NATIONAL QUALITY FORUM

NQF’s Evaluation Criteria: Discussion on updating criteria and guidance

Karen Johnson
Helen Burstin

May 1, 2017

Updating Criteria and Guidance

- NQF staff encounter issues related to measure evaluation that require greater clarity and possible revision
- CSAC input is needed on the following issues:
  - *Evidence requirement for outcome measures*
  - *Use of the evidence exception*
  - *Evidence v validity for evidence*
  - *Performance gap and use/usability*
  - *Use and usability muss pass for maintenance measures*
  - *Validity – move beyond face validity*
  - *Reliability thresholds*
- Opportunity for CSAC to identify other evaluation issues/concerns

Evidence: Convened Ad Hoc Evidence Advisory Panel to Consider Options

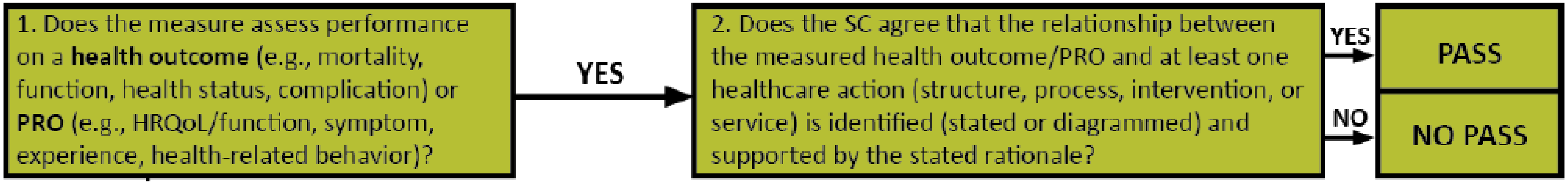
- Follow-up from July 2016 CSAC discussion
- Key questions:
  - *Should we modify the evidence criterion <u>for outcome measures</u> to require at least some empirical evidence?*
    - If so, how?
  - *Should we remove the option to invoke the "exception" to the evidence criterion when there is insufficient evidence to support the measure?*

Current requirements for outcome measures

"A rationale supports the relationship of the health outcome to at least one healthcare structure, process, intervention, or service."

- Applies to health outcomes, patient-reported outcomes
- Does not apply to intermediate clinical outcomes

Algorithm 1. Guidance for Evaluating the Clinical Evidence



1. Does the measure assess performance on a **health outcome** (e.g., mortality, function, health status, complication) or **PRO** (e.g., HRQoL/function, symptom, experience, health-related behavior)?

YES

2. Does the SC agree that the relationship between the measured health outcome/PRO and at least one healthcare action (structure, process, intervention, or service) is identified (stated or diagrammed) and supported by the stated rationale?

YES

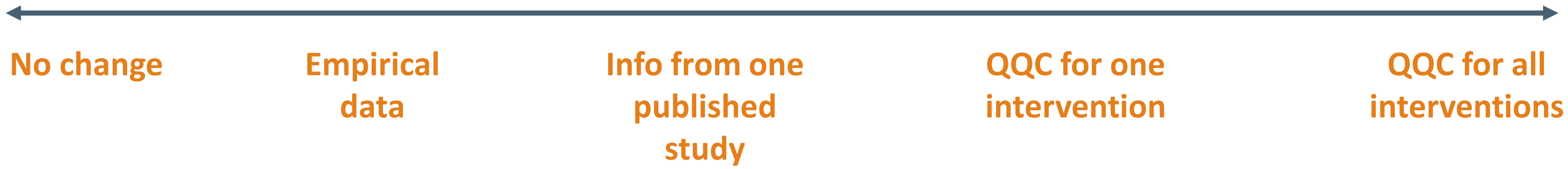
PASS

NO

NO PASS

Ad Hoc Evidence Advisory Panel: Potential Options

- Options could include:



No change

Empirical data

Info from one published study

QQC for one intervention

QQC for all interventions

- Empirical data demonstrate a **relationship** between the outcome and at least one healthcare structure, process, intervention, or service.

OR

- Empirical data demonstrate that at least one healthcare structure, process, intervention, or service **leads to** the desired outcome.

Advisory panel discussion

- Not full consensus among the group
  - *Some want to require evidence*
    - Given their use in accountability applications, outcomes should no longer "get a pass"
    - Concern that things that may sound reasonable could have negative consequences for patient care
  - *Some believe evidence may not be necessary for all outcomes*
    - Some outcomes (e.g. PROs, experience) may be inherently meaningful to patients
    - But there should be actionable interventions on the part of those being measured
- Agree that we need to be careful about how we frame our language (i.e., not a lower bar for outcome measures)

Advisory panel discussion: Discussion Questions

- Is it a “meaningful” outcome?
- Is it “actionable?”
- Is it an “appropriate” end point for particular processes (e.g., hernia repair and mortality)
  - *There may be published evidence showing associations*

