NATIONAL QUALITY FORUM

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Click to go to the link. ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

Brief Measure Information

NQF #: 3461

Corresponding Measures:

De.2. Measure Title: Functional Status Change for Patients with Neck Impairments

Co.1.1. Measure Steward: Focus on Therapeutic Outcomes

De.3. Brief Description of Measure: This is a patient-reported outcome performance measure (PRO-PM) consisting of a patient-reported outcome measure (PROM) of risk-adjusted change in functional status (FS) for patients aged 14 years and older with neck impairments. The change in FS is assessed using the Neck FS PROM. The measure is adjusted to patient characteristics known to be associated with FS outcomes (risk adjusted) and used as a performance measure (PM) at the patient, individual clinician, and clinic levels to assess quality.

The Neck FS PROM is an item-response theory-based computer adaptive test (CAT) for patients with impairments related to neck problems. Specific ICD-10-CM codes are described in the denominator section.

The Neck PRO-PM is publically available in the CAT version on the FOTO website at no charge. The Neck FS PROM is also available at no charge for public use as a 10-item short form (static/paper-pencil). CAT administration is preferred as it reduces patient response burden by administrating the minimum number of items needed to achieve the targeted measurement accuracy. Scores are reported on a 0 to 100 scale with higher scores indicating better functional status. The Neck FS PROM maps to the Mobility and Self-care constructs within the Activities and Participation domain of the International Classification of Functioning, Disability and Health.

1b.1. Developer Rationale: Neck pain is recognized as a global healthcare burden.1,2 Prevalence estimates from epidemiologic studies on neck pain, defined as pain in the neck with or without pain referred into one or both upper limbs that lasts for at least 1 day, have a mean 1-year prevalence range from 23% 4 to 37% 3 and mean lifetime prevalence of 49%.3 The use of patient-reported outcome measures (PROMs) for assessing functional status in patients with neck pain is an essential step in addressing this burden, provided 1) scores can be interpreted in clinically useful ways to inform patient-centered clinical decision making, 2) performance across providers can be reliably and validly assessed, and 3) a performance gap between providers can be identified setting an opportunity for improvement over time. For example, results (Measure Testing form FIGURE 2b4ii-iii) demonstrate performance gaps that may form a basis for improvements in quality envisioned for this measure; we expect providers ranked as having low quality (mean residuals below zero) to improve over time to average or high quality (mean residuals zero or above) Evidence supporting that NQF measure 3461 can successfully address these 3 elements is outlined above and is described in the Measure Testing form in the Scientific Acceptability section.

PROMs are increasingly advocated as necessary components of an overall strategy to improve healthcare5,6 and are advocated for use in clinical decision making in clinical practice guidelines.7-11 However, prior to the development of the Neck Functional Status (FS) PROM, the literature lacked a neck-specific functional status PROM based on modern scientific measurement methods like item response theory (IRT) and computer adaptive testing (CAT).12-16 Further, when modern measurement theory approaches were applied to previously existing measures, psychometric limitations were discovered including floor and ceiling effects, invalid assumptions of interval scaling, and multi-dimensionality. The Neck FS PROM is free of such limitations.17-21 Furthermore, using the Neck FS PROM reduces patient burden by minimizing the number of functional questions the patient must respond to in order to obtain a precise estimate of the patient's functional ability level.22-23

When combined with robust risk adjustment ,24 the IRT-based Neck FS PROM forms the basis for a valuable patient reported outcome performance measure (PRO-PM). Placing risk-adjusted Neck FS PROM data directly into the hands of the provider embodies the definition of patient-centered healthcare and is consistent with National Quality Forum's vision to achieve performance improvement and accountability through patient-reported outcomes.25 This approach improves quality of care by promoting improved communication between provider and patient and enhancing the provider's understanding of the patient's perception of functional status. The Neck FS PROM and PRO-PM results can even be shared with the patient to further promote patient engagement; as one example, the FOTO Outcomes Measurement system provides a visually pleasing, patient-focused real-time report of the patient's (risk-adjusted) PRO-PM results.

1. Hoy D, March L, Woolf A, et al. The global burden of neck pain: estimates from the global burden of disease 2010 study. Ann Rheum Dis. 2014;73:1309-1315.

2. Hurwitz EL, Randhawa K, Yu H, Cote P, Haldeman S. The Global Spine Care Initiative: a summary of the global burden of low back and neck pain studies. Eur Spine J. 2018;27:796-801.

3. Fejer R, Kyvik KO, Hartvigsen J. The prevalence of neck pain in the world population: a systematic critical review of the literature. Eur Spine J. 2006;15:834-848.

4. Hoy DG, Protani M, De R, Buchbinder R. The epidemiology of neck pain. Best Pract Res Clin Rheumatol. 2010;24:783-792.

5. Black N. Patient reported outcome measures could help transform healthcare. BMJ. 2013;346:f167.

6. Griggs CL, Schneider JC, Kazis LE, Ryan CM. Patient-reported outcome measures: a stethoscope for the patient history. Ann Surg. 2017;265:1066-1069.

