the specific QI) and from user feedback, technical specification updates including International

Classification of Diseases-Ninth Revision-Clinical Modification (ICD-9-CM) coding updates, periodic clinical panel review, the NQF endorsement and maintenance process, and newly available data and methodological advances in the industry. Each of the material maintenance steps must be considered within the broader measure life cycle.

Evidence

Time line: Throughout task

Evidence may arise through continual formal or informal literature review, ongoing validation studies, or user submitted experiences. New evidence suggesting improvements to specifications, implementation or documentation is evaluated.

Technical Specification Updates

Time line: Throughout task

The QI codes and risk adjustment covariates are updated annually to reflect fiscal year ICD-9-CM and Diagnosis-related Group (DRG) changes and currently available comparative data used for the reference population. Additionally, new U.S. Census data on the population of counties is updated, which is relevant to area-level measures.

Panel Review

Time line: Throughout task

When needed a clinical review panel is engaged if the evidence reviewed, user feedback, or coding changes warrant a detailed examination of the indicators.

National Quality Forum Submission and Maintenance

Time line: Throughout task

NQF submission and endorsement is considered for all QIs developed. QIs that meet the NQF evaluation criteria[3] are considered for submission. QIs accepted for endorsement enter a regular maintenance and annual review cycle established by NQF.

Newly Available Data and Methodological Advances

Time line: Throughout task

Measurement creates demand for better data and methods, and in turn these data and methods are incorporated into the measures. Processes employed in new method development work may include work group input, empirical analyses and other efforts.

Phase IV: QI Retirement

Occasionally AHRQ has retired indicators by removing them from the software and documentation.

Evidence

Time line: Throughout task

A variety of inputs can inform retaining or retiring measures. Review of literature relevant to the QIs and feedback from users may suggest that an indicator is no longer scientifically acceptable and should be removed from the QI module.

Remove Coding of Quality Indicators from Software

Time line: Approximately 0.5 month

The software team removes the measure codes that define the retired indicators from the software for release to users.

Testing

Time line: Approximately 0.5 month

Internal and external testing of the resulting module includes ensuring that the removal of the indicators did not introduce any unexpected consequences. Specifically, removal from the composites requires re-evaluating the composites and the weights using three criteria: discrimination, forecasting and construct validity.

User Documentation

Time line: Approximately 0.5 month

User documentation is updated to remove the retired indicator from specifications for each QI, user guides, SAS and WinQI software instruction and logs of changes from the prior QI version to the current version.

Summary

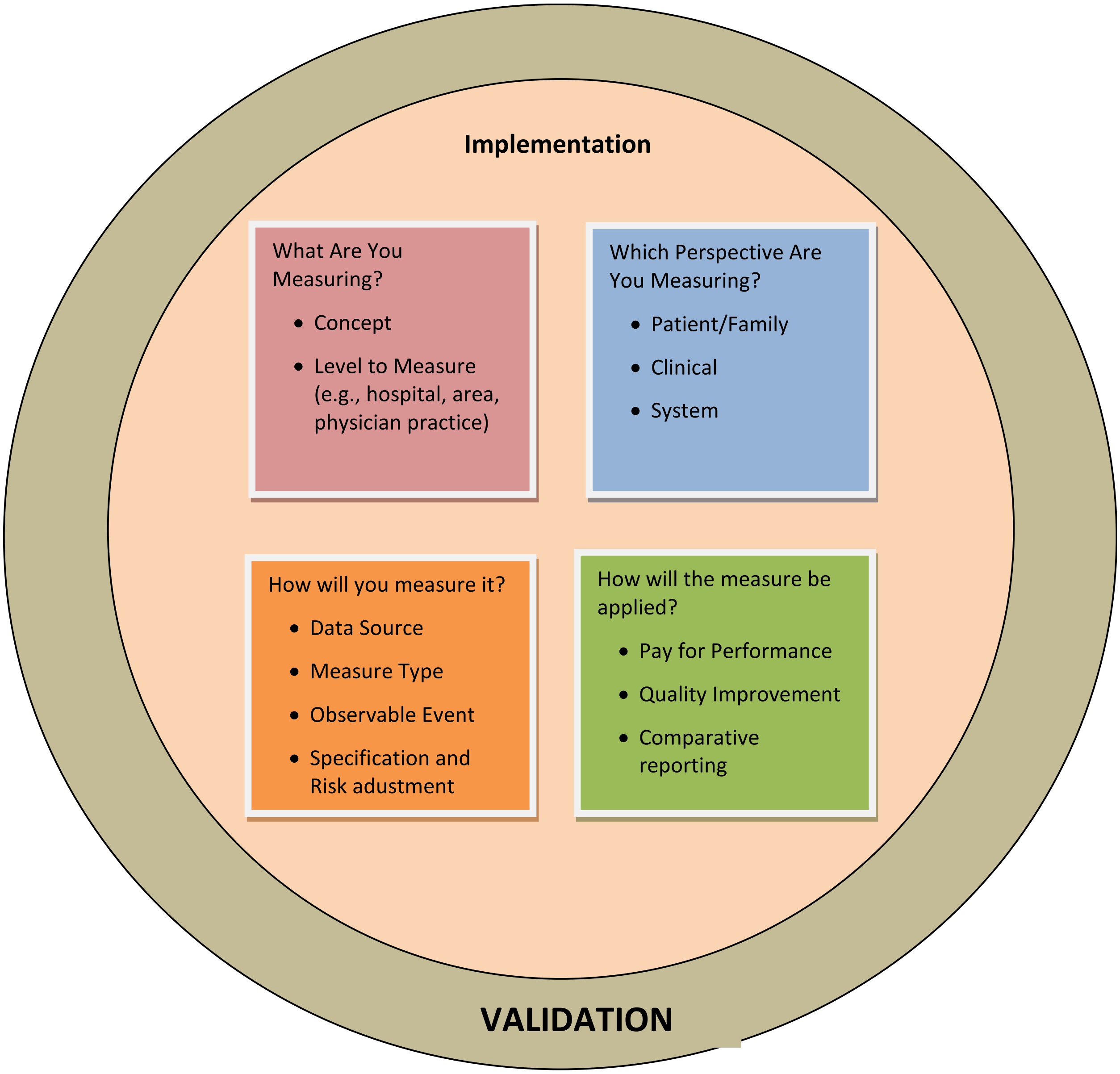
The QI measure development process involves four phases. The first phase is candidate indicator development for an identified topic area of interest. The steps involved in the first phase are: (1) identification of candidate indicators, which includes literature review, expert engagement, and selection of candidate indicators and (2) assessment of candidate indicators, which includes specifications of candidate indicators and existing AHRQ QIs, panel review, risk adjustment, empirical analyses, finalization of specifications, and summary of evidence for each recommended candidate indicator. The second phase is implementation of the QIs into the AHRQ QI software, which involves coding the QIs into the software, testing, and developing user documentation. The third phase is maintenance of the QIs, which involves review of the evidence, technical specification updates, periodic clinical panel review, NQF endorsement submission and maintenance, and newly available data and methodological advances. The final phase is retirement which involves evidence, removing coding from software, testing and user documentation. These phases and processes may require modifications to meet the needs of indicator development in new areas.

Overview

This document summarizes the Agency for Healthcare Research and Quality (AHRQ) Quality Indicator (QI) measure development, implementation, maintenance, and retirement processes. It is intended to convey the general steps and rationale involved across a wide range of healthcare indicators. First, we present a general outline of the aspects of an indicator that affect the development process and scope of work. Figure 1 graphically depicts how these aspects of quality indicators are taken together in implementation efforts, and ultimately validation efforts.

Second, we describe in detail the three key phases of indicator work: Phase I: QI Measure Development, Phase II: QI Implementation and Phase III: QI Maintenance. Within each phase we provide examples from previous development efforts, and an approximate time line for completion of each step, (assumes scopes similar to recent development efforts such as that for "Healthcare Acquired Infections"). The resources and time line for developing QI measures depend on factors such as: number of measures, stakeholder involvement, data requirements, and current state of measurement in that field. Finally, we discuss an additional phase of discontinuing a measure in Phase IV: QI retirement section.

