NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

WHAT GOOD LOOKS LIKE — EVIDENCE — PROCESS EXAMPLE #1

The following example is intended only for illustration of the type of information requested for the Steering Committee's evaluation of the evidence. The examples are not intended as requirements or to be replicated exactly—the key point is to provide substantive information and data in the measure submission evidence attachment so it is clear about the evidence that does or does not exist to support the measure focus.

Please contact NQF staff if you have questions, corrections, or suggestions to improve the example.

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Women with urinary incontinence who receive pelvic floor muscle training
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure title

Date of Submission: 6/27/2013

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

EXAMPLE

1a.1.This is a measure of: (*should be consistent with type entered in De.1*)
Outcome

- ☐ Health outcome: Click here to name the health outcome
- ☐ Patient-reported outcome (PRO): Click here to name the PRO
  PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors
- ☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: pelvic floor muscle training for urinary incontinence in women
☐ Structure: Click here to name the structure
☐ Other: Click here to name what is being measured

_____

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) and at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

_____

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

Key Points

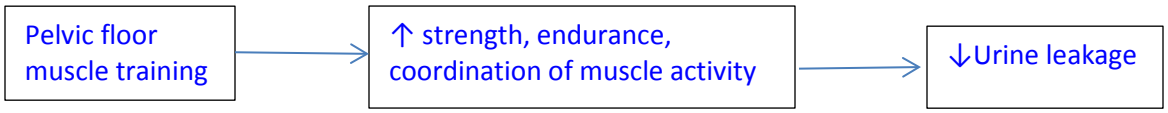
- See NQF guidance for rating quantity, quality, consistency of body of evidence and report from the evidence task force available at the Measure Evaluation webpage.
- A **systematic review** of the evidence is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include quantitative synthesis (meta-analysis), depending on available data (IOM, 2011).
- A body of evidence includes all the evidence for a topic, which is systematically identified, based on pre-established criteria for relevance and quality of evidence.
- Expert opinion is not considered empirical evidence, but evidence is not limited to randomized controlled trials
- There is variability in evidence reviews, grading systems, and presentation of the findings; however, the information should be reported as requested in this form so the Steering Committee can evaluate it according to NQF criteria and guidance.

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

Key Points

- Indicate the causal pathway – do not just make a general statement.

- Do not discuss evidence in this item – it should be presented in the appropriate sections as indicated by the source of the evidence noted in 1a3.1.



PFMT may be prescribed to increase strength (the maximum force generated by a muscle in a single contraction); endurance (ability to contract repetitively, or sustain a single contraction over time); coordination of muscle activity, or timing to suppress urge, or a combination of these.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***
☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***
☒ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***
☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

_____

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Incontinence in women. In: Schröder A, Abrams P, Andersson KE, Artibani W, Chapple CR, Drake MJ, Hampel C, Neisius A, Tubaro A, Thüroff JW. Guidelines on urinary incontinence. Arnhem, The Netherlands: European Association of Urology (EAU); 2009 Mar. p. 28-43.

URL: http://www.uroweb.org/gls/pdf/Urinary%20Incontinence%202010.pdf
http://www.guideline.gov/content.aspx?id=16386&search=urinary+incontinence#Section424

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Key Points
- Do not summarize, paraphrase, or shorten the recommendation

5.2 Initial treatment of urinary incontinence (UI) in women (p.29)
For women with stress, urgency or mixed urinary incontinence, initial treatment includes appropriate lifestyle advice, physical therapy, a scheduled voiding regime, behavioural therapy and medication (Table 7, Figure 3). Some recommendations are based on good and consistent evidence of effect. However, many other recommendations are based on insufficient level 1 or 2 evidence and are essentially hypotheses requiring better evidence of their benefit.

From Table 7: Initial treatment for UI in women
PFMT should be offered as first-line conservative therapy to women with stress, urgency, or mixed UI

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Key Points
• Should include BOTH grade and definition of the grade • Not all grades are on a letter or number scale • Grades for a recommendation are different from grades for quality of evidence (although related) – make sure it is the appropriate grade for a recommendation

A - Based on clinical studies of good quality and consistency addressing the specific recommendations and including at least one randomised trial

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

B - Based on well-conducted clinical studies, but without randomised clinical trials
C - Made despite the absence of directly applicable clinical studies of good quality
Modified from Sackett et al. (2, 3).

1a.4.5. Citation and URL for methodology for grading recommendations and evidence (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***
☒ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations and evidence (*if different from 1a.5.1*)**:**

Complete section 1a.7

_____

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

Cochrane Systematic Review
Dumoulin C, Hay-Smith J; Pelvic floor muscle training versus no treatment or inactive control treatments for urinary incontinence in women; Cochrane Database of Systematic Reviews 2010, Issue 1. Art. No.: CD005654, DOI: 10.1002/14651858.CD005654.pub2.

URL:
http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD005654.pub2/abstract;jsessionid=5B59498CFC062003F250C00B2B8CEFEE.d01t03

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**
The systematic review identified quality of evidence based on risk of bias. System for determining risk of bias was explained in Chapter 8 of Cochrane Handbook for Systematic Reviews for Interventions, 5.0.2, updated September 2009 http://www.mrc-bsu.cam.ac.uk/cochrane/handbook502 .

Complete section 1a.7

_____

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

Key Points

- If more than one systematic review of the evidence identified above (in 1a.4, 1a.5, and 1a.6), you may choose to summarize below the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section.
- If more than one systematic review of the evidence is summarized below, provide a separate response for each review for each question and clearly identify which review is the basis of the response – do not combine systematic reviews.
- If the only systematic review of the body of evidence relevant to your measure does not make details available about the quantity, quality, and consistency of the body of evidence; respond to the following questions with what is known from the systematic review. (For example, it is not useful to report that 5,000 articles were reviewed for an entire guideline because it provides no information on the quantity of studies in the body of evidence for a particular process of care.)

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

The information in the following questions in this section is based on the Cochrane Systematic Review cited in section 1a.6.

The systematic review addressed pelvic floor muscle training for women with urinary incontinence in comparison to no treatment, placebo or sham treatments, or other inactive control treatments.

1a.7.2. Grade assigned for the quality of the quoted evidence <u>with definition</u> of the grade:

Key Points
- Should include BOTH grade and definition of the grade
- Not all grades are on a letter or number scale
- Grades for quality of evidence are different from grades for the recommendation (although related) – make sure it is the appropriate grade for the quality of the body of evidence

An overall grade of methodological quality was not assigned. In the systematic review, individual study quality was graded on a scale for risk of bias.

Based on the reported adequacy of allocation concealment and blinding, two trials appeared to be at low risk (Bø 1999; Castro, 2008), six at moderate risk (Bidmead 2002; Burgio 1998; Burns 1993; Kim 2007; Miller 1998; ; Yoon 2003;), and six at high or possible high risk of bias (Aksac 2003; Henalla 1989; Henalla 1990;Hofbauer 1990; Lagro-Janssen 1991;Wells 1999). Interestingly, the more recent trials tended to be of lower risk for bias based on the trial reports." (p. 20)

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

Risk of bias	Interpretation	Within a study	Across studies
Low risk of bias.	Plausible bias unlikely to seriously alter the results.	Low risk of bias for all key domains.	Most information is from studies at low risk of bias.
Unclear risk of bias.	Plausible bias that raises some doubt about the results.	Unclear risk of bias for one or more key domains.	Most information is from studies at low or unclear risk of bias.
High risk of bias.	Plausible bias that seriously weakens confidence in the results.	High risk of bias for one or more key domains.	The proportion of information from studies at high risk of bias is sufficient to affect the interpretation of results.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: <u>1998-2008</u>

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

14 randomized controlled trials

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

Based on the reported adequacy of allocation concealment and blinding, two trials appeared to be at low risk (Bø 1999; Castro, 2008), six at moderate risk (Bidmead 2002; Burgio 1998; Burns 1993; Kim 2007; Miller 1998; ; Yoon 2003;), and six at high or possible high risk of bias (Aksac 2003; Henalla 1989; Henalla 1990;Hofbauer 1990; Lagro-Janssen 1991;Wells 1999). Interestingly, the more recent trials tended to be of lower risk for bias based on the trial reports." (p. 20)

Methodological quality was evaluated from the trial reports. Therefore, the quality of reporting might have affected the judgement of methodological quality. Two of the included studies were published only as abstracts (Bidmead 2002; Henalla 1990). Limited methodological detail was given, which made it particularly difficult to judge the quality of these trials. In addition, few data were reported.
In one way, it was disappointing that only two trials sufficiently described the randomisation process so that the review authors could be sure there was adequate concealment. On the other hand, it was encouraging, given the difficulties of blinding participants and treatment providers to PFMT, that eight of the 14 studies used blinded outcome assessors. Generally, the proportion of dropout and withdrawals was in the region of 0 to 20%. Sample sizes were small to moderate in 12 of the 14 studies, and only three trials reported an a priori power calculation. Two trials stated that intention to treat principles were used for the primary analysis, and one stated that intention to treat analysis did not change the findings of the primary analysis.