7. Blanpied PT, Gross AR, Elliott JM, et al. Neck Pain: Revision 2017 Clinical Practice Guidelines Linked to the International Classification of Functioning, Disability and Health from the Orthopaedic Section of the American Physical Therapy Association. J Orthop Sports Phys Ther. 2017;47(7):A1-A83. doi:10.2519/jospt.2017.0302.

8. Childs JD, Cleland JA, Elliott JM, Teyhen DS et al. Neck Pain: Clinical Practice Guidelines Linked to the International Classification of Functioning, Disability and Health from the Orthopaedic Section of the American Physical Therapy Association. J Orthop Sports Phys Ther. 2008;38:A1-A34.

9. Baisden J, Easa J, Fernand R, Lamer T, et al. North American Spine Society Evidence-Based Clinical Guidelines for Multidisciplinary Spine Care Diagnosis and Treatment of Cervical Radiculopathy from Degenerative Disorders. North American Spine Society 2010.

10. Bono CM, Ghiselli G, Gilbert TJ, Kreiner S et al. An evidence-based clinical guideline for the diagnosis and treatment of cervical radiculopathy from degenerative disorders. Spine J 2011;11:64-72.

11. Bier JD, Scholten-Peeters WGM, Staal JB, Pool J et al. Clinical Practice Guideline for Physical Therapy Assessment and Treatment in Patients With Nonspecific Neck Pain. Phys Ther 2018:98:162-171.

12. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther. 1991;14:409-415.

13. Jordan A, Manniche C, Mosdal C, Hindsberger C. The Copenhagen Neck Functional Disability Scale: a study of reliability and validity. J Ma¬nipulative Physiol Ther. 1998;21:520-527.

14. Wheeler AH, Goolkasian P, Baird AC, Darden BV, 2nd. Development of the Neck Pain and Disability Scale. Item analysis, face, and criterion-related validity. Spine (Phila Pa 1976). 1999;24:1290-1294.

15. BenDebba M, Heller J, Ducker TB, Eisinger JM. Cervical Spine Outcomes Questionnaire: its de¬velopment and psychometric properties. Spine (Phila Pa 1976). 2002;27:2116-2123; discussion 2124.

16. Leak AM, Cooper J, Dyer S, Williams KA, Turner-Stokes L, Frank AO. The Northwick Park Neck Pain Questionnaire, devised to measure neck pain and disability. Br J Rheumatol. 1994;33:469-474.

17. Hung M, Cheng C, Hon SD, Franklin JD, Lawrence BD, Neese A, Grover CB, Brodke DS. Challenging the norm: further psychometric investigation of the Neck Disability Index. The Spine Journal. 2015;15(11):2440 – 2445.

18. Van der Velde G, Beaton D, Hogg-Johnston S, Hurwitz E, Tennant A. Rash analysis provides new insights into the measurement properties of the Neck Disability Index. Arthritis & Rheumatism. 2009;61(4):544-551.

19. Ailliet L, Knol DL, Rubinstein SM, de Vet HC, van Tulder MW, Terwee CB. Definition of the construct to be measured is a prerequisite for the assessment of validity. The Neck Disability Index as an example. J Clin Epidemiol. 2013; 66(7): 775-782

20. Wang YC, Cook KF, Deutscher D, Werneke MW, Hayes D, Mioduski JE. The Development and Psychometric Properties of the Patient Self-Report Neck Functional Status Questionnaire (NFSQ). J Orthop Sports Phys Ther. 2015;45:683-692.

21. Deutscher D, Cook KF, Kallen MA, et al. Clinical Interpretation of the Neck Functional Status Computer Adaptive Test. J Orthop Sports Phys Ther. 2019;Accepted:

22. Swaminathan H, Hambleton R. Fundamentals of item response theory. Newbury Park [CA]: Sage Publications; 1991. In: Stephanie Nikolaus, JMIR Human Factors. 2014;1(1):e4.

23. Sands WA, Waters BK, McBride JR, editors. Computerized adaptive testing: from inquiry to operation. Washington, DC: American Psychological Association; 1997.

24. Deutscher D, Werneke MW, Hayes D, et al. Impact of Risk Adjustment on Provider Ranking for Patients with Low Back Pain Receiving Physical Therapy. J Orthop Sports Phys Ther. 2018;48(8):637-648.

25. National Quality Forum. Patient Reported Outcomes (PROs) in Performance Measurement. January 10, 2013. https://www.qualityforum.org/publications/2012/12/patient-reported_outcomes_in_performance_measurement.aspx Accessed March 21, 2019.

S.4. Numerator Statement: The numerator is based on residual scores (actual change scores - predicted change after risk adjustment) of patients receiving care for neck impairments and who: a) completed the Neck PRO-PM at admission and at the end of the episode of care; and b) were discharged from care.

S.6. Denominator Statement: All patients 14 years and older with a neck impairment who have initiated an episode of care and completed the neck functional status PROM at admission and discharge.

S.8. Denominator Exclusions: Patients who are not being treated for a neck impairment. Patients who are less than 14 years of age.