Figure 1. The Quality Indicator Development Model



Section 1: The Quality Indicator Development Model

Aspects of Quality Indicators

Each potential indicator has four main considerations that can help define the scope of measure development.

Concept. Each measure is intended to capture a specific aspect of quality. This may be as broad as healthcare associated infection or patient safety, but the concept may also be more granular such as surgical site infection or transitions between healthcare settings. In addition, the concept is applied at a specific level of the healthcare system, such as hospital, area, physician group, or payer. Most of the existing QIs are measured at the hospital level, although the Prevention Quality Indicators (PQI) are measured at the area level.

This consideration relates closely to the NQF criterion[3] of importance. The criterion states "Extent to which the specific measure focus is important to making significant gains in health care quality (safety, timeliness, effectiveness, efficiency, equity, patient-centeredness) and improving health outcomes." The overall concept is determined *a priori*, while more granular concepts may be refined during Phase I of indicator development (see Section 2, Phase I). The importance of these concepts are demonstrated through literature review, empirical testing of prevalence and variation, and panel review.

Perspective. The second consideration is the perspective or perspectives captured by the measure. Healthcare quality can be viewed from multiple, interdependent perspectives. The patient perspective requires asking about the patient experience with their care and its effects on them (e.g., physiologic or psychological well-being). The health professional perspective requires thinking about what the clinical processes are that are expected to produce the desired patient outcomes (e.g., complications, mortality). A third perspective takes a more macro view, and requires asking what a healthcare system manager or policy maker might see as critical for sustainability of care for populations of patients (e.g., healthcare efficiency, access to care). Although for the most part the current AHRQ QIs have focused on the patient and health professional perspective, the PQIs may be more reflective of the system perspective. In addition, as quality measurement moves to new areas, additional perspectives may be considered. For instance, care coordination may incorporate the family/caregiver perspective as an extension of the patient perspective.

Method. To capture the concept and reflect one or more perspectives, the measure will incorporate a specific measurement method. In short, how exactly would one measure the concept? This includes several aspects: data source, measure type, observable events, specification and risk adjustment. First, the data source may vary, and must be defined. For the QIs in general, we have constrained the measures to administrative data, but alternative data sources may include survey or other data collection efforts. The specific measurement approach may include a variety of measure types. The NQF defines the following measure types: outcome, intermediate outcome, process, structure, patient experience, access, and efficiency. Most QIs are outcome measures, although a few process measures (e.g., procedure utilization) and structure measures (e.g., volume) are included. Finally, in order to measure the concept of interest, one must identify an observable event. For instance, patient safety is measured through observable complications, or access to quality outpatient care is measured through potentially avoidable healthcare encounters. The data source and measure type are often determined *a priori*, but also may be the result of Phase I development efforts (See Section 2). The identification of observable events is usually part of Phase I development efforts.

Once the data source, measure type and observable event are defined, the indicator can be fully specified. The NQF Measure Evaluation Criteria state that a measure must be ”well defined and precisely specified so that it can be implemented consistently within and across organizations and allow for comparability.”[3] The measure specification is created and refined during the development process. Finally, risk adjustment must be defined and specified, if necessary. The need for risk adjustment is determined in Phase II of indicator development (See Section 3).

Application. The final measure development consideration is the anticipated application of the indicator. A measure may be designed for use as a quality improvement tool or honed for application that allows for the comparison of entities (e.g., comparative reporting or pay for performance). Although measures may be useful in more than one application, some development may require refining indicator definitions for a specific application. For instance, NQF-Endorsed® Standards[1] must be applicable to comparative reporting uses. Thus, issues of bias must be considered when refining specifications and risk adjustment. The appropriate application of an indicator is partially informed by validation efforts. In general, the QIs are designed for multiple applications and validity testing has resulted in additional guidance regarding the most appropriate applications.

Implementation Efforts

Together the above four considerations fully define an indicator. However, an indicator definition alone does not make an indicator “useful.” Beyond the definition, the implementation of the measure impacts its validity and usefulness. The development of implementation tools is important to ensure consistent application of the indicator definition. Although many aspects of implementation could be discussed, this document will describe three areas that are salient to the development process: (1) data collection guidelines, (2) software tool development, and (3) implementation guidance.

Data collection guidelines. The collection of data that feed into an indicator must be consistent across users. In general, since the QIs have relied on administrative data, these guidelines are all ready established in the form of including International Classification of Diseases-Ninth Revision-Clinical Modification (ICD-9-CM) coding guidelines, other administrative data guidelines, and Healthcare Cost and Utilization Project (HCUP) database variable definitions. However, when expanding beyond administrative data or when considering *de novo* data collection efforts, development activities would also include the creation of data collection guidelines.

Software tool development. The development of software tools allows for the consistent application of the indicators as well as improved usefulness and availability of the indicators.

The AHRQ QI Software is an essential component of indicator development, and is discussed in Section 3.

Implementation guidance. In some cases, issues remain despite the best efforts to establish fully refined definitions and data collection guidelines. For instance, during development and validation residual bias may be identified. In this case, implementation guidelines can highlight this and caveat that comparisons may be biased. In addition, when concerns arise regarding the usefulness of an indicator for specific applications, this can be noted in implementation guidance. The implementation guidance is offered in the document, "AHRQ Summary Statement on Comparative Hospital Public Reporting"[4] and the AHRQ QI User Guides for each module[5-8]

Validation Efforts

Since the validity of an indicator in actual use depends on the indicator definition and specification, all validation efforts in essence focus on the full package of the definition (including the four components) and implementation efforts. For instance, any assessment of the criterion validity of an administrative data based indicator incorporates both the assessment of the specification as well as adherence to ICD-9-CM coding guidelines. As quality indicator development expands beyond administrative data, it is essential that validation efforts include assessments of implementation efforts. Validation activities begin in Phase II during indicator development (see Section 2) and continue in Phase III during indicator maintenance (See section 3).

The Development/Validation/Use cycle

Indicator development is dynamic rather than stepwise. For instance, during the course of validation, improvements to indicator specifications may be noted. Validation may inform the most appropriate application, and continued use of the indicators will ultimately result in a richer knowledge base about the indicators and continued improvement or retirement. This dynamic nature is a key attribute of the QI development process.

Section 2: Phase I: QI Measure Development

Task 1: Identification of Candidate Indicators

Overview of Task 1

The purpose of Task 1 is to understand the current state of measurement in the field of interest. By doing this, the scope of development as it relates to the four primary measure components (concept, perspective, method, and application) can be refined to meet the needs of the field. Typically, during this phase, potential indicators are identified and the steps described below assume the task is to identify, refine and evaluate existing indicators. However, when no specified indicators exist, this phase could identify indicator concepts that might lead to *de novo* indicator development. In addition, when the data source is not determined *a priori*, this phase is useful in identifying potential data sources.

To focus this phase, a list of research questions is developed. These may include:

- What indicators exist in the area of interest, what concepts do they cover, what is their measurement approach and perspective (if applicable)?
- Which data are used in existing indicators or if no indicators are yet specified, what are potential data sources?
- To what extent are the existing indicators specified and used?
- To what extent have the existing indicators been validated according to the criteria included in the NQF measure evaluation framework[3]?
- Are the existing indicators risk adjusted and if so, by what method?

For examples of candidate indicator products stemming from Phase I activities see:
http://www.ahrq.gov/qual/careatlas/index.html
http://www.qualityindicators.ahrq.gov/modules/psi_resources.aspx

Literature Review: Identification of Candidate Indicators

Time line: Approximately 4 months, assuming existing indicators, and a priori determination of data source

The first step in QI measure development is to conduct a literature review of both peer reviewed and, when applicable, grey literature on the topic area to identify candidate indicators. In addition to the literature review, two additional strategies may provide candidate indicators. First, the current AHRQ QIs are also reviewed, given that their areas of focus may be easily adapted to the current topic area (e.g., slight modification to the denominator population to focus on a particular area of health care). Second, the NQF®-Endorsed Standards[1] or the National Quality Measures Clearinghouse[2] are scanned for applicable measures.