Sensitivity analysis on the basis of trial quality was not considered appropriate in view of the small number of trials contributing to each comparison. It is not known to what extent the variable quality of the trials has affected the findings of the review. It is interesting to note that of all the studies contributing data to the analysis, the largest treatment effect (for cure and improvement, and leakage episodes) was observed in a trial at the high risk of bias. This might be an example of the apparent overestimation of treatment effect (about 30%) observed in trials with inadequate or unclear concealment of random allocation (Egger 2002)." (p. 20)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Meta-analysis was not possible due to study heterogeneity.

Comparison of PFMT versus no treatment, placebo, or control was studied for a variety of outcomes as follows:

Outcome: Patient Perceived Cure – 2 studies with consistent direction in favor of PFMT but differences in magnitude of effect (risk ratio 2.34-16.80)

Outcome: Patient Perceived Cure or Improvement – 3 studies with consistent direction in favor of PFMT but differences in magnitude of effect (risk ratio 2.26-20.0). The authors concluded "Overall, the

differences in likelihood of cure or improvement after PFMT compared to control suggested by the review are sufficient to be of interest to women." (p.18)

Outcome: QoL – 2 studies
Hopkins Symptom Checklist, for psychological distress (SCL-90-R)
Global severity: 50.8 (12.8) vs. 51.4 (10.9); mean difference -0.6, 95% CI -5.3 to 4.1

Norwegian Quality of Life Scale
90.1 (9.5) vs. 85.2 (12.1); mean difference 4.9, 95%CI -1.1 to 10.9

The authors concluded "Based on evidence from single trials, there is improved condition specific QoL in women treated with PFMT compared to controls, but there might be less or no effect on generic QoL." (p.18)

Outcome: Leakage Episodes – 5 studies with consistent direction in favor of PFMT but differences in magnitude of effect. "there were statistically significantly fewer leakage episodes (-0.77 to -2.92) with PFMT" (p.18)

Outcome: Number of Voids per Day – 1 study with significantly fewer (-3.1) with PFMT

Outcome: Number of Voids per Night – 1 study with no significant difference

Outcome: Short pad Test Number Cured – 3 studies with consistent direction in favor of PFMT but differences in magnitude of effect (risk ratios 5.54-16.24)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

Three of four studies that reported adverse events stated there were none with PFMT. The other trial recorded a few minor effects of PFMT (for example discomfort with training), and all of which were reversible with cessation of training. Although randomized trials are probably not the most appropriate way to address safety, neither these data nor the content of PFMT suggest that PFMT is likely to be unsafe. (p. 19)

The authors concluded that "PFMT is better than no treatment, placebo, drug, or inactive control for women with stress, urge, or mixed incontinence. Women treated with PFMT were more likely to report cure or improvement, report better QoL, have ewer leakage episodes per day and have less urine leakage on short pad tests than controls. (p.21)

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

_____

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

WHAT GOOD LOOKS LIKE — EVIDENCE — PROCESS EXAMPLE #2

The following example is intended only for illustration of the type of information requested for the Steering Committee's evaluation of the evidence. The examples are not intended as requirements or to be replicated exactly—the key point is to provide substantive information and data in the measure submission evidence attachment so it is clear about the evidence that does or does not exist to support the measure focus.

Please contact NQF staff if you have questions, corrections, or suggestions to improve the example.

Measure Number (*if previously endorsed*)**:** 38T
Measure Title: Periconception folic acid supplementation
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: 38T

Date of Submission: 6/27/2013

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

EXAMPLE

1a.1.This is a measure of: (*should be consistent with type entered in De.1*)
Outcome
- ☐ Health outcome: 38T
- ☐ Patient-reported outcome (PRO): 38T
  PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors
- ☐ Intermediate clinical outcome (*e.g., lab value*): 38T

☒ Process: Folic acid supplements for women who may become pregnant and in early pregnancy to prevent neural tube defects

☐ Structure: 38T

☐ Other: 38T

_____

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) and at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*)..

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

_____

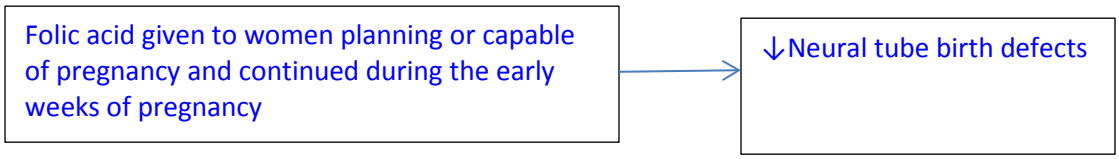
INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

Key Points
- See NQF guidance for rating quantity, quality, consistency of body of evidence and report from the evidence task force available at the Measure Evaluation webpage.
- A **systematic review** of the evidence is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include quantitative synthesis (meta-analysis), depending on available data (IOM, 2011).
- A body of evidence includes all the evidence for a topic, which is systematically identified, based on pre-established criteria for relevance and quality of evidence.
- Expert opinion is not considered empirical evidence, but evidence is not limited to randomized controlled trials
- There is variability in evidence reviews, grading systems, and presentation of the findings; however, the information should be reported as requested in this form so the Steering Committee can evaluate it according to NQF criteria and guidance.

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

Key Points
- Indicate the causal pathway – do not just make a general statement.

- Do not discuss evidence in this item – it should be presented in the appropriate sections as indicated by the source of the evidence indicated in 1a3.1.



Folic acid supplementation in women planning or capable of pregnancy,and continued during the early weeks of pregnancy reduces the occurence of neural tube birth defects.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***
☒ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***
☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***
☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

American College of Obstetricians and Gynecologists (ACOG). Neural tube defects. Washington (DC): American College of Obstetricians and Gynecologists (ACOG); 2003 Jul. 11 p. (ACOG practice bulletin; no. 44). [81 references]

URL: http://www.guideline.gov/content.aspx?id=3994&search=folic+acid+supplement

The American College of Obstetricians and Gynecologists (ACOG) reaffirmed the currency of the guideline in 2008.

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Key Points
- Do not summarize, paraphrase, or shorten the recommendation

Practice bulletin no. 44
Periconceptional folic acid supplementation is recommended because it has been shown to reduce the occurrence and recurrence of neural tube defects (NTDs).

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Key Points

- Should include BOTH grade and definition of the grade
- Not all grades are on a letter or number scale
- Grades for a recommendation are different from grades for quality of evidence (although related) – make sure it is the appropriate grade for a recommendation

Level A - Recommendation is based on good and consistent scientific evidence

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

Level B - Recommendations are based on limited or inconsistent scientific evidence.
Level C - Recommendations are based primarily on consensus and expert opinion.

1a.4.5. Citation and URL for methodology for grading recommendations and evidence (*if different from 1a.4.1*)**:**

Same URL as in 1a.4.1 http://www.guideline.gov/content.aspx?id=3994&search=folic+acid+supplement
See tab for related content

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***
☒ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

_____

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION
1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

U.S. Preventive Services Task Force. Folic Acid for the Prevention of Neural Tube Defects: U.S. Preventive Services Task Force Recommendation Statement. AHRQ Publication No. 09-05132-EF-2, May 2009. http://www.uspreventiveservicestaskforce.org/uspstf09/folicacid/folicacidrs.htm
Folic Acid for the Prevention of Neural Tube Defects: U.S. Preventive Services Task Force Recommendation Statement U.S. Preventive Services Task Force Ann Intern Med May 5, 2009 150:626-631.

URL: http://www.uspreventiveservicestaskforce.org/uspstf/uspsnrfol.htm

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

No numbering provided; date is May 2009
The USPSTF recommends that all women planning or capable of pregnancy take a daily supplement containing 0.4 to 0.8 mg (400 to 800 µg) of folic acid.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

A - The USPSTF recommends the service. There is high certainty that the net benefit is substantial. Offer or provide this service.