De.1. Measure Type: Outcome: PRO-PM

S.17. Data Source: Instrument-Based Data

S.20. Level of Analysis: Clinician : Group/Practice, Clinician : Individual

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? NA

Preliminary Analysis: New Measure

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcome measure include providing empirical data that demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service; if these data not available, data demonstrating wide variation in performance, assuming the data are from a robust number of providers and results are not subject to systematic bias. For measures derived from patient report, evidence also should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.

Evidence Summary

- Brief background: This is a patient-reported outcome-based performance measure (PRO-PM) that uses data from the Neck FS PROM to assess change in functional status (FS) for patients ages 14+ who have neck impairments.
- In the logic model provided by the developers, they link information collected via the Neck FS PROM to clinician decision-making and communication needed to establish goals of care and a plan of care. This model does not explicitly identify clinical interventions that can lead to improvements in functional status for those with neck impairments.
- The developers provided data indicating that administering interim functional status assessments early in the episode of care is associated with statistically significant improvement in functional status. Developers suggest that administration of interim assessments allow clinicians to continue/modify treatment interventions based on patient report of improvement in function.
- The developer described how they determined that patients with neck pain find the physical activity question on the Neck FS PROM to be meaningful vis-à-vis their neck pain. It appears the developers did not explicitly discuss with patients the meaningfulness of the measured outcome itself (i.e., change in functional status).
    - Results from their analysis suggest that most sampled patients found at least some of the questions to be meaningful. Developers note that older patients found the questions more meaningful than did younger patients, but no differences by sex, treatment status, or current neck pain status.

Questions for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*
- *Given the evidence described by the developer, do you believe patients with neck pain values the outcome of change in functional status and finds it meaningful?*

Guidance from the Evidence Algorithm

Measure assesses performance on a patient-reported outcome (Box 1) → Empirical data suggest that early, interim assessments of functional status are associated with greater improvement in functional status (Box 2) → PASS

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

Data provided by the developer from for patients who completed the Neck FS PROM at intake and discharge between 2016-2017 indicate the following: (NOTE: a higher value is better)

- Clinician groups (n=1,378 clinics; 123,194 patients):
    - Average risk-adjusted change: -0.5
    - Range: -14.0 to 22.1
- Individual clinicians (n=4,537 clinicians; 112,178 patients)
    - Average risk-adjusted change: -0.4
    - Range: -14.1 to 22.1

Disparities

- The developer provided beta coefficients (for a model predicting patient-level functional status at discharge) for age group, sex, and insurance status (after controlling for other variables included in the measure's risk-adjustment model).
    - Results indicated that younger patients have better outcomes compared older patients; that women have worse outcomes than men; and that outcomes vary substantially by insurance type.

Questions for the Committee:

- *Do you need additional explanation from the developer to help interpret the values presented to demonstrate opportunity for improvement?*
- *Is there a gap in care (or opportunity for improvement) that warrants a national performance measure?*
- *Are you aware of evidence that other disparities exist in this area of healthcare, beyond those indicated by the developer?*

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:
Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence: *For all measures (structure, process, outcome, patient-reported structure/process), empirical data are required. How does the evidence relate to the specific structure, process, or outcome being measured? Does it apply directly or is it tangential? How does the structure, process, or outcome relate to desired outcomes? For maintenance measures –are you aware of any new studies/information that changes the evidence base for this measure that has not been cited in the submission? For measures derived from a patient report: Measures derived from a patient report must demonstrate that the target population values the measured outcome, process, or structure.*

- Evidence supports the value of the measure - majority of patients indicated the questions were meaningful to them (92% said some or most); treatment plans can be created and adjusted based on the PROM and does have some data to show it is impact on the patients outcome.
- It appears as if the evidence base for this measure is from a single study, albeit a well-constructed one. I cannot tell if this study was published in a peer-reviewed journal or not. The committee is asked to consider if there is at least one thing that the clinician can do to change the measure results and it seems logical that one could, but specific clinical methodologies that have been tested are not provided (at least I

could not find them). It does seem reasonable that a patient would value the outcome of a positive change in this measure.

1b. Performance Gap: *Was current performance data on the measure provided? How does it demonstrate a gap in care (variability or overall less than optimal performance) to warrant a national performance measure? Disparities: Was data on the measure by population subgroups provided? How does it demonstrate disparities in the care?*

- Gap in case is demonstrated as there was not a similar meaure to help determine next steps in treatment prior to this one. As this was a new measure starting in 2016, preliminary performance data does indicate improvements with use.
- The data are cited as being preliminary, but there does appear to be a performance gap.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications **and** Testing

2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability; Missing Data

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

2a2. Reliability testing demonstrates if 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 enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Composite measures only:

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

Complex measure evaluated by Scientific Methods Panel? ☒ **Yes** ☐ **No**

Evaluators: NQF Scientific Methods Panel

Methods Panel Review (Combined)

Methods Panel Evaluation Summary:

This measure was reviewed by the Scientific Methods Panel and discussed on their subgroup call. A summary of the measure and the Panel's discussion is provided below.