Although the format of the literature review may vary in resource intensity, generally a semi-structured approach is used. Medline is the primary database searched, and may be augmented by additional databases of peer reviewed literature (e.g., PsychInfo) or metasearch engines (e.g., Google Scholar) when necessary. We first develop search strategies by identifying MeSH terms, key words and limits based on the background information and scope of work. Key articles known to the research team are identified and also used to develop and validate search strategies (e.g., ensuring that the search string captures key articles). When available, consultation with a librarian is useful to refine search strategies. Publication time frame parameters may be specified in the event that a large number of resources are identified in initial searches. The reference lists of identified resources are also reviewed to identify additional articles.

Articles for full extraction are identified using title screens, abstract screen and finally article screening according to their relevance to the key research questions. Abstract forms or databases allow for the systematic gathering of information about candidate indicators. See Appendix 1 for an example list of database fields. Information abstracted generally includes:

- Article information (citation information)
- Indicator characteristics (measure type, data source, level of measurement, concept)
- Measure specification (observable event or outcome of interest, denominator or population at risk)
- Risk adjustment (methods, stratification)
- Validation performed and accompanying data (reliability, psychometrics, calibration, discrimination)

In preparation for abstraction, a training stage ensures consistency in abstraction. Two or more members of the technical team first abstract one or more articles and differences are reviewed and resolved.

Information learned from the literature review is then summarized, including a list of identified indicators, potential challenges in measurement, and evidence gaps.

Based on this review, the list of potential indicators is narrowed to the final candidate list. In general, candidate measures must be:

- Defined using the available data source (thus far generally administrative data) OR
- Adaptable to the available data source (e.g., measures defined using laboratory or clinical data, but could be adapted to use administrative data)

- Without demonstrated poor performance in initial literature scan (e.g., poor sensitivity)

Development of Conceptual Model

During the first phase, it is useful to develop a conceptual model of the area of interest. The conceptual model includes: the clinical pathways for the area of interest, the multiple perspectives (when applicable), or observable events. The conceptual model is useful to highlight gaps in currently developed measures, refine the scope of projects and provide direction of future research. The conceptual model is not intended to be a detailed representation of the entire area of interest, but rather a general guide to indicator development and validation. As such, factors outside the designated scope of the development project may be omitted from the model. It is tailored to the task, and spans only the scope specified for that task.

See Chapter 3 in the Care Coordination Measure Atlas (http://www.ahrq.gov/qual/careatlas/index.html) for an example of a conceptual model.

Expert Engagement

Time line: Throughout task

Experts are important to the QI measure development process, as they enhance the scientific acceptability of the QIs. Expert engagement helps facilitate the development of the conceptual model to inform the entire QI measure development process. As the QI measure development process proceeds (e.g., after the literature review or consultation with current experts), additional experts may be identified to enhance the understanding of the team in specific areas related to the topic of interest.

Experts already engaged. Experts already engaged with the QI measure development process include the primary technical team and subcontractors. These experts may have expertise related to measure development and/or expertise related to the specific topic area.

Additional experts to engage. Individuals with expertise in a specific area, or a group of experts in a specific topic area may be engaged beyond the current team of experts. Such additional experts may or may not have published in the area. Knowledge sharing groups are designed to link together researchers from other government agencies and outside organization in the topic of interest. Representatives convene for one or more webinars to hear about the current development efforts and discuss specific topics.

Example: Knowledge sharing groups consisting of representatives from Centers for Medicare and Medicaid Services (CMS), Centers for Disease Control and Prevention (CDC) and the QI Development team met in a series of conference calls to discuss Healthcare Acquired Infection measurement. Topics included planned methods for QI measure development methodology, existing efforts and indicator specifications.

Task 2: Assessment of Candidate Indicators

Task two focuses on the evaluation of candidate indicators. The evaluation follows the NQF evaluation framework: Importance, Scientific Acceptability, Usability, and Feasibility.[3] Details regarding the NQF evaluation framework can be found at http://www.qualityforum.org/docs/measure_evaluation_criteria.aspx. Evidence arises from literature review, empirical analysis and expert panel evaluation. The findings from this task inform the final specification of the indicator, the final selection of indicators, and guidance regarding indicator use.

Again, following the NQF framework,[3] the research questions for this task focus on the specification of the indicator and the validity of that specification. By this stage, concept and perspective have been determined. The evidence review is focused on the level and specification being assessed, rather than expanding the review to alternative levels or data sources (unless directly applicable). For instance, when assessing indicators of hospital mortality, construct validity can be demonstrated by interventions that reduce hospital level mortality. There is often a large literature of randomized controlled trial (RCT) and observational studies examining interventions impacting patient-level mortality. This represents an additional level, and thus is not directly applicable to this task.

The research questions for this phase may include:

- What is the prevalence of the health event or condition in the population of interest?
- What are important clinical considerations for this health event/condition in the population of interest (if population is specialized)?
- Is there evidence of poor quality care is related to the health event/condition in the population of interest/level of measurement?
- What is the frequency of the event/condition in the population of interest?
- What is the evidence for prevention of the health event or hospitalization at the level of measurement?
- What factors impact hospitalization for the health event/condition in the population of interest?
- What is known about documentation of and coding related to the health event/condition, including sensitivity and specificity?
- What is known about the timing of the event in relation to the observed time period (e.g., timing of complications in relation to hospitalized days)?

Initial Specifications of Candidate Indicators and Existing QIs

Time line: Approximately 1 month, assuming existing measures primarily specified using available data source.

The candidate indicator list must then be fully specified using the available data source (e.g., in most cases administrative data). The initial specification is used for initial empirical analyses (e.g., application to HCUP databases) and panel evaluations. The starting point is always the specification identified in the literature. Specifications include:

1) Numerators
2) Denominators
3) Exclusion criteria

Modifications to those specifications are made to:

1. Adapt any specification to the available data source. In cases when the existing indicator is entirely or partially defined using different data sources (e.g., laboratory or clinical data), the indicator is adapted to the available data source (e.g., administrative data). This is done by identifying the intent of the specification, and creating an alternative definition using available data. The alternative definition is tested empirically and modified as necessary.

   Example: Patient Safety Indicator (PSI) 04. Death among surgical patients with serious treatable complications. This indicator was originally specified using clinical data. The research team, in conjunction with clinical panel review, adapted the clinical data based complications to be specified using ICD-9-CM codes.

2. Improve the specification based on evidence available in the literature review, initial empirical analyses, and expertise. In cases where the literature based evidence or expertise identifies shortcomings in the specification (such as poor sensitivity), specification may be modified. In addition, if initial empirical analyses demonstrate shortcoming, the indicator specification would be improved. Improvements are tested empirically and modified as necessary.

   Example: A published study shows a non-specific code included in an indicator's numerator definition has a high false positive rate and omitting that code would miss few cases. The code is removed from the initial specification.

3. Align and harmonize similar measures or within the current QI measure framework.

   Example: Definitions of pneumonia differed in measures of pneumonia mortality, including which specific ICD-9-CM codes were included. The definitions were compared and harmonized. In some cases, panel review or collaboration between organizations may aid in this process.

4. Update the specifications based on changes to the data source (e.g., coding updates).

   Example: The base specifications for many PSI indicators were based on the work by Lisa Iezzoni several years prior to the PSI development efforts. Since that time, many changes impacting the indicators were made to the ICD-9-CM coding system. The project team examined these changes and modified the definitions as necessary.

Literature Review: Evidence Base for Candidate Indicators

A second literature review focuses on the abstraction of evidence supporting the indicators. This search includes not only the evidence from articles identifying candidate indicators, but also articles that discuss the outcome or event of interest without proposing the event as a quality indicator. For instance, some articles may assess the impact of policy changes by examining reductions in area level hospitalization rates. These studies inform the validity of the indicator, although the articles are not specifically proposing a "quality indicator."

The process mirrors the initial literature review. In fact, often the articles identified in Task 1 Literature Review can be fully abstracted for Task 2 simultaneously. However, the search strings may need to be expanded depending on the scope of the initial literature search. For instance, if the initial search string specified a MeSH term of "quality indicators," this term may need to be removed and the search string modified to improve the specificity of the search.