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

Grade	Definition	Suggestions for Practice
A	The USPSTF recommends the service. There is high certainty that the net benefit is substantial.	Offer or provide this service.
B	The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial.	Offer or provide this service.
C	*Note: The following statement is undergoing revision.* Clinicians may provide this service to selected patients depending on individual circumstances. However, for most individuals without signs or symptoms there is likely to be only a small benefit from this service.	Offer or provide this service only if other considerations support the offering or providing the service in an individual patient.
D	The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits.	Discourage the use of this service.
I Statement	The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined.	Read the clinical considerations section of USPSTF Recommendation Statement. If the service is offered, patients should understand the uncertainty about the balance of benefits and harms.

1a.5.5. Citation and URL for methodology for grading recommendations and evidence (*if different from 1a.5.1*)**:**

URL: http://www.uspreventiveservicestaskforce.org/uspstf/grades.htm

Complete section 1a.7

_____

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

_____

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

Key Points

- If more than one systematic review of the evidence identified above (in 1a.4, 1a.5, and 1a.6), you may choose to summarize below the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section.
- If more than one systematic review of the evidence is summarized below, provide a separate response for each review for each question and clearly identify which review is the basis of the response – do not combine systematic reviews.
- If the only systematic review of the body of evidence relevant to your measure does not make details available about the quantity, quality, and consistency of the body of evidence; respond to the following questions with what is known from the systematic review. (For example, it is not useful to report that 5,000 articles were reviewed for an entire guideline because it provides no information on the quantity of studies in the body of evidence for a particular process of care.)

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

The information in the following questions in this section is based on the USPSTF review cited in section 1a.5 unless the ACOG is specifically identified.

Folic acid supplementation in pregnant women was studied for its effect on the occurrence of neural tube defects in newborns.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

Key Points

- Should include BOTH grade and definition of the grade
- Not all grades are on a letter or number scale
- Grades for quality of evidence are different from grades for the recommendation (although related) – make sure it is the appropriate grade for the quality of the body of evidence

High Certainty of Net Benefit - The available evidence usually includes consistent results from well-designed, well-conducted studies in representative primary care populations. These studies assess the effects of the preventive service on health outcomes. This conclusion is therefore unlikely to be strongly affected by the results of future studies.

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

Levels of Certainty Regarding Net Benefit

Level of	Description

Certainty*	
High	The available evidence usually includes consistent results from well-designed, well-conducted studies in representative primary care populations. These studies assess the effects of the preventive service on health outcomes. This conclusion is therefore unlikely to be strongly affected by the results of future studies.
Moderate	The available evidence is sufficient to determine the effects of the preventive service on health outcomes, but confidence in the estimate is constrained by such factors as: • The number, size, or quality of individual studies. • Inconsistency of findings across individual studies. • Limited generalizability of findings to routine primary care practice. • Lack of coherence in the chain of evidence. As more information becomes available, the magnitude or direction of the observed effect could change, and this change may be large enough to alter the conclusion.
Low	The available evidence is insufficient to assess effects on health outcomes. Evidence is insufficient because of: • The limited number or size of studies. • Important flaws in study design or methods. • Inconsistency of findings across individual studies. • Gaps in the chain of evidence. • Findings not generalizable to routine primary care practice. • Lack of information on important health outcomes. More information may allow estimation of effects on health outcomes.

* The USPSTF defines certainty as "likelihood that the USPSTF assessment of the net benefit of a preventive service is correct." The net benefit is defined as benefit minus harm of the preventive service as implemented in a general, primary care population. The USPSTF assigns a certainty level based on the nature of the overall evidence available to assess the net benefit of a preventive service.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: 1992-2009

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

Initially 1 large randomized, controlled trial (RCT) for the 1996 review.
The recent evidence synthesis included 4 studies published since 1996: 1 cohort study, 2 case control studies, and 1 meta-anlalysis.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

One cohort study rated as fair quality.
Studies will be graded "fair" if any or all of the following problems occur, without the fatal flaws noted in the "poor" category : Generally comparable groups are assembled initially but some question remains

whether some (although not major) differences occurred with follow-up; measurement instruments are acceptable (although not the best) and generally applied equally; some but not all important outcomes are considered; and some but not all potential confounders are accounted for. Intention to treat analysis is done for RCTs.

Two case control studies – one rated fair quality and one rated good quality.
Good: Appropriate ascertainment of cases and nonbiased selection of case and control participants; exclusion criteria applied equally to cases and controls; response rate equally to or greater than 80 percent; diagnostic procedures and measurements accurate and applied equally to cases and controls; and appropriate attention to confounding variables.
Fair: Recent, relevant, without major apparent selection or diagnostic work-up bias but with response rates less than 80 percent or attention to some but not all important confounding variables.

One meta-anlaysis rated fair quality.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

The Czeizel cohort study reported that 1 NTD and 9 NTDs occurred in the supplemented and unsupplemented women, respectively, for an adjusted odds ratio (aOR) of 0.11 (95% CI, 0.01–0.91); the odds ratio (OR) was adjusted for birth order, chronic maternal disorders, and history of previous fetal death or congenital abnormality. The meta-analysis also found a protective effect of folic acid-containing multivitamins in NTDs with an OR of 0.67 (95% CI, 0.58–0.77) in case-control studies and an OR of 0.52 (0.39–0.69) in RCTs and cohort studies. Both the Czeizel study and the meta-analysis found a statistically significant association between folic acid supplementation and a reduction in cardiovascular congenital abnormalities. In addition, there was a significant effect of folic acid-containing multivitamin use on congenital limb defects in the meta-analysis. No consistent effect of folic acid-containing multivitamins, either on orofacial clefts or on urinary tract congenital abnormalities, was seen in the Czeizel study or the meta-analysis.

The 1995 case-control study reported an OR of 0.65 (95% CI, 0.45–0.94) for use of folic acid-containing supplements in the 3 months before conception, and an OR of 0.60 (95% CI, 0.46–0.79) for supplement use in the 3 months after conception. The 2003 study by Thompson and colleagues reported an OR of 0.55 (0.25–1.22) for regular use, and an OR of 0.92 (0.55–1.55) for some use of folic acid-containing supplements, but neither of these findings was statistically significant. Several differences in these case-control studies may explain differences in results. The 2003 Thompson study was smaller and adjusted for dietary folate intake. Additionally, the exposure timeframes were different: the Shaw study measured exposure in 2 time frames, 3 months before and 3 months after conception, while the Thompson study combined these same 6 months of periconception time into one measure of exposure.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

USPSTF:
The recommendation statement concluded: “Adequate evidence suggests that folic acid from supplementation at usual doses is not associated with serious harms. In its current review, the USPSTF found no evidence on drug interactions, allergic reactions, or carcinogenic effects.”

The evidence synthesis found one fair quality retrospective cohort study that addressed whether folic acid supplementation in women of childbearing age increases the risk of harmful outcomes for either the woman or the infant. After adjusting for age and parity, the authors reported an OR of 1.59 (95% CI 1.41–1.78) for twin delivery after preconceptional folic acid supplementation. In a subgroup analysis of women who did not report IVF, the risk of twinning was lower and non-significant (OR 1.13, 95% CI 0.97–1.33), as expected given the increase in multiple gestation associated with IVF and other assisted reproductive technologies. The odds of having twins of unlike sex, an outcome used as a proxy for dizygotic twinning, were increased in women taking folate, (OR 1.43, 95% CI 1.12–1.83). The authors then adjusted for both a 45% underreporting of supplementation as well as an estimated 12.7% of unidentified IVF pregnancies. When the likely underreporting for folic acid use and IVF were accounted for, the OR for twin delivery after preconceptional supplementation fell to 1.02, and was no longer statistically significantly greater than the risk for women who did not take folic acid (95% CI, 0.85–1.24).

ACOG:
Risks of folic acid supplementation. The risks of higher levels of folic acid supplementation are believed to be minimal. Folic acid is considered nontoxic even at very high doses and is rapidly excreted in the urine. There have been concerns that supplemental folic acid could mask the symptoms of pernicious anemia and thus delay treatment. However, folic acid cannot mask the neuropathy typical of this diagnosis. Currently, 12% of patients with pernicious anemia present with neuropathy alone. With folic acid supplementation, this proportion may be increased, but there is no evidence that initiating treatment after the development of a neuropathy results in irreversible damage. A small number of women taking seizure medication (diphenylhydantoin, aminopterin, or carbamazepine) may have lower serum drug levels and experience an associated increase in seizure frequency while taking folic acid supplement. Monitoring drug levels and increasing the dosage as needed may help to avert this complication.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

_____

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

WHAT GOOD LOOKS LIKE — EVIDENCE — HEALTH OUTCOME EXAMPLE

The following example is intended only for illustration of the type of information requested for the Steering Committee's evaluation. The examples are not intended as requirements or to be replicated exactly—the key point is to provide substantive information in the measure submission evidence attachment so it is clear about the relationship of the health outcome to healthcare structures, processes, interventions, or services**.**

Please contact NQF staff if you have questions, corrections, or suggestions to improve the examples.