Scientific Methods Panel Votes: Measure passes reliability; consensus not reached for validity

- Reliability: H-0, M-5, L-0, I-0
- Validity: H-0, M-3, L-1, I-1

Background

This measure did not pass on reliability and validity in the previous review during the Fall 2018 Scientific Methods Panel (SMP) evaluation. Developers resubmitted the measure for the Spring 2019 cycle, after receiving technical assistance from NQF staff. In their Fall 2018 evaluation, SMP subgroup members expressed several concerns regarding both reliability and validity, and requested the measure developer perform several actions:

- Add specificity on data sources and definitions for key data elements
- Clarify the definition of neck impairments that are included in the measure. "Not limited to" as a description of included conditions is not specific and lacks the necessary clarity for an implementer to run the measure.
- Clarify the numerator statement and descriptor regarding whether this is a change score
- Explain how proxy responses are used
- Clarify the episode definition and how a discharge is determined or captured
- Clarify whether measure testing for validity was done at the score level and present results of that testing
- Clarify how incomplete surveys are handled
- Address concerns related to limited testing of the risk-adjusted change score and lack of differentiation of the clinic- versus clinician-level validity

The developers also reviewed the formula used for reliability testing and compared it to the description of that formula to ensure they are aligned. They corrected the testing method used, ensuring that it is based on mixed effects model. They also clarified that the analysis is based on raw change scores or the residuals (residual = actual change score minus the risk-adjusted predicted change score).

As part of its evaluation of the measure during the Spring 2019 cycle, the SMP felt that the current submission addressed their concerns from the previous cycle. While the SMP also had concerns regarding specifications, reliability, and validity during the current review cycle, these were resolved in the course of review. However, for score level validity testing, the SMP also wanted to see change scores versus some external criterion. (NOTE that the testing provided for the Spring 2019 submission used a comparator measure that included elements of the measure itself, and consequentially, the SMP felt that the demonstration of construct validity lacked compelling evidence of an association between the measure scores and some independent measure or concept of quality of care. This is why the SMP did not achieve consensus on its vote). The SMP provided specific instructions to the measure developer on how this issue (of an external criterion measure) could be addressed, and requested the developer provide that testing to the full Standing Committee. This additional testing has been included as an appendix to the measure developer's submission. It will be fully incorporated into the developer's submission materials before the end of the evaluation cycle.

Reliability

- Testing included score-level (via signal to noise analysis) and data element testing
- Data element level testing (i.e., demonstrating reliability of the instrument) was assessed via internal consistency analysis (Cronbach's alpha) and Item Response Theory (IRT) person reliability analysis.
  - Cronbach's alpha=0.98
  - IRT-based person reliability=0.96.
- Score-level testing, clinics: Average reliability= 0.79 (for clinics meeting the FOTO unique threshold of number of patients per clinic for quality reporting)
- Score-level testing, clinicians:
  - Average reliability=0.64 for clinicians with 10 or more patients per calendar year
  - Average reliability=0.76 for clinicians with 20 or more patients per calendar year
- In addition, developers also assessed reliability of individual scores via the Standard Error of Measurement and analysis of minimal detectable improvement.

Validity

- Because this is an instrument-based measure, both data element and score-level testing are required.
- Developers performed several types of data element validity testing, including content validity, structural validity, person-item match, differential item functioning of the scale, known groups construct validity, sensitivity to change, responsiveness, and functional staging. Results can be found here.
- The developers, per instruction from the SMP, provided additional score-level testing results using two external comparators (a global rating of change score and the neck disability index).
  - Developers examined the correlation between performance level scores of the measure compared to two external measures: the global rating of change (GROC) assessed at discharge, and the neck disability index (NDI) as change from admission to discharge.
  - Developers reported Pearson correlation coefficients of mean risk-adjusted residual scores and GROC and NDI mean scores.
  - Absolute correlations for the two measures and provider levels ranged from 0.64 to 0.73 and were statistically significant (P<0.001).

Questions for the Committee regarding reliability:

- *Do you have any questions regarding the specifications of the measure?*
- *Do you have any concerns that the measure can be consistently implemented?*
- *Do you have any concerns regarding the reliability of the Neck FS PROM or the performance measure derived from this instrument?*
- *The Scientific Methods Panel is satisfied with the reliability testing for the measure. Does the Committee think there is a need to vote on reliability?*

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the Neck FS PROM?*
- *Do you have any questions or concerns regarding the newly-added score-level validity testing results?*
- *Are you satisfied that the newly-added score-level validity testing demonstrates the validity of this measure?*
- *Do you have any concerns regarding potential threats to the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Preliminary rating for validity: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Combined Methods Panel Scientific Acceptability Evaluation

Measure Number: 3461

Measure Title: Functional Status Change for Patients with Neck Impairments

Type of measure:

☐ **Process** ☐ **Process: Appropriate Use** ☐ **Structure** ☐ **Efficiency** ☐ **Cost/Resource Use**

☐☐ **Outcome** ☐☒ **Outcome: PRO-PM** ☐ **Outcome: Intermediate Clinical Outcome** ☐ **Composite**

Data Source:

☐ **Claims** ☐ **Electronic Health Data** ☐ **Electronic Health Records** ☐ **Management Data**

☐ **Assessment Data** ☐ **Paper Medical Records** ☐☒ **Instrument-Based Data** ☐ **Registry Data**

☐ **Enrollment Data**

Level of Analysis:

☒ **Clinician: Group/Practice** ☒ **Clinician: Individual** ☐ **Facility** ☐ **Health Plan**

☐ **Population: Community, County or City** ☐ **Population: Regional and State**

☐ **Integrated Delivery System** ☐☐ **Other:** individual patient level

Measure is:

☒ **New** ☐ **Previously endorsed (**NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.)