The abstraction form includes fields to abstract data relating to importance, including prevalence and variation, reliability, criterion validity, construct validity, bias and risk adjustment, usability, including adverse effects, and feasibility.

Following abstraction, the evidence is summarized and assessed. Evidence that suggests diminished validity is noted, and some indicators may be dropped from further development at this point. Evidence gaps are also identified and when possible these are addressed during further development activities (e.g., empirical analyses, panel evaluation).

Panel Review

Time line: Approximately 3.5 months

The panel review provides clinical face validity (i.e., the QI measure assesses what it "looks like" it will) for the indicators. The panel process primarily addresses the scientific acceptability of the indicator, specifically face validity, although the panel review also addresses other concepts. The NQF criteria specifically require "If face validity is the only validity addressed, it is systematically assessed."[3] In this case, we use a modified RAND/UCLA Appropriateness Method[9] to establish the consensual validity of the indicators, as stated by Green and Lewis, extending "face validity from one expert to a panel of experts who examine and rate the appropriateness of each item…."[10] The panel assesses this validity from the clinician perspective, although a similar process could be used to assess other perspectives.

Beyond the primary purpose of establishing consensual validity, the panel has two secondary purposes. First, since the process engages a variety of professional organizations, it increases the transparency of the development process. Second, by engaging clinicians from a variety of practice settings and specialties, we are able to refine indicator definitions to best reflect the intended purpose of the indicators. The method itself does not demand consensus, but rather encourages sharing of information. The only exception is for definitional modifications, when consensus is sought.

The RAND/UCLA Appropriateness Method[9] has been termed a "modified Delphi" process or a Nominal Group Technique. Traditionally the Delphi Technique uses multiple rounds of independent ratings (e.g., by mailed questionnaire), with written summaries of responses distributed between rating rounds. The RAND/UCLA Appropriateness Method[9] also uses an initial independent rating, followed by the distribution of summarized results. At this point the panel then meets, traditionally in person and in some cases via conference call, to discuss opinions regarding the indicators. Panelists then re-rate the indicators independently.

The composition of panels is decided *a priori* (i.e., number and type of specialties) based on the indicators being assessed. For instance, panels that include for review indicators of cardiac procedures will include cardiologists and/or cardiovascular surgeons. The panels are limited in size, since it is difficult to moderate larger groups; the recommended panel size for the RAND/UCLA Appropriateness Method[9] is eight to 12 individuals, although we have accommodated up to 15 individuals. Typically, the panels are intended to obtain a clinician's viewpoint so all panelists must practice 30% FTE in direct patient care. In addition, in order to create diverse panels that are not overly influenced by any single panelist, the panels are populated with multiple individuals from a variety of specialties (e.g., essential specialties are represented by multiple panelists), practice settings (e.g., urban vs. rural, teaching vs. community), and regions.

Once the desired panel composition has been determined, the first step in the process is to seek nominations for panelists from national professional organizations in order to obtain a diverse panel of qualified clinicians, as well as to engage these organizations. Organizations representing the desired specialties are identified and provided with a summary of the project and asked to provide nominations. Typically, we ask for three nominations per slot. Panelists are not considered official representatives of the organizations.

Following the nomination process, the panelists are contacted, provided with a project summary, and asked to respond with their interest. The panelists are also asked to provide background information (e.g., education) and practice information (e.g., specialty, practice setting, population served, and academic affiliation) to aid in creating diverse panels. Project team members assigned nominees to panels. If more nominees are available than required for a given specialty, nominees are assigned based on maximizing the diversity in the panel. When two nominees provide similar diversity, a panelist is randomly chosen.

The final planning stage involves the creation of the panel evaluation materials. The panel methods strive to combat common cognitive errors encountered during group processes. This includes initial independent assessment via questionnaire to avoid "group think," followed by the exchange of information to avoid “silos” and then the final independent assessment. The questionnaire itself is both quantitative and qualitative in nature, allowing for the quantitative assessments while obtaining information useful in improving the indicators and interpreting the ratings. The questionnaire is anchored around an overall rating, which is the question for which final analyses are centered. The other questions are intended to prime the panelists, to ensure the panelists consider similar issues when assigning the overall rating. See Appendix 2 for an example questionnaire.

To ensure that all panelists have similar access to the evidence surrounding the indicators, an initial packet is distributed. This packet includes summaries of literature based evidence and initial empirical analyses (e.g., overall rates of draft indicators, analyses impacting indicator specific questions). In addition, project summaries and information on the available data source can be helpful to ensure that each panelist understands the purpose and limitations of the indicators. See Appendix 3 for an example indicator evaluation information sheet.

Following the initial ratings, the results are summarized and redistributed to panelists. The panelists are given the opportunity to exchange opinions during a conference call. The call is moderated by a neutral moderator, who seeks to encourage information exchange rather than consensus. The agenda is set based on two factors: first, items of strong disagreement during the initial rating are highlighted and discussed; second, any areas of ambiguity in the initial ratings are addressed. Finally, indicator specific questions developed by the team a priori are discussed. For instance, when two alternative definitions are being considered, we may ask the panel to weigh in on the issue.

Following the call, the results from the call analyzed and any analyses that could address any panelist questions are performed. Where definitional changes were identified during the call, these changes are made and rates re-run. All this information is provided back to panelists to rate the indicators.

The final ratings are analyzed using the RAND/UCLA methodology.[9] The level of support for the indicators is assigned based on median score of the overall rating of the indicator as well as a measure of agreement. Agreement or disagreement is assigned based on the dispersion of final ratings. The indicators are tiered based on this final analysis.

Example of tiering from PSI development:

Acceptable	Median falls between 7 and 9 (inclusive of both), agreement
Acceptable (-):	Median falls between 7 and 9 (inclusive of both), indeterminate agreement
Unclear:	Median falls between 7 and 9 (inclusive of both), disagreement, OR
	Median falls between 5 and 7 (inclusive of neither), agreement or indeterminate agreement
Unclear (-):	Median between 4 and 5 (inclusive of both), agreement, indeterminate agreement or disagreement, OR
	Median falls between 1 and 3.9 with disagreement.
Unacceptable:	Median falls between 1 and 3.9, agreement or indeterminate agreement.

Alternative Panel Processes

In some previous indicator development the process described above has been altered. This is generally done when the needs of the project dictate a larger number of panelists (because a large number of specialties must be represented), or the panel provides input at a different stage in the development process (e.g., identifying important concepts). The hybrid panel approach that allows for larger panel size can be found in the report "Expanding Use of the AHRQ Prevention Quality Indicators: Report on the Clinical Expert Review Panel"[11] and an example of a panel process to identify concepts can be found in "Developing Measures of Hospital Emergency Preparedness: The Identification of Key Topics for Measurement".[12]

A less resource intensive alternative to a formal panel process is an unstructured feedback mechanism (e.g. a workgroup). However, this method does not provide robust evidence of the face validity of indicators. In this case, experts may be identified to participate in a conference call(s) and review of the indicators. Feedback is provided in an open ended manner.

Risk Adjustment

Time line: Approximately 3 months

Risk adjustment is particularly applicable to outcomes measures. Since outcomes often vary by factors outside the control of the system, such as comorbidities, and these factors often vary systematically, risk adjustment allows for fair comparisons between entities, such as hospitals, as well as more focused quality improvement efforts.

Risk adjustment development begins with a standard model. This includes age and gender, and when available comorbidities and reason for admission. During literature review and panel assessment, potential risk factors for the outcome of interest are identified. First, the standard model is assessed to ascertain whether the important risk factors are adequately addressed. When the factors are theoretically included in the model, but there is question whether or not the model actually accounts for the additional risk, empirical testing is used to assess the residual bias in high-risk groups. For example, APR-DRGs theoretically incorporate severity of illness, but in some cases when we examine high-risk groups identified using ICD-9-CM codes; we have identified residual bias suggesting the risk adjustment is inadequate. At this point, these risk factors, identified using ICD-9-CM codes, can be added to the model. When the model does not include the risk factors of interest, then those risk factors must be specified using the available data if possible. First established algorithms are scanned for applicable definitions (e.g., Clinical Classification Categories[16]). The established algorithms are modified if necessary, or definitions are created *de novo* and tested. It is particularly important to avoid adjusting for the outcome of interest when the definitions of risk factors are confounded with the outcome of interest. For instance, when present on admission data are not available, severity of illness scores may take into account complications of care, making these score unusable in risk adjusting complications outcomes.