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: hospital readmission
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/title

Date of Submission: 6/27/2013

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus
The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

EXAMPLE

1a.1.This is a measure of: (*should be consistent with type entered in De.1*)
Outcome

- ☒ Health outcome: hospital readmission
- ☐ Patient-reported outcome (PRO): Click here to name the PRO
  PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors
- ☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name the process
☐ Structure: Click here to name the structure
☐ Other: Click here to name what is being measured

_____

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

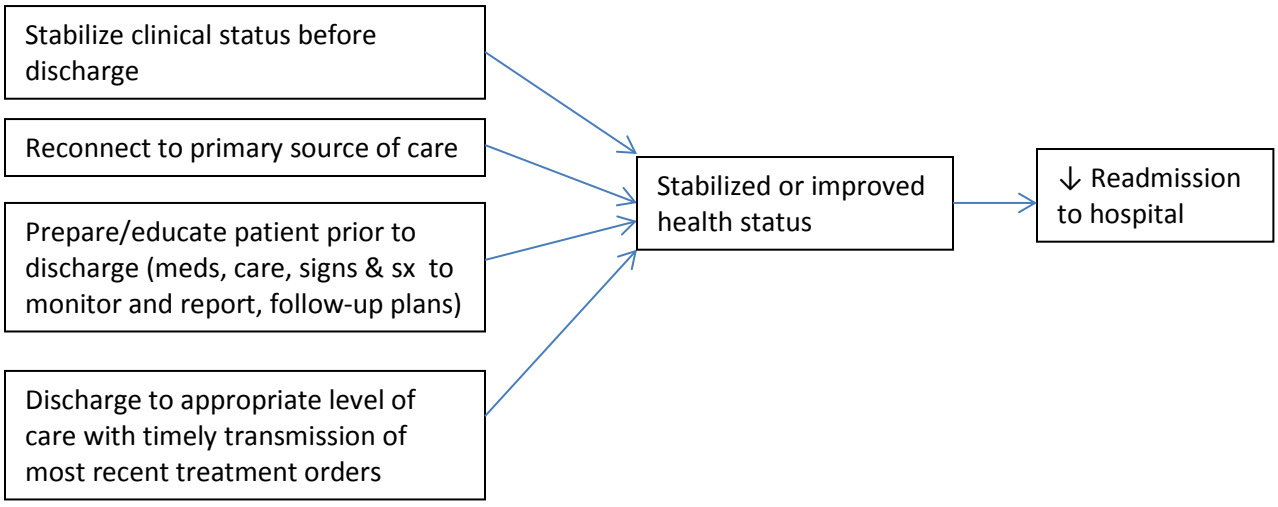
Key Points

- A health outcome is an end-result (e.g., mortality, complication, function, health status; or sometimes a proxy for health outcome such as hospital admission).
- The health outcome or PRO must be linked to at least one healthcare structure, process, intervention, or service.
- Indicate the causal pathway – do not just make a general statement.
- Multiple processes may influence a health outcome or PRO – not all need to be included – focus on those with the strongest rationale.
- Do not include rationale or evidence in this item – address rationale in the next item (1a.2.1).

EXAMPLE 1

Hospital readmission is considered a proxy for the health outcome of deterioration in health status.

Multiple care processes can influence deterioration in health status after discharge resulting in hospital readmission.



EXAMPLE 2
Comprehensive care transition management/care coordination can lead to decreased hospital readmissions as described below.

Comprehensive care transition management/ care coordination

Leads to
↓

Early reconnection to primary care; appropriate level of follow-up care; patient understanding of self-monitoring, self-management, & follow-up care

Leads to
↓
Continuity of treatment plan; early identification & intervention for adverse changes

Leads to
↓

Stable/improved health status

Leads to
↓

Decreased likelihood of readmission

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) and at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Key Points
- The rationale should support the linkages described in 1a.2 above.
- The rationale may refer to the evidence that exists but systematic reviews of evidence are not required.
- For health outcomes, you are not required to complete following items about a systematic review of body of evidence, but may do so for any of the healthcare structures, processes, interventions, or services identified above as influencing the outcome.

EXAMPLE (adapted from NQF # 1789, CMS**)**
Randomized controlled trials have shown that improvement in the following areas can directly reduce readmission rates: quality of care during the initial admission; improvement in communication with patients, their caregivers and their clinicians; patient education; predischarge assessment; and coordination of care after discharge. Evidence that hospitals have been able to reduce readmission rates through these quality-of-care initiatives illustrates the degree to which hospital practices can affect readmission rates. Successful randomized trials have reduced 30-day readmission rates by 20-40% [4-14].

Since 2008, 14 Medicare Quality Improvement Organizations have been funded to focus on care transitions, applying lessons learned from clinical trials. Several have been notably successful in reducing readmissions. The strongest evidence supporting the efficacy of improved discharge processes and enhanced care at transitions is a randomized controlled trial by Project RED (Re-Engineered Discharge), which demonstrated a 30% reduction in 30-day readmissions. In this intervention, a nurse was assigned to each patient as a discharge advocate, responsible for patient education, follow-up, medication reconciliation, and preparing individualized discharge instructions sent to the patient's primary care provider. A follow-up phone call from a pharmacist within 4 days of discharge was also part of the intervention [4].

Given that studies have shown readmissions to be related to quality of care, and that interventions have been able to reduce 30-day readmission rates, it is reasonable to consider an all-condition readmission rate as a quality measure.

References:

1. Jencks SF, Williams MV, Coleman EA. Rehospitalizations among patients in the Medicare fee-for-service program. New England Journal of Medicine 2009;360(14):1418-28.
2. Benbassat J, Taragin M. Hospital readmissions as a measure of quality of health care: advantages and limitations. Archives of Internal Medicine 2000;160(8):1074-81.
3. Medicare Payment Advisory Commission (U.S.). Report to the Congress promoting greater efficiency in Medicare. Washington, DC: Medicare Payment Advisory Commission, 2007.
4. Jack BW, Chetty VK, Anthony D, Greenwald JL, Sanchez GM, Johnson AE, et al. A reengineered hospital discharge program to decrease rehospitalization: a randomized trial. Ann Intern Med 2009;150(3):178-87.
5. Coleman EA, Smith JD, Frank JC, Min SJ, Parry C, Kramer AM. Preparing patients and caregivers to participate in care delivered across settings: the Care Transitions Intervention. J Am Geriatr Soc 2004;52(11):1817-25.
6. Courtney M, Edwards H, Chang A, Parker A, Finlayson K, Hamilton K. Fewer emergency readmissions and better quality of life for older adults at risk of hospital readmission: a randomized controlled trial to determine the effectiveness of a 24-week exercise and telephone follow-up program. J Am Geriatr Soc 2009;57(3):395-402.
7. Garasen H, Windspoll R, Johnsen R. Intermediate care at a community hospital as an alternative to prolonged general hospital care for elderly patients: a randomised controlled trial. BMC Public Health 2007;7:68.
8. Koehler BE, Richter KM, Youngblood L, Cohen BA, Prengler ID, Cheng D, et al. Reduction of 30-day postdischarge hospital readmission or emergency department (ED) visit rates in high-risk elderly medical patients through delivery of a targeted care bundle. J Hosp Med 2009;4(4):211-218.
9. Mistiaen P, Francke AL, Poot E. Interventions aimed at reducing problems in adult patients discharged from hospital to home: a systematic metareview. BMC Health Serv Res 2007;7:47.
10. Naylor M, Brooten D, Jones R, Lavizzo-Mourey R, Mezey M, Pauly M. Comprehensive discharge planning for the hospitalized elderly. A randomized clinical trial. Ann Intern Med 1994;120(12):999-1006.
11. Naylor MD, Brooten D, Campbell R, Jacobsen BS, Mezey MD, Pauly MV, et al. Comprehensive discharge planning and home follow-up of hospitalized elders: a randomized clinical trial. Jama 1999;281(7):613-20.
12. van Walraven C, Seth R, Austin PC, Laupacis A. Effect of discharge summary availability during post-discharge visits on hospital readmission. J Gen Intern Med 2002;17(3):186-92.

13. Weiss M, Yakusheva O, Bobay K. Nurse and patient perceptions of discharge readiness in relation to postdischarge utilization. Med Care 2010;48(5):482-6.
14. Krumholz HM, Amatruda J, Smith GL, et al. Randomized trial of an education and support intervention to prevent readmission of patients with heart failure. J Am Coll Cardiol. Jan 2 2002;39(1):83-89.