MP#2: The numerator is based on residual scores (actual change scores - predicted change after risk adjustment) of patients receiving care for neck impairments and who: a) completed the Neck PRO-PM at admission and at the end of the episode of care; and b) were discharged from care.

The denominator is all patients 14 years and older with a neck impairment who have initiated an episode of care and completed the neck functional status PROM at admission and discharge.

RELIABILITY: SPECIFICATIONS

1. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?** ☐☒ **Yes** ☒☐ **No**

   Submission document: "MIF_xxxx" document, items S.1-S.22

   NOTE*: NQF staff will conduct a separate, more technical, check of eCQM specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

2. **Briefly summarize any concerns about the measure specifications.**

   MP#1: A few concerns:

   - The specification needs more clarity regarding age (age at admission? Age at discharge?)
   - What is the timeframe for the denominator? Calendar year?
     - If the timeframe is the most recent calendar year, does the admission and discharge need to occur in the same year?
   - Reliability testing results indicate that the measure is insufficient reliable with fewer than 20 patients in the denominator per reporting entity. The specifications should provide that information.

   MP#4: The link cited in S.1 has much more detail with regard to specifications than was submitted on the NQF MIF. For example, episode of care is defined as either rehabilitation therapy, medical or chiropratctic episode. No mention of modes of administration or in what languages. No sampling procedure reported. Missing data is not addressed.

RELIABILITY: TESTING

Submission document: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

3. **Reliability testing level** ☒ **Measure score** ☒ **Data element** ☐ **Neither**

4. **Reliability testing was conducted with the data source and level of analysis indicated for this measure**

   ☒ **Yes** ☐ **No** Clarify whether at group level or clinic level. At times its used interchangeably.

5. If score-level and/or data element reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical VALIDITY testing** of patient-level data conducted?

   ☐ **Yes** ☐ **No** N/A

6. **Assess the method(s) used for reliability testing**

   MP#2:

a. **Patient-level: Internal consistency (alpha) for 10-item SF and IRT-based person reliability for CAT score. These are appropriate with Cronbach's alpha of 0.98 and an IRT-based person reliability of 0.96.**
b. **Clinician: the average reliability of clinics meeting the FOTO unique threshold of number of patients per clinic for quality reporting was 0.79. At the clinician level, average reliability for clinicians with 10 or more, or 20 or more patients per calendar year was 0.64 and 0.76, respectively.**

Submission document: Testing attachment, section 2a2.2

MP#4: The methods use are appropriate. Testing was conducted at the data source and level of analysis indicated. See #4 above.

MP#3: Internal consistency of FS PROM was assessed through both classical analysis and IRT. Reliability of individual scores were assessed using SEM.

Performance score reliabilities were assessed using HLM. Because these measures are specified for both individual clinician and group practice, separate analyses should have been conducted, the specifications of HLM models are important and relevant but seem to be missing.

MP#1: The developer evaluated internal consistency of the 28 items using Cronbach's alpha, the reliability of baseline and change functional status scores using standard error of measurement analyses, and conducted a STN analysis of change scores at the clinician and practice levels. These tests are appropriate, particularly the STN analysis. However, the materials submitted do not indicate that the score testing, particularly the STN reliability testing, was completed using risk adjusted scores at the provider and practice level.

MP#5: The methods used were acceptable and technically appropriate, at both the data element and measure score levels.

7. **Assess the results of reliability testing**

MP#4: Internal consistency (Cronbach's alpha) conducted at the patient level—suggesting excellent internal consistency. At the clinic level, scores are sufficiently reliable if the threshold of number of patients is used as provided.

MP#2:

a. **Reliability of the patient-level data elament (score on FS measure) was well described and justified. The CAT administration and 10-item short form appear to produce reliable scores across most of the measurement continuum. Removal of some items served to increase reliability but may also have changed the nature of what is being measured.**

Submission document: Testing attachment, section 2a2.3

MP#3: Cronbach's alpha 0.98, IRT based person reliability was 0.96, both were very high, indicating excellent internal consistency.

Performance score reliabilities in general were moderate and acceptable. For individual clinician with small number of patients, performance score reliability could be an issue.