Following the specification, an analytic plan is established to assess the model. This may include estimating covariates by applying the model to available data, and testing the calibration and explanatory power of the model.

Alternative Risk Adjustment Approaches

At times there may be important risk factors which are appropriate to include in risk adjustment models for some applications, but may mask disparities in other instances. For instance, socioeconomic status (SES) is a strong predictor of hospitalization as measured by the PQIs, even in the absence of disparities in access to care. Some users may wish to include SES in the risk adjustment model, but other users may not want to mask disparities. In this case, "optional" risk adjustment models may be developed.

A second alternative approach to accounting for risk is risk stratification. Stratification provides additional insight into which patients for which hospitals are performing well, and is often a preferred method for clinical quality improvement.

Example: Specific Pediatric Quality Indicators (PDI) are stratified by high risk, intermediate risk and low risk groups. Panelists preferred the granularity for quality improvement applications of the indicators.

Empirical Analyses

Time line: Approximately 2 months

The empirical analyses serve to provide information regarding the performance of the specification, fill evidence gaps, guide risk adjustment development, and inform guidance for the application of the indicator.

There is a standard set of empirical analyses that are conducted on most indicators. These include:

1. Initial indicator rates
2. Mean hospital or area level rate and variation
3. Measures of precision including signal ratio
4. Measures of reliability including persistence
5. Relationship between the indicator and other quality indicators

In addition, several tailored analyses may be conducted, including:

- Numerator breakdown
- Regression analyses
- Impact of definitional changes
- Exploration of qualifying cases

Information from the literature review (e.g., cases of coding bias), abstraction, previous empirical methods, and expertise (from the team and panel review) are used to develop an analytic plan. Empirical analyses should use a source and type of data as similar as possible to that proposed for the final measure implementation, if possible. For example, for many of the existing AHRQ QIs, the analytic plan was executed with HCUP data provided by AHRQ, which is similar in content and structure to administrative data used by hospitals to assess their performance. An additional component to the empirical analyses may involve validation activities. The validation activities involve medical record abstraction and review to determine the utility of using certain codes, as well as the rigor with which the codes are identifying the information relevant to the topic of interest. These additional validation activities add time and resources requirements to the overall measure development process.

Finalization of Specifications

Time line: Approximately 1 month

The initial specifications developed prior to the panel review are finalized to include evidence from the literature review, panel review, risk adjustment, and empirical analyses. The strengths and weaknesses of each candidate indicator are evaluated, and recommendations to strengthen the candidate indicators are proposed. The strongest candidate indicators are recommended for implementation by the development team. Generally, recommended candidate indicators have high face validity, confirmatory evidence of validity (e.g., from the literature), acceptability to the clinical panel, and adequate performance on empirical analyses.

Summary of Evidence for each Recommended Candidate Indicator

Time line: Approximately 2 months

Using the information from the finalized specifications, a summary of evidence is created for each recommended candidate indicator. The summary includes all relevant evidence gathered over the course of the QI measure development process. The summary of evidence for the recommended candidate indicators helps to facilitate the review and decision process on the candidate indicators.

AHRQ Review and Decision on Indicators

Time line: Approximately 1 month

AHRQ uses the finalized specifications and summary of evidence on the candidate indicators to determine if some or all of the recommended indicators warrant an additional development phase before inclusion in a publicly released AHRQ QI module.

Section 3: Phase II: QI Implementation

If AHRQ endorses the advancement of recommended indicators to Phase II of QI measure development, the focus changes to implementation.

Coding Quality Indicators into Software

Time line: Approximately 1 month

The software team codes the indicators into the software for release to users as a QI module. The QI module is incorporated into the software in a user-friendly manner that is consistent with the implementation of previous QI modules.

Testing

Time line: Approximately 2 months

The newly coded QI module is tested according to current software testing processes to ensure accuracy and consistency. Testing includes identifying and deploying appropriate test datasets for use with the AHRQ QIs. The testing occurs both internally and by an external entity as well. The SAS software is tested side by side with the WinQI software to evaluate the consistency of results produced by both sets of software.

User Documentation

Time line: Approximately 1.5 months

Throughout the coding and testing process, user documentation is developed that includes specifications for each QI, user guides, SAS and WinQI software instruction and establish logs of changes for future revisions from the prior QI version to the current version.

Technical specifications document the full definition of the indicator, which when used with the QI Empirical Methods report, can facilitate the reproduction of an indicator. In the case of the QIs, that includes ICD-9-CM code level definitions of numerators, denominators, and exclusion criteria. The technical specifications also include the specification of risk adjustment factors and the assigned covariates and coefficients. The document is updated with each software release. The specifications are compared with the software at each software release to ensure consistency between the two products.

Each module user guide for summarizes the evidence base for each measure and generally provide brief summaries of the measure definition, and summaries of the literature based evidence, panel review, and empirical analyses. It also includes any indicator specific guidance or caveats of use. The user guide is created during the development process, usually adapted from the final report of indicator development. It is updated as necessary during Phase III.

The SAS and WinQI software instructions provide guidance for formatting a user's dataset for use with the software, including variable specifications and assumed values, explanation of the structure of the software programs, explanation of the intermediate and final output of the program, guidance on interpreting rates and troubleshooting information. The documentation is updated with each annual release.

Two change logs are maintained, and include changes to both the indicator specifications and documentation. The first logs changes to specifications resulting from the annual updates to the ICD-9-CM coding system and to the DRGs. These changes impact the specification when applied to the data the year of the change and forward. The second logs changes to specifications resulting from the other changes to the indicators based on new information, such as scientific evidence, recently convened expert panels on the indicators and user feedback. These changes usually impact the indicators for all years. The logs include which document or software module/indicator was impacted by the change, a description of that change, and reason for the change. This information is intended to help users interpret longitudinal applications of the indicators.

The software and documentation process are subject to quality assurance processes including, but not limited to, internal independent comparison of software syntax and documentation for consistency and external beta testing of software.

Section 4, Part 1: Phase III: QI Maintenance – Preserving Scientific Acceptability

In order for the QIs to remain scientifically acceptable and useful, they must be maintained and potentially enhanced on a regular cycle. QIs need to be updated based on such factors as: recent evidence published in the literature (particularly as publications are made available using the specific QI) user feedback, ICD-9-CM and DRG coding updates, periodic clinical panel review, the NQF endorsement and maintenance process, newly available data and methodological advances in the industry. Each of the material maintenance steps must be considered within the broader measure life cycle.

Evidence

Time line: Throughout task

Continued review of literature relevant to the QIs needs to occur to incorporate current evidence as appropriate. This literature review can take two forms. Periodic systematic literature review (e.g., every 3-5 years) ensures comprehensive evidence review. These literature updates follow the same format as the evidence focused literature review described in Phase II. Generally, the same or similar search strings, selection criteria, and abstraction forms are used. The new evidence is added to the literature review summary created during Phase II, and any needed modifications to the indicators or implementation guidance is added to a log of potential changes for the next software release. This systematic review is also particularly useful during the course of NQF measure maintenance. In addition to systematic reviews team members also informally note literature-based evidence that may arise.

A second line of new evidence stems from user feedback through the QI user support system or through presentations and meetings on the QIs. This source provides rich information regarding the validity, usefulness, and feasibility and potential modifications. Users often reported false positive cases that may spur further investigation or offer suggestions for improving the indicators or making the indicators more useful. When user comments are relevant to potential updates to indicators or implementation guidance, the comment is flagged and logged for consideration in future software releases.