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***
☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***
☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***
☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***
☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

_____

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

_____

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

_____

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

_____

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

WHAT GOOD LOOKS LIKE — MEASURE TESTING — VARIOUS EXAMPLES

The following examples are only for illustration of the type of information requested for the Steering Committee's evaluation of reliability and validity. The examples are not intended as requirements or to be replicated exactly—the key point is to provide substantive information and data in the measure submission testing attachment so it is clear about the testing that was conducted and the results.

The examples in this form represent multiple measures, not just one measure submission. They may be from actual measure submissions in whole or part, fictitious, or a combination. Please note that the example testing results shown should not necessarily be construed as demonstrating adequate demonstration of reliability or validity.

Please contact NQF staff if you have questions, corrections, or suggestions to improve the examples.

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Examples for measure testing form
Date of Submission: 6/27/2013
Type of Measure: Check only one – the examples in this form derive from process and outcome measures

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high

proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. 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; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):
- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR
- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;
OR
there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes
10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). 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.
12. 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.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

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

Key Points
- See NQF guidance for rating reliability and validity available at the Measure Evaluation webpage.
- Testing is about the measure ***as specified***.
- Testing is evaluated based on whether: 1) the method is appropriate for the specified measure, 2) the sample is representative and of sufficient size, and 3) the results demonstrate adequate results (e.g., reliability and validity).
- Be sure to explain what you are testing, not just name an analysis.
- Be sure to interpret the results in light of norms for the particular analysis.
- *Please refer to the Measure testing Task Force Report for more information or contact NQF staff.*

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE
Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

Key Points
- The data type(s) and levels of analysis checked below should be consistent with the measure specifications.
- The samples used for testing should be representative of the entities whose performance will be measured and the patients served.
- The sample sizes should be of sufficient size for the statistical tests that are used.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☒ clinical database/registry	☒ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Example:
Medicare Home Health Outcome and Assessment Information Set (OASIS)

1.3. What are the dates of the data used in testing? March-June 2009

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

Example:
20 home health agencies representing various types, locations, and sizes were included in the testing. Agencies were recruited through national and state associations and were selected based on capacity to do the reliability study with 20-40 patients and representing various ownership, locations, and size.

Ownership
4 – private, for-profit
2 – public for-profit chain
6 – private nonprofit
1 – health dept.
5 – hospital-based
2 – visiting nurse associations

Location
Located in 4 states: AZ, MO, NY, TX

Size
3 – less than 10,000 visits/year
10 – 10,000-30,000
7 – greater than 30,000

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

Example:
20-40 patients per agency for a total of 500 patients

Patient case-mix characteristics in the testing sample were similar to national averages with few significant differences.

Characteristic	National Mean	Sample Population
	N=3,289,067	N=500
Age	72.78	70.75*
Female	62.9%	69.4%**
Prior short-stay acute care hospitalization	27.2%	27.3%
Lives alone	32.4%	33.3%
Multiple hospitalizations	3.3%	1.8%
History of falls	56.8%	62.9%
No primary caregiver	34.5%	34%

* p<.01; ** p<.001

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Example:
The dataset described above was used for all aspects of testing.

2a2. RELIABILITY TESTING

***<u>Note</u>**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

Key Points

- Empirical reliability testing of the measure as specified is required. However, empirical validity testing of the data elements (compared to an authoritative source) may also be used for reliability testing.
- Reliability testing addresses random error in measurement.
- Reliability testing should be consistent with the measure specifications (including all specified data types and levels of analysis).
- Reliability testing could be conducted for the critical data elements, or the performance measure score, or both.
- Reliability testing at the data element level must include ALL critical data elements for numerator

and denominator (e.g., interrater agreement).

- Reliability testing at the level of the performance measure score addresses measurement error relative to the quality signal (e.g., signal-to-noise, interunit reliability, ICC).
- Some testing may not be applied as intended (e.g., percent agreement without kappa to adjust for random agreement; inter-rater agreement of only the final score does not address all critical data elements and does not adequately address error relative to quality signal).
- Some methods may not be applicable to the context of performance measures (e.g., consistency/stability of performance measure scores over time based on different patients and in the context of performance improvement).

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*) **Example 1**
☒ **Performance measure score** (e.g., *signal-to-noise analysis*) **Example 2**

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Example 1 (data elements):
Inter-rater reliability was assessed for the critical data elements used in this measure to determine the amount of agreement between 2 different nurses' assessments of the same patient.

Patients were randomly selected from the planned visits for start or resumption of care and discharge assessments for each day of the study.

The first nurse assessed the patient on the usual visit. The second nurse visited the patient and conducted the same assessment within 24 hours of the first assessment so as not to confound differences in assessment with real changes in the patient.

Data analysis included:

- Percent agreement
- Kappa statistic to adjust for chance agreement for categorical data

Example 2 (performance measure score):
Reliability was calculated according to the methods outlined in a technical report prepared by J.L. Adams titled "The Reliability of Provider Profiling: A Tutorial" (RAND Corporation, TR-653-NCQA, 2009). In this context, reliability represents the ability of a measure to confidently distinguish the performance of one physician from another. As discussed in the report: "Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of variability in measured performance that can be explained by real differences in performance. There are 3 main drivers of reliability; sample size, differences between physicians, and measurement error."

According to this approach, reliability is estimated with a beta-binomial model. The beta-binomial model is appropriate for measuring the reliability of pass/fail measures such as those proposed.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Example 1:

Data Element	N	Percent Agreement	Kappa (95% CI)
Functional status score for ambulation	500	85%	0.62 (.55,.69)
Functional status score for ambulation prior to this start/resumption of care	495	83%	0.55 (.39, .71)
Primary diagnosis major diagnostic category	500	90%	0.70 (.62, .78)
Pain scale	500	88%	0.69 (.64, .74)
Location prior to this start/resumption of care	500	91%	0.72 (.60, .84)

Example 2:

Clinic-specific reliability results for the "Prescription of HIV antiretroviral therapy" measure are detailed in Table 1 below.

Table 1: Clinic-Specific Reliability for ART Measure – Year 2010

Between-clinic variance: 0.0040

Clinic	Number of patients	Performance measure rate	Reliability Statistic from signal-to-noise analysis (95% CI)
A	2930	83.7	0.99 (.98,.995)
B	366	98.6	0.99 (.97, .995)
C	2099	79.2	0.98 (.97,.99)
D	438	92.9	0.96 (.94, .97)
E	1586	90.8	0.99 (.98, .995)
F	595	89.6	0.96 (.95, .97)
G	1552	83.1	0.98 (.97, .99)
H	1739	91.3	0.99 (.98, .995)
I	2149	92.6	0.99 (.98,.995)
J	527	88.2	0.95 (.94, .96)
K	4116	90.1	0.99 (.985, .994)
Peds	595	76.5	0.93 (.91, .94)
Median (Range)			**0.98 (0.93-0.99)**

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Example 1:

A statistical measure of inter-rater reliability is Cohen's Kappa, which ranges generally from 0.0 to 1.0 (although negative numbers are possible), where large values mean better reliability and values near zero suggest that agreement is attributable to chance alone. It indicates the proportion of agreement not expected by chance alone (e.g., kappa of 0.6 means that raters agreed 60% of the time over and above what would be expected by chance alone).

Landis & Koch, 1977 offers the following classification of Kappa interpretation.
< 0 Poor agreement
0.00 – 0.20 Slight agreement
0.21 – 0.40 Fair agreement
0.41 – 0.60 Moderate agreement
0.61 – 0.80 Substantial agreement
0.81 – 1.00 Almost perfect agreement

Other authors (Cicchetti & Sparrow; Fleiss) have suggested additional classifications for interpreting the Kappa statistic, but all seem to indicate kappa >0.60 is desirable.

The results of our interrater analysis ranged from Kappa = 0.55 to 0.72. All values except one were in the range of substantial agreement based on the Landis & Koch categorization; only one value was below .61 with a fairly wide confidence interval (0.39, .71). We believe these results demonstrate acceptable reliability of the assessment data used in the performance measure.

Example 2:
Clinic-specific reliability is consistently greater than 0.9, and thus can be considered to be very good.

Reliability scores vary from 0.0 to 1.0, with a score of zero indicating that all variation is attributable to measurement error (noise, or variation across patients within providers) whereas a reliability of 1.0 implies that all variation is caused by real difference in performance across accountable entities.

There is not a clear cut-off for minimum reliability level. Values above 0.7, however, are considered sufficient to see differences between some physicians (or clinics) and the mean, and values above 0.9 are considered sufficient to see differences between pairs of physicians (in this case clinics) (see RAND tutorial, 2009).