MP#1: Results of testing indicate that the 28 item bank achieves a Cronbach's alpha of .98, indicating that the item bank questions have a high level of internal consistency. At the individual provider level, the STN analysis showed that among providers with at least 20 patients in the denominator, the measure is generally reliable, with a mean reliability coefficient of 0.76, with nearly 3/4 of providers with a reliability coefficient of 0.70 or higher. However, fewer than one third of all providers in the testing database had 20 or more patients in the denominator. For that reason, this measure should probably be reported at the practice level to promote more reliable measurement. Still, without understanding whether the reliability estimates were calculated using risk-adjusted values we cannot interpret the findings appropriately.

MP#5: Results for reliability testing at the data element level were strong and positive – both in terms of the Cronbach's alpha analysis and in terms of the reliability calculation using the CAT approach. The measure seems to be very reliable at the individual patient level. At the measure score level, the measure also seems reliable, as long as a minimum sample size of 20 patients is achieved (and also presuming that the 20 cases are representative of the practice being evaluated). Larger samples produce higher levels of reliability, but samples of 20 allow for reliability at the .7 threshold or beyond.

8. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? NOTE: If multiple methods used, at least one must be appropriate.

MP#2:

a. I was unable to locate information regarding this criterion

Submission document: Testing attachment, section 2a2.2

☐☒ **Yes**

☒☐ **No**

☐ **Not applicable** (score-level testing was not performed)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

MP#2:

a. Could not determine signal-to-noise ratio, thought FOTO reports that 10 patients per provider is inadequate and they will explore 20/provider

Submission document: Testing attachment, section 2a2.2

☐☒ **Yes**

☒☐ **No**

☐ **Not applicable** (data element testing was not performed)

10. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications and <u>all</u> testing results):

☐☒ **High** (NOTE: Can be HIGH <u>only if</u> score-level testing has been conducted)

☒☐ **Moderate** (NOTE: Moderate is the highest eligible rating if score-level testing has <u>not</u> been conducted)

☐ **Low** (NOTE: Should rate <u>LOW</u> if you believe specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

☒☐ **Insufficient** (NOTE: Should rate <u>INSUFFICIENT</u> if you believe you do not have the information you need to make a rating decision)

11. **Briefly explain rationale for the rating of OVERALL RATING OF RELIABILITY and any concerns you may have with the approach to demonstrating reliability.**

MP#4: Reliability testing at the patient level and performance score level are appropriate. No proxy analysis done.

MP#3: Performance score reliability could be an issue for clinician or clinic with small number of patients.

MP#1: Without understanding whether the reliability estimates were calculated using risk-adjusted values (i.e., residuals) we cannot interpret the findings appropriately.

MP#2:

a. **Could not appraise signal-to-noise beyond patient-level data**

MP#5:

As noted above, the methods used were appropriate, and the results of reliability testing at both data element and measure score levels were strong.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

12. **Please describe any concerns you have with measure exclusions.**

MP#3: No concern

MP#2:

a. No concerns

Submission document: Testing attachment, section 2b2.

MP#4: None

MP#5: None.

13. **Please describe any concerns you have regarding the ability to identify meaningful differences in performance.**

MP#2:

a. **Factorial validity was not firmly established surrounding assumption os a single dimension. Item content is somewhet diverse in terms of different neck-related functions. Fit ststistics were not generally supportive of a single dimension: CFI and TLI fit statistics were 0.84 and 0.98, respectively. RMSEA was 0.16. Four items (sleeping more than 1 hour, sleeping through the night, lying flat on your back for 30 minutes, running a block) were removed due to high infit and outfit statistics. This raises concern as to whether a single definable thing is being measured across patients, clinicians, and practices.**

Submission document: Testing attachment, section 2b4.

MP#1: The mean patient score (residual) was 0, with a SD of 12.1. The developer did not provide this information at the clinician or clinic level, however a plot of average clinician (and clinic-level) residuals and 95% CI is provided, showing a modest distribution as the range of performance was fairly narrow (the majority of clinician residual scores ranged between approximately -5 and 3, for clinics the majority of residual scores had a similar spread). This suggests that performance on the measure is likely to be highly clustered.

MP#4: None

MP#5: No concerns – the measure developers have done an unusually good job of linking the distributions of measure scores at the clinic and individual clinician levels to empirically-derived estimates of minimum clinically significant differences.

14. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**

Submission document: Testing attachment, section 2b5.
MP#4: NA
MP#1: N/A
MP#5: None
MP#3: No concern
MP#2: NA

15. **Please describe any concerns you have regarding missing data.**

MP#3: No concern

MP#2:

a. **Discussed and managed well in application**

Submission document: Testing attachment, section 2b6.

MP#4: What constitutes a “complete”? Not really defined.

MP#1: Missing assessment data appears to be fairly high (about 33%) in the testing data. No bias was evident in the baseline FS score comparing patients with complete and incomplete data, so this is less of a concern.