Technical Specification Updates

Time line: Throughout task

The QI technical specifications, risk adjustment covariates and coefficients are updated annually to reflect fiscal year ICD-9-CM and DRG changes, newly introduced or revised data elements per the uniform bill (UB-04), currently available comparative data used for the reference population and various classification systems. The classification systems come from a number of sources. External parties maintain some classification systems used in the QIs, including: (1) 3M All Patient Refined DRGs (APR DRGs),[13] which is used in the Inpatient Quality Indicator (IQI) mortality measures, and (2) Risk Adjustment for Congenital Heart Surgery -1 (RACHS-1),[14] which is maintained by Children's Hospital in Boston and is used in the Pediatric Heart Surgery Mortality (PDI 6) measure. Two systems that are used in the QIs are maintained by AHRQ (HCUP): Comorbidity Software[15] and the Clinical Classification System (CCS).[16] The QI support team maintains several classification systems that are used in the AHRQ QIs. These include: modified DRGs, birth weight, congenital anomalies and several specific classification systems used with select PDIs for stratification and in risk adjustment (i.e. procedure type risk category, pressure ulcer risk category, wound class procedure type, immune-compromised risk category and bloodstream infection risk category).

Area-level QIs draw on population of the county the person resides in to serve as the denominator. Each year updated populations counts by county are provided by the U.S. Census Bureau. The QI update process involves obtaining the currently available county level population and integrating it in to the forthcoming AHRQ QI release.

Composite Updates

Annual updates occur to not only the individual AHRQ QIs, but the composites as well. One update germane to both individual measures is the computation of the signal variance for each measure. In regard to composites, the updated signal variance and data from the current reference population file is used in the weighting of measures within a composite. The user has a number of weighting options when calculating a composite. One of these options is the "NQF weights", which is the weighting system appearing in the composite endorsed by NQF. For the PSI and PDI composites, the weighting of indicators in the composite is based on numerators – i.e., the relative frequency of the numerators of the composite indicators. In the IQI composites, the weights for the Mortality for Selected Procedures and the Mortality for Selected Conditions Composites are based on denominators – i.e., the relative frequency of the denominators of the component indicators. Each year the updated reference data is used to calculate the numerator and denominator weights.

Panel Review

Time line: Throughout task

A periodic clinical review panel is engaged if the evidence reviewed, user feedback, or coding changes warrant a detailed examination of the indicators. For example, a panel may be convened if it becomes apparent that there may be alternate uses for the QI. These panels may take the form of a formal panel review process, as described above or a less formal process, such as a work group.

Example: Following literature review and empirical analyses, the team recommended that the Gastrointestinal Hemorrhage Mortality Indicator be restricted to esophageal varices. Since this change was extensive, an ad hoc panel was formed based on nominations from the QI Listserv. This panel discussed the face validity of this change in an informal format.

National Quality Forum Submission and Maintenance

Time line: Throughout task

NQF submission and endorsement is considered for all QIs developed. QIs that meet the NQF evaluation criteria are given consideration for submission. NQF submission requires a summary of the literature by NQF evaluation criteria,[3] summaries of empirical analyses or other studies (e.g., chart reviews) which provide evidence for indicators. QIs accepted for endorsement enter a regular maintenance and annual review cycle established by NQF. Measure maintenance also requires evidence summaries, including up-to-date literature reviews. Table 1 demonstrates how development and validation tasks feed into NQF evaluation.

Table 1. Relationship of Development and Validation Efforts to National Quality Forum Criteria

CRITERIA	DESCRIPTION OF CRITERIA	DEVELOPMENT AND VALIDATION ACTIVITIES
Importance	• Is the concept important to measure? • Is there opportunity for improvement?	• Structured Panel Review • Literature review: Discussion of importance, rates and variation • Empirical Analyses: Overall rate and variation
Usability	• Does the measure foster true quality improvement instead of gaming or adverse consequences? • Is the measure harmonized with similar measures? • Is the measure meaningful, understandable and useful?	• Structured Panel Review: Assessment of adverse consequences and overall usefulness • Literature Review: Indicator scan, current use
Feasibility	• Does the measure minimize burden? • Is the data collection and implementation feasible?	• Structured Panel Review • Literature review: Current use
Scientific Acceptability	• Is the measure precisely defined? • Is it reliable (test-retest and inter-rater)? • Does the measure demonstrate face validity, construct validity, and predictive validity? • Is there systematic bias and can that bias be address with adjustment? • Does it detect meaningful differences in performance?	• Structured panel review: Review of specification, overall usefulness • Literature review: Reliability, criterion validity, construct validity, potential risk factors • Empirical analyses: rates, reliability, potential bias and risk adjustment evaluations, relatedness of indicators, year to year chart review of criterion validity or construct validity

Newly Available Data and Methodological Advances

Time line: Throughout task

Measurement creates demand for better data, additional data elements and methods, and in turn these data and methods are incorporated into the measures. Recent examples include the addition of Present on Admission for a greatly increased portion of claims and hierarchical modeling.

Members of the QI support team monitor efforts to enhance available data and to improve available methods. In addition, AHRQ and the QI support team have sought out potential data sources (e.g., electronic laboratory values) and convened workgroups of researchers and users to advance methodological approaches.

Section 4, Part 2: Phase IV: QI Retirement

Occasionally AHRQ has retired indicators by removing them from the software and documentation.

Evidence

Time line: Throughout task

Review of literature relevant to the QIs and feedback from users through the QI user support system or through presentations on the QIs occasionally may suggest that an indicator is no longer scientifically acceptable and should be removed from the QI module.

The determination of which QIs are relevant for retirement is an evolving process. Going forward, the QI retirement criteria may include the following:

1. New evidence showing that the measure is no longer scientifically acceptable

    a. Loss of content validity – i.e., the process of care has been shown to be irrelevant or even harmful

    b. Loss of criterion validity – i.e., the available data cannot be used for the intended purpose, and cannot easily be fixed.

    c. Loss of predictive validity – i.e., an outcome no longer matters because it doesn't predict anything important to patients.

    d. Increase in residual bias without the ability to address the bias with improvements to risk adjustment. These indicators may be retained but recommended that they not be used for comparative purposes.

2. Evidence of unanticipated/undesirable consequences of implementing the measure, particularly as a result of manipulation or gaming by providers.

Removal of Quality Indicators from a Module

Time line: Approximately 0.5 month

The software team removes the codes for the retired indicators from the respective module for release to users. The QI is removed from the software and documentation in a user-friendly manner that is consistent with the implementation of previous QI modules.

In some instances, the measures may be reassigned to the "Experimental Quality Indicators" set. These include indicators that have been included in the indicator set for some time, but have concerns that can only be addressed by major development efforts. These indicators may be of interest to researchers or others for further development or trending over time.

Testing

Time line: Approximately 0.5 month

In the event of removing a measure from a composite, the QI are tested according to current software testing processes to ensure accuracy and consistency. Testing includes ensuring that the removal of the indicators did not introduce any unexpected consequences. The composites are re-evaluated for discrimination[1], forecasting[2] and construct validity[3]. The testing occurs both internally and by an external entity as well. The SAS software is tested side by side with the WinQI software to evaluate the consistency of results produced by both sets of software.

User Documentation

Time line: Approximately 0.5 month

User documentation is updated to remove the retired indicator from specifications (i.e., the definition, numerator, denominators, and risk adjustment coefficient tables) for each QI, user guides (i.e., the evidence summaries for each measure) and SAS and WinQI software documentation from the prior QI version to the current version. The retirement of the indicator is noted in the log of changes made to the measure software.

[1] Discrimination is the ability of the composite measure to differentiate performance as measured by statistically significant deviations from the average performance.

[2] Forecasting is the ability of the composite measure to predict performance for each of the component indictors. Ideally, the forecasting performance will reflect the weighting of the components, in the sense that forecasting will maximize differences for the most highly weighted components.

[3] Construct validity is the degree of association between the composite and other aggregate measures of quality. Specifically, our focus is on the consistency in the composites with one another.