2b2. VALIDITY TESTING

Key Points

- Empirical validity testing of the measure as specified is preferred over face validity.
- Validity testing could be conducted for the critical data elements, or the performance measure score, or both.
- Validity testing at the data element level should include ALL critical data elements for numerator and denominator and often is based on assessing agreement between the data elements used in the measure compared to the data elements in an authoritative source (e.g., sensitivity, specificity, positive predictive value, negative predictive value).
- Validity at the level of the performance measure score refers to the correctness of conclusions about quality that can be made based on the score (i.e., a higher score on a quality measure reflects higher quality) and generally involves testing hypotheses based on the theory of the construct (e.g., correlation with performance measures hypothesized to be related or not related; testing the difference in performance measure scores between groups known to differ on quality as assessed

by some other performance measure).

- Face validity is the weakest demonstration of validity and therefore, is subject to challenge by other groups of experts. Face validity is an option ONLY IF: 1) it is systematically assessed; AND 2) it is assessed for the performance measure score resulting from the measure as specified (will scores on the performance measure distinguish quality; not just that the measure concept is a good idea or the data elements are appropriate or feasible to collect).

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

☒ **Empirical validity testing** **Example 1**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **Example 2**

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Example 1 (Empirical validity testing of the measure score):
Validity testing of the hospital score on the process measure of timely reperfusion in AMI patients was conducted by correlation analysis to determine the association between the process and the outcome of 30-day mortality. The hypothesized relationship is that better scores on timely reperfusion should be associated with lower scores on risk-adjusted mortality. Because different numbers of patients were eligible for the process measures at different hospitals, all analyses in which the hospital was the unit of analysis were weighted by the total number of patients from that hospital who were included in the calculation of process measures. We report the correlation coefficient and the percentage of the hospital-specific variation in risk-standardized mortality rates explained (i.e., the square of the correlation coefficient) as indicators of the strength of the associations.

The mortality measure used was the risk standardized 30-day mortality rate using a hierarchical generalized linear model (HGLM).

We conducted various secondary analyses to help interpret results, including correlation with six other process measures (beta blocker at admission/discharge; aspirin at admission/discharge, ACEI at discharge, smoking cessation counseling).

See published report provided in appendix: Bradley EH, Herrin J, Elbel B, et al., Hospital quality for acute myocardial infarction: correlation among process measures and relationship with short-term mortality, JAMA, 2006;296(1):72-78.

Example 2 (Systematic assessment of face validity of the performance measure score):
Face validity of the measure score as an indicator of quality was systematically assessed as follows. After the measure was fully specified, a group of experts (other than those who advised on measure development) was assembled to rate face validity. The 20 experts included 15 clinicians who would be

evaluated on this performance measure, 2 patients, and 3 purchasers (the list of experts is included in the attached appendix).

We provided the detailed measure specifications to the experts and asked them to rate their agreement with the following statement: The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality

The rating scale had 5 levels (1-5) with the following narrative anchors:
1=Disagree; 3=Moderate Agreement; 5=Agree

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Example 1 (Validity testing at level of measure score):
Descriptive statistics for timely reperfusion therapy:
N=709 hospitals
mean=54.5 (SD=13.3); 25th percentile=45.5; median=53.9; 75th percentile=63.9

Correlation coefficient between hospital rates for timely reperfusion and 30-day mortality = -0.18 ($p<.001$);
Percentage of variation in mortality explained by timely reperfusion = 3.2%

Additional analyses
Timely reperfusion was positively correlated with the other process measures (except for smoking) with $p<.001$

- beta blocker at admission (0.17)
- beta blocker at discharge (0.30)
- aspirin at admission (0.21
- aspirin at discharge (0.37)
- ACEI at discharge (0.30)
- smoking cessation counseling (0.13)

Example 2 (Face validity):
The results of the assessment of face validity indicate that an independent group of experts (i.e., different from those who advised on measure development) had high levels of agreement with the statement: "The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality."

Mean rating = 4.75 (out of 5)

Rating Scale	Number who selected the rating
1 - Disagree	0
2	0
3 – Moderate agreement	1
4	3
5 - Agree	16
Total	20

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Example 1 (Empirical testing of the measure score)
The correlation with 30-day mortality (-.18) is negative and significant. It is in the hypothesized direction – as hospital performance on timely perfusion goes up, mortality goes down. The squared correlation (.0324) means that 3.2% of the variability in risk standardized 30-day mortality is shared with timely perfusion. This means that other factors are associated with mortality but does not negate the use of this performance measure as one indicator of quality (not the only one).

Timely reperfusion was also positively correlated with other evidence-based processes of care.

Although this one process measure score for timely reperfusion cannot be used to infer 30-day mortality, the results do not negate the importance of continuing to measure it given the strong evidence base and until research identifies process performance measures with stronger links to outcomes. The positive but modest correlations also indicate the measure is not providing redundant information.

Example 2 (Face Validity):
This measure was examined through a group of experts. Out of the 20 participants, 16 (80%) agreed at the highest level that the scores from the measure as specified would provide an accurate reflection of quality and none disagreed.

_____

2b3. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b4*

Key Points
- Exclusions can affect the validity of the performance measure as well as burden (feasibility) and use.
- Analysis of exclusions should identify to what extent patients from the target population are excluded (overall frequency) and to what extent exclusions vary across the measured entities (frequency distribution).

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

Example:
To identify a homogeneous cohort of patients undergoing elective primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA) procedures, the measure excludes patients who had a principal discharge diagnosis on the index admission indicative of a non-elective arthroplasty (e.g., hip fracture, mechanical complication). The measure also excludes patients who had a procedure code for an arthroplasty procedure that is not an elective primary arthroplasty (e.g., partial hip arthroplasty, revision procedures) or represents a different procedure (e.g., hip resurfacing, removal of implanted device).

After excluding the above admissions to select elective primary THA/TKA procedures, the following exclusions were analyzed for frequency and variability across providers:

- Without at least 12 months pre-index admission enrollment in Medicare FFS
- Without at least 30 days post-discharge enrollment in Medicare FFS
- Who were transferred in to the index hospital
- Who were admitted for the index procedure and subsequently transferred to another acute care facility
- Who leave the hospital against medical advice (AMA)
- With more than two THA/TKA procedures codes during the index hospitalization
- Who die during the index admission

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Example:
We examined overall frequencies and proportions of the admissions excluded for each exclusion criterion in all THA/TKA admissions in 2008-2010 Medicare fee-for-service data. The initial cohort included 1,404,143 admissions. After excluding patients that were not undergoing elective THA/TKA (i.e. patients with concurrent procedures or diagnoses that were indicative of non-elective primary arthroplasty), the cohort included 1,027,565 admissions. The final cohort, after additional patient exclusions, included 897,321 admissions. Categories are not mutually exclusive.

From among 1,027,565 admissions in 1000 hospitals

Exclusion	**Overall Occurrence N**	**Overall Occurrence %**	**Distribution Across Hospitals 25th, 50th, 75th percentile**
In-hospital deaths	1,208	0.12%	0.09, 0.11, 0.13
Patients with incomplete administrative data in 12 months prior to index hospitalization	115,632	11.25%	10.0, 11.0, 12,0
Transfer-out patients	10,851	1.06%	1.03, 1.06, 1.08
Transfer-in patients	186	0.02%	0.0, 0.02, 0.03
Without at least 30 days post-discharge or claim end date information	38,227	3.72%	3.69, 3.73, 3.77
Patients who leave hospital against medical advice (AMA)	209	0.02%	0.0, 0.02, 0.03
Patients with more than two THA/TKA procedure codes	1	0.00%	
Additional admission for THA/TKA within 30 days of prior index admission	1,440	0.14%	0.10, 0.13, 0.16

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis.* Note: ***If patient preference is an exclusion****, the measure must be specified*

so that the effect on the performance score is transparent, e.g., scores with and without exclusion)

Example:
The overall frequency of the exclusions is low, with incomplete data for the prior 12 months resulting in the most exclusions (11.25%). The distribution of exclusions across hospitals is very narrow indicating that the occurrence is random and likely would not bias performance results. However, we think the exclusions should be retained for the following reasons.

Who die during the index admission
Rationale: Patients who die during the initial hospitalization are not eligible for readmission.

Without at least 12 months pre-index admission enrollment in Medicare FFS
Rationale: Appropriate risk adjustment requires uniform data availability of pre-operative comorbidity. The value of proper severity adjustment outweighs the burden of increased data collection and analysis.