MP#5: None

16. **Risk Adjustment**

16a. **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

16b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

☐ Yes ☐ No ☒☐ Not applicable

16c. **Social risk adjustment:**

16c.1 Are social risk factors included in risk model? ☐☒ Yes ☒☐ No ☐ Not applicable

16c.2 Conceptual rationale for social risk factors included? ☐☒ Yes ☒☐ No

16c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☐☒ Yes ☒☐ No

16d.**Risk adjustment summary:**

16d.1 All of the risk-adjustment variables present at the start of care? ☒ Yes ☐ No

16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☒☐ Yes ☐ No

16d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☐ No

16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☒☐ Yes ☐ No
(adjusted R-sq of .33)

16d.5.Appropriate risk-adjustment strategy included in the measure? ☐☒ Yes ☒☐ No

16e**. Assess the risk-adjustment approach**

MP#4: Appropriate

MP#1: The risk adjustment approach demonstrates a very high adjusted r-square (r^2=.33). However, there are 21 risk adjusters, raising the concern of overfitting.

The data source for the variables are unclear – this measure is intended to be reported using patient data, but some of the covariates would likely need to come from a medical record/EHR given the nature of some of the variables (exercise history, medication use, BMI). Presumably these variables are included on the patient survey.

MP#2: The model identified 11 constructs that explained 33.3% of the variance in discharge FS, with FS at admission, acuity, payer type and age being the most important predictors. R-squared shrinkage was less than 1% for both methods used to assess shrinkage. Risk adjustment seems reasonable.

MP#5: The strategy is reasonable and the results suggest that the model does a good job of controlling for possible biasing or confounding factors that would create unfair comparisons among providers**.** It appears that dual-eligible status is included in the model as a social risk factor, although there is a little ambiguity in the text about that. Education may be included in the future.

MP#3: acceptable

For cost/resource use measures ONLY:

17. **Are the specifications in alignment with the stated measure intent?**

☐ **Yes** ☐ **Somewhat** ☐ **No (If "Somewhat" or "No", please explain)**

18. **Describe any concerns of threats to validity related to attribution, the costing approach, carve outs, or truncation (approach to outliers):**

VALIDITY: TESTING

19. **Validity testing level:** ☐☒ **Measure score** ☐☒ **Data element** ☒☐ **Both**

20. **Method of establishing validity of the measure score:**

    ☒☐ **Face validity**

    ☒ **Empirical validity testing of the measure score**

    ☐ **N/A (score-level testing not conducted)**

21. **Assess the method(s) for establishing validity**

    MP#4: Appropriate (Face validity, construct validation and empirical testing of measure score).

    MP#2:

    a. **Structural validity through CFA, IRT analyses with fit statistics, known groups comparisons, global change ratings.**

    Submission document: Testing attachment, section 2b2.2

    MP#3: Externsive psychometrics evaluations were performed to assess the validity of FS PROM including content validity, structural validity, construct validity and others.

    Validity of performance scores were assessed by comparing % MCII across 3 performance levels of clinician or practice. Similar comparisons were done across performance score deciles.

    MP#1: Data element testing included analysis of face validity and construct validity. DIF analyses were also conducted to evaluate the appropriateness of PRO-PM items. The measure score was also evaluated using known groups differences (construct validity). The developer also evaluated sensitivity to change, functional staging and minimally clinically important improvement (empirically derived). Construct validity testing of the PRO-PM items and measure score were the most illuminating analyses of those presented.

    MP#5:

    The methods used were reasonable and appropriate, particularly at the data element level. At the measure score level, there was careful analysis of the variations in scores at clinic and individual clinician levels, but no apparent attempt to link the provider-level scores to any independent measure or concept of quality of care.

22. **Assess the results(s) for establishing validity**

    MP#4: Adequate sample size to generalize for implementation. Sufficient validity so that conclusions about quality can be made. Agree that score is a validy indicator of quality.

    MP#2:

    a. **Patient-level assessment appears to have sufficielt validity, notwithstanding concerns over unidimensionality and model fit statistics applied to response data.**
    b. **Clinician-level validity less clear**

    Submission document: Testing attachment, section 2b2.3

    MP#3: Results on FS PROM validity assessment were positive and acceptable.

    There were significant differences across 3 performance levels or across deciles.

    MP#1: Overall, the measure's CFA testing (data element testing) and factor loadings supported the one factor structure for the measure. The known group differences provided evidence for the measure score validity.

    MP#5:

    The validity results at the data element level are strong and positive. At the measure score level, there is evidence of variability in performance among clinics and individual providers, but there is no empirical evidence that would clearly demonstrate that the variations represent differences in quality of care.

Validity at the measure score level will depend heavily on the judgement of the standing committee about face validity.

23. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**

    Submission document: Testing attachment, section 2b1.

    ☐☒ **Yes**

    ☐ **No**

    ☐ **Not applicable** (score-level testing was not performed)

24. **Was the method described and appropriate for assessing the accuracy of ALL critical data elements?** *NOTE that data element validation from the literature is acceptable.*

    Submission document*: Testing attachment, section 2b1.*

    ☐☒ **Yes**

    ☒☐ **No**

    ☐ **Not applicable** (data element testing was not performed)

25. **OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.**

    ☐ **High** (NOTE: Can be HIGH only if score-level testing has been conducted)

    ☐☒ **Moderate** (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

    ☒☐ **Low** (NOTE: Should rate LOW if you believe that there are threats to validity and/or relevant threats to validity were not assessed OR if testing methods/results are not adequate)

    ☒☐ **Insufficient** (NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT.)