Advisory panel: Consideration for CSAC

- Some interest in strengthening the evidence requirement for outcomes
  - *Empirical data demonstrate a **relationship** between the outcome and at least one healthcare structure, process, intervention, or service.*
  - *Consider wide variation as an option if data not available (consensus around this point)*
- Agreement to potentially add some discussion points
  - *Is it meaningful?*
  - *Is it appropriate?*

Evidence Exception

- Panel discussed several options for the exception
  - *Drop option completely?*
  - *Limit its use to certain topic areas or types of measures?*
  - *Interpret current algorithm more stringently?*
  - *Provide more guidance to achieve more consistency in application?*
- May need exception for outcomes measures if we change evidence requirement for outcome measures
- Recommendation: Maintain current approach

Evidence: Importance v Validity

- Evidence currently is considered under two criteria:
  - Evidence subcriterion: process can be linked to desired health outcome
  - Validity subcriterion: measure specifications are consistent with evidence presented
  - For measures that specify a **particular timeframe or threshold**, there may be less evidence for the timeframe/threshold
    - Should this fail a measure on the evidence subcriterion or should this be more appropriately discussed under validity?
    - Example: %SMI discharges w/follow-up visits with a mental health practitioner within 7 and 30 days of discharge. Guidelines address consistent and continuous management of mental illnesses, but not follow-up after hospitalization or appropriate time intervals.
  - Committee members sometimes view validity evidence sub-criterion as another opportunity to fail measure on evidence (opportunity for simplification?)

Performance Gap, Usability and Use

- For maintenance measures, we now have a greater emphasis on Gap and Use/Usability
  - *Less focus on evidence, reliability, validity if previous information meets current requirements*
- Information about current performance and improvement usually missing when:
  - *A steward/developer is not the implementer*
  - *When a measure is not being used*
- Without information on current and past performance
  - *It is very difficult to pass the Gap sub-criterion (must-pass)*
  - *Difficult to be responsive to the improvement portion of the use/usability criterion (although not must-pass, could still fail the criterion)*

Usability and Use: Should this become must-pass for maintenance measures?

- Four subcriteria:
  - *In use in accountability program within 3 years and publicly reported within 6 years*
  - *Demonstrated improvement*
  - *Benefits outweigh evidence of unintended negative consequences to patients*
  - *Measure has been vetted by those being measured or others*

Usability and Use: Should this become must-pass for maintenance measures?

- Potential pros
  - Measurement should drive improvement
  - “Aligns” with current process of greater emphasis on use and usability
  - Probably decrease number of endorsed measures
- Potential cons
  - Developers or stewards that are not implementers may not know if measure in use or cannot obtain improvement data
  - Subjectivity in evaluating benefits over harms and vetting
  - Vetting is still relatively new, and was included in U&U because it is aspirational
    - » Recent appeal of readmission measure by Association of Rehabilitation Nurses due to inability to access patient-level data for improvement

Face validity

- Definition: The subjective determination by experts that, on the face of it, the measure appears to reflect quality of care.
  - *Weakest form of validity testing*
- Current guidance: Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.
  - *Applies to both new and maintenance measures*

Face validity: Should we strengthen evaluation requirements?

- Potential change to criteria
  - *Discontinue face validity option for both new and maintenance measures*

    OR
  - *Continue to allow face validity for initial endorsement but require empirical testing of maintenance measures*

- Both options would be more burdensome for developers
  - *Would likely result in loss of endorsement for potentially large number of measures*
  - *Second option might be reasonable given NQF's strategic direction of prioritizing measures and reflect more graduated approach*

Face validity: Should we strengthen guidance to Committees?

- Should face validity testing results be ignored if empirical results are available?
  - *Some seem to think that if there has been a face validity assessment, then the measure should pass validity*
  - *Others think empirical results should <u>always</u> trump subjective assessments*
    - However, not all testing is equally strong, so this may be too restrictive
    - Consider differentiating between data-element testing and score-level testing

Validity: Strengthen guidance for exclusions?

- Committees interpret exclusion guidance differently
- Clinical/providers tend to support more inclusions for face validity and lower risk of misclassification
- Greater clarity is needed to guide committee decision-making
- Current exclusion criteria:
  - *Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion*
- Current exclusion guidance:
  - *Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion*

Reliability: Consider establishing thresholds?

- Recent readmissions appeal related to reliability results
- In general, NQF is not prescriptive for how measures should meet our criteria
  - *Examples: no particular type of evidence required, no thresholds for testing samples, testing methods, or testing results*
- NQF Measure Testing Task Force (2010) did not set minimum thresholds, but provided basic principles, noted common approaches and “rules of thumb”
- CSAC has previously noted difficulty with determining thresholds and wanted committees to have flexibility to make judgments
  - Most commenters agreed that it is difficult or impossible to identify minimum thresholds that are applicable to all testing situations
- Potential opportunity to emphasize consistent use “rules of thumb” and principles with committees and CSAC