Summary

The QI measure development process involves four phases. The first phase is candidate indicator development for an identified topic area of interest. The steps involved in the first phase are: (1) identification of candidate indicators, which includes literature review, expert engagement, and selection of candidate indicators and (2) assessment of candidate indicators, which includes specifications of candidate indicators and existing AHRQ QIs, panel review, risk adjustment, empirical analyses, finalization of specifications, and summary of evidence for each recommended candidate indicator. The second phase is implementation of the QIs into the AHRQ QI software, which involves coding the QIs into the software, testing, and developing user documentation. The third phase is maintenance of the QIs, which involves review of the evidence, technical specification updates, periodic clinical panel review, NQF endorsement submission and maintenance, and newly available data and methodological advances. The final phase is retirement which involves evidence, removing coding from software, testing and user documentation.

The length of the QI measure development process can vary widely depending on the scope of the development efforts. However, on average it is approximately 20 months for development and three months for implementation, with a variable maintenance schedule (Table 2).

Table 2. Quality Indicator Measure Development Time Line

TASK	AVERAGE COMPLETION TIME
Phase I: QI Measure Development	**Approximately 20 months**[a]
Task 1: Identification of Candidate Indicators	
Literature Review	2.5 months
Expert Engagement	Throughout task
Selection of Candidate Indicators	1 month
Task 2: Assessment of Candidate Indicators	
Initial Specifications of Candidate Indicators and Existing QIs	1 month
Panel Review	3.5 months
Risk Adjustment	3 months
Empirical Analyses	2 months
Finalization of Specifications	1 month
Summary of Evidence for each Recommended Candidate Indicator	2 months
AHRQ Review and Decision on Candidate Indicators	1 month
Phase II: Implementation	**Approximately 3 months**[a]
Coding QIs into Software	1 month
Testing	2 month
User Documentation	1.5 months

TASK	AVERAGE COMPLETION TIME
Phase III: QI Maintenance	**Variable**[a]
Evidence	Throughout task
Technical Specification Updates	Once each year
Panel Review	As needed
NQF Submission and Maintenance	As needed
Newly Available Data and Methodological Advances	As needed
Phase IV: QI Retirement	**Variable**[a]
Evidence	Throughout task
Removal of QI from a Module	0.5 months
Testing	0.5 months
User Documentation	0.5 months

[a] This represents the approximate **total** time for QI measure development, given certain tasks run in parallel with each other.

References

1. National Quality Forum Endorsed Standards. http://www.qualityforum.org/Measures_List.aspx. Updated Last Updated Date.
2. National Quality Measures Clearinghouse. http://www.qualitymeasures.ahrq.gov. Updated Last Updated Date.
3. National Quality Forum. Measure Evaluation Criteria. http://www.qualityforum.org/docs/measure_evaluation_criteria.aspx. Accessed March, 2011.
4. AHRQ Summary Statement on Comparative Hospital Public Reporting. http://qualityindicators.ahrq.gov/news/AHRQSummaryStatement.pdf.
5. Guide to the AHRQ Prevention Quality Indicators. http://www.qualityindicators.ahrq.gov/modules/pqi_resources.aspx.
6. Guide to the AHRQ Pediatric Quality Indicators. http://www.qualityindicators.ahrq.gov/modules/pdi_resources.aspx.
7. Guide to the AHRQ Inpatient Quality Indicators. http://www.qualityindicators.ahrq.gov/modules/iqi_resources.aspx.
8. Guide to the AHRQ Patient Safety Indicators. http://www.qualityindicators.ahrq.gov/modules/psi_resources.aspx.
9. Fitch K, Bernstein SJ, Aguilar MD, et al. *The RAND/UCLA Appropriateness Method User's Manual.*: RAND; 2001.
10. Green L, Lewis F. *Measurement and Evaluation in Health Education and Health Promotion*. Mountain View, CA: Mayfield Publishing Company; 1998.
11. Davies S, Schmidt E, Romano PS, Geppert J, McDonald KM. *Expanding the Use of the AHRQ Prevention Quality Indicators.* November, 2009.
12. McDonald KM, Davies S, Chapman T, et al. *Developing Measures of Hospital Emergency Preparedness: The Identification of Key Topics for Measurement* 2010.
13. 3M Corporation. All Patient Refined Diagnostic Related Groups. http://solutions.3m.com/wps/portal/3M/en_US/3M_Health_Information_Systems/HIS/Products/APRDRG_Software/. Accessed March, 2011.
14. Jenkins KJ, Gauvreau K. Center-specific differences in mortality: preliminary analyses using the Risk Adjustment in Congenital Heart Surgery (RACHS-1) method. *J Thorac Cardiovasc Surg.* Jul 2002;124(1):97-104.
15. Agency for Healthcare Research and Quality. Comorbidity Software. http://www.hcup-us.ahrq.gov/toolssoftware/comorbidity/comorbidity.jsp. Accessed March, 2011.
16. Healthcare Cost and Utilization Project. Clinical Classification System for ICD-9-CM. http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp.
17. Billings J, Zeital L, Lukomnik J, Carey T, Blank A, Newman L. Analysis of variation in hospital admission rates associated with area income in New York City: Unpublished Report; 1992.
18. Weissman JS, Gatsonis C, Epstein AM. Rates of avoidable hospitalization by insurance status in Massachusetts and Maryland. *Jama.* 1992;268(17):2388-2394.
19. Millman M, ed. *Committee on Monitoring Access to Personal Health Care Services.* Washington, D.C.: National Academy Press; 1993. Access to health care in America/ Committee on Monitoring Access to Personal Health Care Services, Institute of Medicine.

Appendix 1: Sample Abstraction Fields for Indicator Scan

Indicator Abstraction Database Fields

Items in bold are database fields; other items comprised the associated pull-down menus. The item marked with an asterisk was added after information collection was underway in order to address evolving needs.

- **Indicator**

- **Topic from Final Eval***
  Indicator is associated with which topic?

- **Indicator type**
  Structure
  Process
  Proxy-outcome
  Outcome
  Don’t know

- **Calculation type**
  Rate
  Count
  Continuous
  Yes/No
  Don’t know

- **Sampling**
  No sampling
  Sampling required, and specified
  Sampling required, but not specified
  Don’t know

- **Level of measurement**
  Physician/Department
  Hospital
  Integrated, interface between hospital and other organization
  Population
  Don’t know

- **Time line**
  Pre-incident: (mitigation/preparedness planning)
  Recognition: (recognition, notification)

Response: (mobilization)
Recovery: (demobilization)
Don't know

- **Primary function framework**
  Capability planning (increased complexity or specialized care)
  Capacity planning (increased volume)
  Continuity planning (maintaining non-event related essential services)
  Interoperability planning (partnerships and communication with outside organizations)
  Multiple function areas
  Don't know

- **Substantive framework [fed into concept areas used in evaluation process]**
  Surge capacity/alternate care sites
  Emergency management procedures and plan-making
  Communication systems (including redundant capabilities, both technical and personal)
  Continuity of operations
  Decontamination
  Evacuation/shelter-in-place
  Security/facility access control
  Disease reporting/surveillance
  Countermeasures/medical supplies/personal protective equipment management
  Behavioral health
  Fatality management
  Volunteer/personnel management
  Staff training
  Patient management
  Community integration
  Other
  Don't know

- **Resource framework:**
  Personnel (management and other)
  Supplies (equipment acquisition)
  Information (communication protocols, information gathering)
  Multiple framework areas
  Don't know

- **Hazard**
  All hazard
  Chemical
  Biological
  Radiological
  Natural disaster (flood, hurricane, earthquake)
  Fire or explosion
  Mass casualty accident (transportation accident, workplace accident)
  Other

- **Data source**
  Survey data – self-report
  After Action Report data
  Administrative data
  Audit data
  Other national dataset of existing data
  Chart data
  Don't know

- **Data source available on what % of hospitals**
- **Current use**
  Currently required by federal government of all hospitals
  Currently required by state/local government of all hospitals
  Currently required by accreditation organization
  Currently required of participants in specific program
  Currently used, but not required by federal government
  Currently used, but not required by other organizations
  Proposed in the literature by a subject matter expert
  Don't know