Without at least 30 days post-discharge enrollment in Medicare FFS
Rationale: The 30-day readmission outcome cannot be assessed for the standardized time period.

Who were admitted for the index procedure and subsequently transferred to another acute care facility
Rationale: Attribution of readmission to the index hospital would not be possible in these cases, since the index hospital performed the procedure but another hospital discharged the patient to the non-acute care setting.

Who were transferred in to the index hospital
Rationale: If the patient is transferred from another acute care facility to the hospital where the index procedure occurs, it is likely that the procedure is not elective or that the admission is associated with an acute condition.

Who leave the hospital against medical advice (AMA)
Rationale: Hospitals and physicians do not have the opportunity to provide the highest quality care for these patients.

With more than two THA/TKA procedures codes during the index hospitalization
Rationale: Although clinically possible, it is highly unlikely that patients would receive more than two elective THA/TKA procedures in one hospitalization, and this may reflect a coding error.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

Key Points
This section should justify:
- the factors selected for the statistical risk model or for stratification; and
- the adequacy of the risk model or stratification in controlling for patient factors present at the start of care

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**
☒ **Statistical risk model with 37 risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

Key Point
- There may be some outcome performance measures that do not need to be risk adjusted; however, that will need to be demonstrated.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

Key Points
- Indicate the clinical criteria for identifying candidate risk factors (e.g., identified in a literature review, identified by clinical experts).
- Describe the statistical analyses and criteria for selection of final risk factors (e.g., correlation with outcome, significance level).
- Analyses to select risk factors are relevant for both statistical risk adjustment and stratification approaches.

Example (for statistical risk adjustment):
Candidate variables were identified through a literature review for factors that influenced improvement in ambulation and then discussed with a clinical expert panel. Candidate variables were analyzed in a logistic regression model. For dichotomous factors, the odds ratio indicates the strength of the association with the outcome while controlling for all other risk factors. The larger or the smaller the odds ratio (>1.0 or <1.0) the greater the influence on the outcome. Confidence intervals for odds ratios that do not include 1.00 indicate a significant association with the outcome. Variables with at least one category with a confidence interval that does not include 1.00 were included in the final risk-adjustment model.

2b4.4. What were the statistical results of the analyses used to select risk factors?

Key Points
- Report data that supports final risk factors. Data on candidate factors not selected can be provided in an appendix.
- Analysis used to select risk factors are relevant for both statistical risk adjustment and stratification approaches.
- The following example is for a dichotomous outcome variable. If the outcome variable is continuous, other statistics would be appropriate (e.g., correlations with outcome and standard regression

coefficients).

Example (for statistical risk adjustment):

Table 1: Improvement in Ambulation / Locomotion

Risk Factor Measured at SOC/ROC[1],[2]	Coefficient[3]	Odds ratio[3]	OR (95% CI)
Risk for Hospitalization: History of falls	-0.169	0.844	(0.820-0.869)
Risk for Hospitalization: Two or more hospitalizations past year	-0.088	0.916	(0.888-0.945)
Supervision and Safety Assistance: None needed	-0.107	0.899	(0.851-0.950)
Supervision and Safety Assistance: Caregiver currently provides	-0.099	0.905	(0.860-0.953)
Status of Surgical Wound: Fully granulating	0.288	1.334	(1.254-1.418)
Status of Surgical Wound: Early/partial granulation	0.578	1.783	(1.709-1.860)
Status of Surgical Wound: Not healing	0.657	1.929	(1.802-2.066)
Number of therapy visits: 1-2	0.021 c	1.021	(0.969 -1.075)
Number of therapy visits: 3-4	-0.052 c	0.949	(0.886 -1.018)
Number of therapy visits: 5-6	0.047 a	1.049	(1.001 -1.099)
Number of therapy visits: 7-8	0.106	1.112	(1.062-1.164)
Number of therapy visits: 9-10	0.125	1.133	(1.082-1.187)
Number of therapy visits: 11-12	0.169	1.184	(1.127-1.244)
Number of therapy visits: 13-14	0.167	1.182	(1.121-1.246)
Number of therapy visits: 15-16	0.130	1.139	(1.072-1.210)
Number of therapy visits: 17-18	0.176	1.193	(1.108-1.283)
Number of therapy visits: 19-20	0.140	1.150	(1.058-1.251)
Episode Timing: Early	0.261	1.298	(1.233-1.367)
Toilet Hygiene Assistance: Needs supplies laid out	0.089	1.093	(1.053-1.134)
Toilet Hygiene Assistance: Needs assistance	-0.037 c	0.964	(0.917-1.013)
Toilet Hygiene Assistance: Entirely dependent	-0.202	0.817	(0.739-0.903)
Vision: Partially impaired	-0.071	0.931	(0.901-0.962)
Vision: Severely impaired	-0.168	0.845	(0.763-0.936)
Constant	-0.359	0.698	

1 SOC = Start of Care, ROC = Resumption of Care after inpatient stay. Risk factors values are based on SOC/ROC assessment values for the episode of care.

2 Most risk factors take on the values 0 and 1; 1 denotes the presence of the attribute and 0 denotes its absence. In virtually all cases, each response option for an OASIS item is a risk factor. In a small number of cases (e.g., number of pressure ulcers at Stage III), the actual number provided in the assessment is used.

3 Because all response option values for an OASIS item are included as risk factors in a model even if only one option value is statistically significant, the following superscripts are used to denote the statistical significance for each risk factor coefficient:

Blank = significant at probability<.01
a = significant at 0.01<probability<.05
b = significant at 0.05<probability<.10
c = not significant, probability>.10

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what*

statistical analysis was used)

Key Points
• Indicate whether the accepted practice of randomly dividing the data into development and validation samples to compare discrimination and calibration was followed. • Describe the type of analysis used (e.g., logistic regression, linear regression) • Demonstration of the adequacy of the approach is relevant for both statistical risk adjustment and stratification. • Identify if any additional analyses were performed (sensitivity analyses for problem or missing data), provide the information in question 2b4.11

Example (for statistical risk adjustment):
The analytic file for the sample described above was randomly divided into development and validation samples, each with 250,000 cases. The outcome variable is dichotomous so logistic regression analysis was conducted to make the final selection of risk factors and analyze model performance.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
if stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

Key Point
• The following example is for a dichotomous outcome variable. If the outcome variable is continuous, the R-squared statistic would be appropriate.

Example (for statistical risk adjustment):

Statistic	**Development Sample**	**Validation Sample**
N	500,000	1,000,000
Area Under Receiver Operator Curve (C-statistic)	0.651	0.648

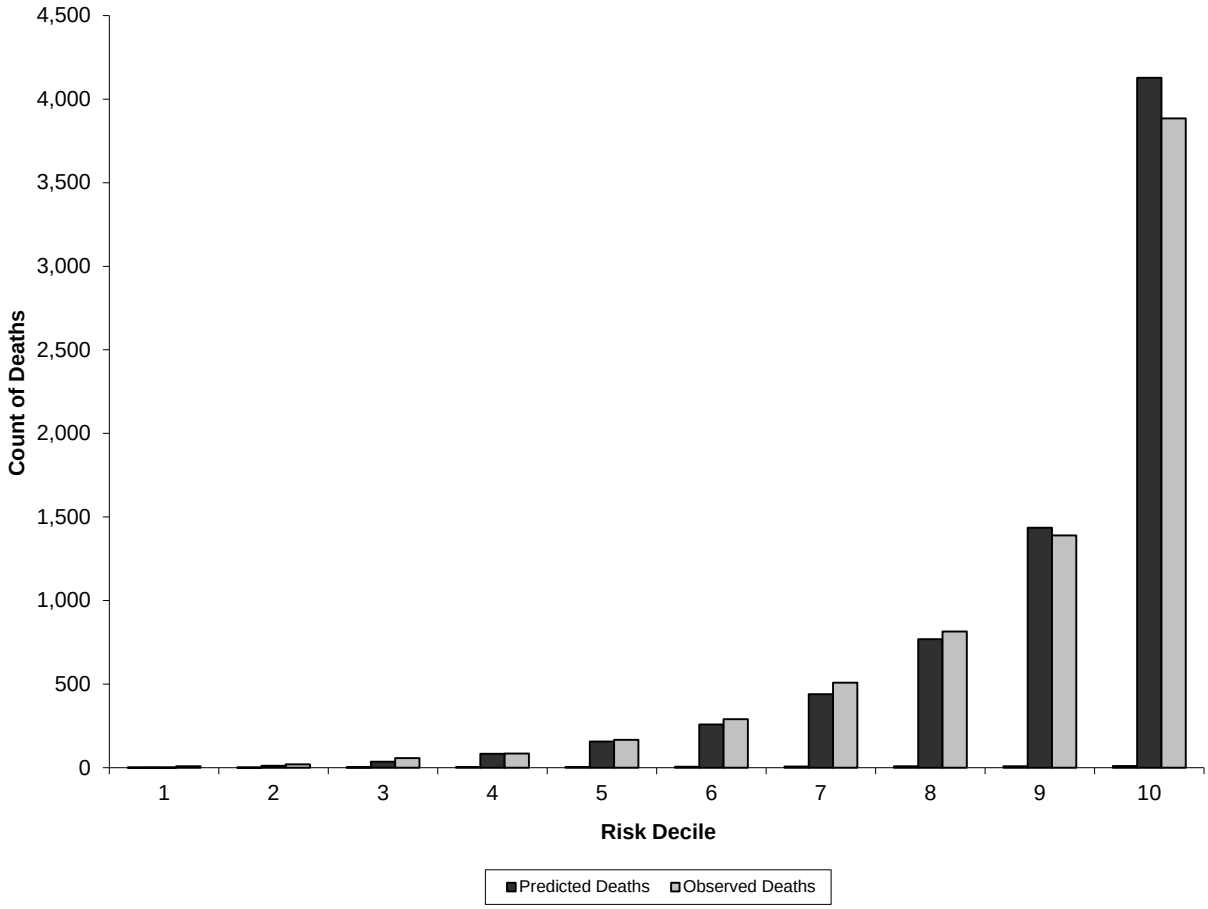
2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