- **Current state of operationalization**
  Operationalized using existing and available data
  Operationalized using existing, but not available data
  Operationalized, but no data exists or available
  Operationalized in concept only
  Operationalization required
  Don't know

- **Risk adjustment**
  Potentially required
  None specified
  Specified but not operationalized
  Operationalized and tested using unavailable data
  Operationalized and tested using available data
  Operationalized and untested using unavailable data
  Operationalized and untested using available data
  Don't know

- **Risk adjustment may not be desired in all applications**
  Yes/No

- **Reliability information available**
  Reliability has been tested and information available
  Reliability has been tested, but information not available
  Reliability has not been tested/unknown

- **Face validity information available**
  Face validity has been tested and information available
  Face validity has been tested, but information not available
  Face validity has not been tested/unknown

- **Construct validity available**
  Construct validity has been tested and information available
  Construct validity has been tested, but information not available
  Construct validity has not been tested/unknown

- **Criterion validity available**
  Criterion validity has been tested and information available
  Criterion validity has been tested, but information not available
  Criterion validity has not been tested/unknown

- **Current level of performance**
  Current performance known
  Current performance estimated
  Past and Current performance unknown
  Past performance known, but current performance unknown
  Current performance known for a subset and unrepresentative cohort of hospitals

- **Hyperlinks**
- **Notes**

Appendix 2: Sample Panel Evaluation Questionnaire

Indicator name: Diabetes Short-term Complications Admission Rate

1. Access barriers may relate to geographic access (i.e., distance, lack of local transportation), temporal access (i.e., after hours care), economic access (i.e., Medicaid providers), or cultural access (i.e., interpreting services). To what extent is this event likely to reflect *poor access* to outpatient care?

1 2 3 4 5 6 7 8 9

Not at all likely Very likely

Comments:

2. Poor quality care may affect such specific domains as screening, diagnosis, treatment, patient education, and follow-up. To what extent is this event likely to reflect *poor quality* outpatient care?

1 2 3 4 5 6 7 8 9

Not at all likely Very likely

Comments:

3. How often are these diagnoses, when they are responsible for the admission, clearly charted in medical records by physicians (e.g., as opposed to using different terminology)?

1 2 3 4 5 6 7 8 9

Never charted Always charted

Comments:

4. To what extent is this indicator subject to bias (meaning that some areas/organizations will be judged as low quality because they systematically differ from other areas/organizations in some aspect, such as the prevalence of a related chronic disease that is not due to poor quality care or poor access to care)?

1 2 3 4 5 6 7 8 9

Not at all biased Very biased

What are the factors that contribute to the bias?

5. Are there ways that areas/organizations could easily appear to better their performance on this indicator, without actually improving the accessibility or quality of care that they provide?

6. Are there adverse outcomes that could result from implementing this indicator? If so, please explain

7. Geographic areas include the states, counties, cities, and zip codes in which patients reside. What is your overall rating of the usefulness of this indicator, for publicly reporting rates at the level of geographic areas?

1 2 3 4 5 6 7 8 9

Highly discourage use Highly recommend use

Please discuss you reasons for assigning the overall rating above.

8. Payor organizations include state Medicaid agencies and their contracted managed care plans, State Children's Health Insurance Program (SCHIP) agencies and their contracted managed care plans, Medicare Advantage plans, and private managed care plans. What is your overall rating of the usefulness of this indicator, for publicly reporting rates at the level of payor organizations?

1 2 3 4 5 6 7 8 9

Highly discourage use Highly recommend use

9. Pay-for-performance programs have been implemented by some state Medicaid and SCHIP agencies to reward contracted managed care plans that facilitate higher quality or more efficient care. What is your overall rating of the usefulness of this indicator, for pay for performance at the level of payor organizations?

1 2 3 4 5 6 7 8 9

Highly discourage use Highly recommend use

Please discuss you reasons for assigning the overall ratings above (q. 8 and 9).

10. Provider organizations include capitated physician organizations and similar entities that provide comprehensive inpatient and outpatient care for a defined population. What is your overall rating of the usefulness of this indicator, for quality improvement within provider organizations?

1 2 3 4 5 6 7 8 9

Highly discourage use Highly recommend use

11. What is your overall rating of the usefulness of this indicator, for comparative public reporting amongst provider organizations?

1 2 3 4 5 6 7 8 9

Highly discourage use Highly recommend use

12. Pay-for-performance programs have been implemented by some managed care organizations to reward contracted physician organizations that provide higher quality or more efficient care. What is your overall rating of the usefulness of this indicator, for pay for performance at the level of provider organizations?

1 2 3 4 5 6 7 8 9

Highly discourage use Highly recommend use

Please discuss you reasons for assigning the overall ratings above (q. 10, 11, 12).

13. Some indicators definitions limit the denominator to exclude patients for which admissions are likely to be unpreventable even with good quality of care, or to focus the indicator on those truly at risk for hospitalization. Are there any patients that should be excluded from this indicator? Do you have any other input on the denominator for this indicator?

14. Would you suggest any changes to the definition of this indicator? Please specify changes and give rationale supporting proposed changes.

15. Is there anything else that you would like us to know about this indicator?

Appendix 3: Example Indicator Information Sheet for Panel Evaluation

DEHYDRATION ADMISSION RATE

Indicator definition:

Number of patients admitted for dehydration.

Included Admissions:

Numerator:

All non-maternal discharges of age 18 years and older with ICD-9-CM principal diagnosis code for hypovolemia (see below).

Volume depletion [276.5]

Exclude: hypovolemic shock – postoperative & traumatic

Volume depletion, unspecified [276.50]

Dehydration [276.51]

Hypovolemia [276.52]

Exclude patients transferring from another institution, MDC 14 (pregnancy, childbirth, and puerperium), or MDC 15 (newborns and neonates)

Denominator:

Area applications: Population in Metro Area or county, age 18 years and older.

Payor/provider applications: All patients, age 18 years and older.

Age and sex risk adjustment is currently incorporated for this indicator. A risk adjustment system for provider organization or health plan applications has not been developed. However, potential risk adjustment could take into account prior hospitalizations, prior ED utilization, diagnosis codes from outpatient records, and potentially pharmacy data. Note that clinical results from laboratory tests are not likely to be available. Please see the risk adjustment rating questionnaire for more details.

Clinical Rationale

This indicator is intended to identify hospitalizations for dehydration. With early interventions including oral rehydration therapy, this complication can often be managed in an outpatient basis.

This indicator was developed as part of the Prevention Quality Indicator measure set, and is adapted from an indicator developed by John Billings[16] and colleagues after favorable evaluation by a physician panel.

Literature Based Evidence

Precipitating events leading to admission may include physiologic causes, as discussed above, or the cessation of treatment due to access to care or non-compliance issues. Evidence that such causes are or are not due to access to care contributes to the construct validity of this indicator. However, such evidence has not been strongly shown. Access to care in relation to admissions has been explicitly studied and reported. Weissman et al.[17] found that uninsured patients had a higher risk of admission for DKA and coma than privately insured patients (age 0-64) (adjusted O.R. 2.18 – 2.77). Two studies of ACSC indicators reported validation work for diabetes independent of measure sets. Millman[18] reported that low-income zip codes had 4.1 times more diabetes hospitalizations per capita (age 0-64) than high-income zip codes in 11 states in 1988. Billings et al.[16] found that low-income zip codes in New York City (where at least 60% of households earned less than $15,000 in 1988, based on adjusted 1980 Census data) had 6.3 times more diabetes hospitalizations per capita (age 0-64) than high-income zip codes (where less than 17.5% of households earned less than $15,000). Household income explained 52% of the variation in short term diabetes complication hospitalization rates at the zip code level. These findings suggest that this indicator may be marker for poor access to outpatient care.

Additional Questions to Consider

Although we are not asking you to state your opinion on this form, there are some questions that we will be discussing in our conference calls on each of the indicators. For this indicator, we will be discussing whether a code for “Uncontrolled diabetes should be included in this indicator.” This code is used in the Healthy People 2010 indicator. We will also be discussing whether the age span identified in the included population is appropriate.