Key Point
• Calibration statistics (e.g., Hosmer-Lemeshow in logistic regression) typically are not very informative and risk decile plots or calibration curves should be provided.

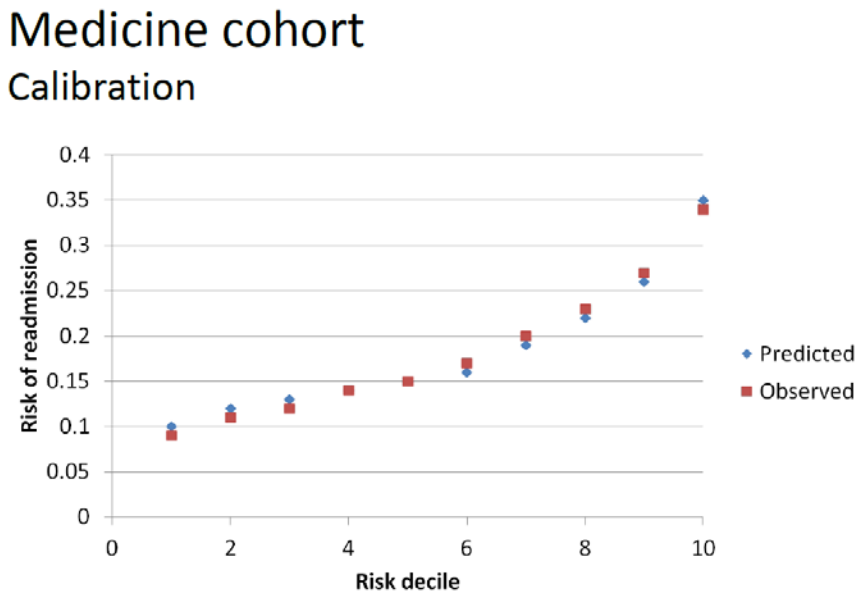
Example (for statistical risk adjustment):
Hosmer-Lemeshow statistic

Chi-Square	Degrees of Freedom	Probability
9.8690	8	0.2743

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
Example 1 (for statistical risk adjustment):



Example 2 (for statistical risk adjustment):



2b4.9. Results of Risk Stratification Analysis:

Key Points

- Provide data that demonstrate the risk categories represent different levels of risk and support the stratification of performance results to facilitate fair comparisons.

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

Key Points

- Explain the meaning of the discrimination statistics and calibration curve or risk decile plot, or analysis of stratification.
- Explain how the results compare to norms for the analyses.
- Provide a rationale as to why the results indicate the model is adequate.

Example (for statistical risk adjustment):

The C statistic is a measure of the extent to which a statistical model is able to discriminate between a patient with and without an outcome. The c-statistic ranges from 0.5 to 1.0. A c-statistic of 0.50 indicates the model is no better than random prediction, implying that the patient risk factors do not predict variation in the outcome; conversely, a c-statistic of 1.0 indicates perfect prediction, implying patients' outcomes can be predicted completely by their risk factors, and physicians and hospitals play little role in patients' outcomes. In the context of healthcare performance assessment, the purpose of the risk model is to reduce bias due to case mix characteristics present at the start of care (i.e. to risk adjust), not to totally explain variation in outcomes, which would require also including variables about quality of care. Variables related to quality of care are purposely not included in risk models for performance measures used to assess quality. This result (0.648) is comparable to c-statistics for risk models for this outcome in other settings (0.65 to 0.75) (references). Although a higher c-statistic is desirable, it is possible for a risk model to exhibit low discrimination (using only patient factors) and still perform well at reducing bias due to differences in case mix. The H-L statistic was not significant indicating the model does fit the data. The risk decile plots indicate that the risk model performs well across all deciles of risk with predicted deaths similar to observed deaths, except for the 10% with the highest risk.

(Note: Sometimes with very large samples the H-L may be significant but not as useful as examining the decile plot for concluding a model does not fit the data.)

***2b4.11. Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

Key Points

- This question may overlap with data provided in section 1b on performance gap but ideally goes beyond providing a simple distribution of performance measure scores.
- If this is analysis is based on different data than the rest of testing (possibly more current), then that

should be explained in question 1.7 above.

- Ideally, analyses to distinguish better from poorer quality should be provided (e.g., statistical/clinically meaningful differences from average performance).
- At a minimum, frequency distribution of performance measure scores for measured entities should be provided.

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Example:
To examine differences in performance, we used the development dataset. The final results (unadjusted and risk-standardized hospital-level 30 day readmission rates) were calculated across 4,742 hospitals. We excluded hospitals with fewer than 25 cases total across the five condition cohorts since estimates for hospitals with fewer cases are less reliable. A confidence interval was computed for each provider's score and if it did not contain the average, the provider is identified as better as or worse than average.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Example:
Distribution of standardized risk ratios: Hierarchical logistic regression model results for the 2007-2008 development sample

SRR	Mean	SD	Minimum	10th percentile	Lower quartile	Median	Upper quartile	90th percentile	Maximum
Combined	1.00	0.07	0.75	0.92	0.96	0.99	1.04	1.09	1.36
Medicine	1.00	0.09	0.75	0.90	0.94	0.99	1.05	1.12	1.51
Surgery/ Gynecology	1.00	0.08	0.73	0.91	0.96	1.00	1.04	1.10	1.52
Cardiorespiratory	1.00	0.08	0.75	0.90	0.95	1.00	1.05	1.11	1.44
Cardiovascular	1.00	0.05	0.78	0.94	0.97	1.00	1.03	1.06	1.32
Neurology	1.00	0.06	0.76	0.94	0.97	1.00	1.03	1.07	1.34

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

Example:
The measure was able to detect those with better and worse than average performance.
The unadjusted readmission rates ranged from 0 percent to 36.56 percent, with a median of 16.26 percent. Fifty percent of hospitals fell within the interquartile range of 14.01-18.64. The mean ± standard deviation (SD) hospital unadjusted readmission rate was 16.35 percent ± 4.20. The risk-

standardized rates (RSRRs) had a much narrower range, from 12.58 percent to 22.76 percent, with a median of 16.58 percent. Fifty percent of hospitals fell within the interquartile range of 15.96-17.30. The mean ± SD hospital RSRR was 16.69 percent ± 1.15 percent. This distribution of performance is similar to other endorsed outcome measures, specifically condition-specific readmissions and all-cause readmissions. Of the 4,742 hospitals, 15 percent (711hospitals) were statistically significantly better than average and 12 percent (569 hospitals) were worse than average.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note: *This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

Key Points
- Performance measures that are specified with more than one: data source, risk model, or use of assessment instrument need to demonstrate that performance measure scores derived with the various approaches result in comparable results so that the differences do not bias results.

2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

Key Points
- Required for eMeasures, composites, and PRO-PMs, but could have relevance for other measures.
- Applies to missing data and nonresponse for patient-reported data.
- At a minimum, frequency and distribution of missing/nonresponse data must be provided.
- Ideally analyses demonstrate that missing data does not bias results or handling of missing data minimizes bias in the performance measure scores; if not empirical analysis follow directions for what should be discussed.

- Will add example when available.

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)