26. **Briefly explain rationale for rating of OVERALL RATING OF VALIDITY and any concerns you may have with the developers' approach to demonstrating validity.**

    MP#3: It would be more desirable to use some external criterion (not based on chane score) to performance score validity testing.

    MP#1: Although the items appear to be validly assembled and the score demonstrated meaningful known group difference, performance variation was highly constricted, there was minimal variation at the provider and clinic level despite the large testing sample.

    MP#2:

    a. **Unclear score-level validity as it would apply to comparing clinicians and practices.**

    MP#5: As noted above, the validity results are generally strong and positive. At the measure score level, there is evidence of variability in performance, but not compelling evidence of an association between the measure scores and some independent measure or concept of quality of care. Face validity is very important here, and the measure developers don't provide data on some formal effort (e.g., expert panel) to establish face validity.

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

27. **What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?**

☐ **High**

☐ **Moderate**

☐ **Low**

☐ **Insufficient**

28. **Briefly explain rationale for rating of EMPIRICAL ANALYSES TO SUPPORT COMPOSITE CONSTRUCTION**

ADDITIONAL RECOMMENDATIONS

29. **If you have listed any concerns in this form, do you believe these concerns warrant further discussion by the multi-stakeholder Standing Committee? If so, please list those concerns below.**

Committee Pre-evaluation Comments:
Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Specifications: *Which data elements, if any, are not clearly defined? Which codes with descriptors, if any, are not provided? Which steps, if any, in the logic or calculation algorithm or other specifications (e.g., risk/case-mix adjustment, survey/sampling instructions) are not clear? What concerns do you have about the likelihood that this measure can be consistently implemented?*

- No concerns about reliability
- I agree with the questions that MP#1 raised with regard to the specifications needing some additional clarity.

2a2. Reliability testing: *Do you have any concerns about the reliability of the measure?*

- No
- Yes, I am concerned that it is not clear whether the testing was conducted at the individual clinician level or the group level. It seems like it would be a more appropriate measure to use a the group level given that individual clinicians may have a more limited number of patients in their denominators.

2b2. Validity testing: *Do you have any concerns with the testing results?*

- No
- I need more clarity as to why the SMP did not reach consensus on their validity vote. Was that just during their earlier review and they are now satisfied? That seems to be the case, but want to understand it better.

Validity- Threats to Validity: *Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data). 2b4. Meaningful Differences: How do analyses indicate this measure identifies meaningful differences about quality? 2b5. Comparability of performance scores: If multiple sets of specifications: Do analyses indicate they produce comparable results? 2b6. Missing data/no response: Does missing data constitute a threat to the validity of this measure?*

- Would like to hear more about newly added score level validity testing and how missing data is handled.
- No additional comments.

Other Threats to Validity: *Other Threats to Validity (Exclusions, Risk Adjustment). 2b2. Exclusions: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? 2b3. Risk Adjustment: If outcome (intermediate, health, or PRO-based) or resource use performance measure: Is there a conceptual relationship between potential social risk factor variables and the measure focus? How well do social risk factor variables that were available and analyzed align with the conceptual description provided? Are all of the risk-adjustment variables present at the start of care (if not, do you agree with the rationale provided)? Was the risk adjustment (case-mix adjustment) appropriately developed and tested? Do analyses indicate acceptable results? Is an appropriate risk-adjustment strategy included in the measure?*

- No concerns
- No concerns.

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- Some of the data used in the measure are collected by the provider during the provision of care; the remainder is collected from patients (it is not clear if these patient-reported data are collected during a medical encounter).
- The developer notes that data may be collected electronically or via paper.
- The typical amount of time needed by patients to complete the Neck FS PROM is 5 minutes; providers may have to enter some data, but this is minimized if the PROM is integrated with the EHR.
- The developer provides three tiers of access/services and costs related to the use of this measure (i.e., free, minimal services with relatively lower cost; additional services with higher costs)

Questions for the Committee:

- Do you have any questions or concerns about how the data for the measure are collected?
- Is the data collection strategy ready to be put into operational use?
- Do you have any questions or concerns about the costs involved in using the Neck FS PROM or calculating the measure?

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. Feasibility: *Which of the required data elements are not routinely generated and used during care delivery? Which of the required data elements are not available in electronic form (e.g., EHR or other electronic sources)? What are your concerns about how the data collection strategy can be put into operational use?*

- Measure is feasible as it can be administered via paper or computer and takes 5 minutes -thus it is not a burden on the patient. Minimal concerns about the costs for facilities if they are small, could it be a barrier?
- More clarity is needed on the time and cost that clinicians may encounter when using this measure. I am concerned about the fact that clinicians (or groups) need to pay extra in order to be able to use the data for quality improvement rather than simply reporting it for others to judge.

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4a.1. Accountability and Transparency. Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No** ☐ **UNCLEAR**

Accountability program details

- The measure currently is used in the following:
  - 2019 MIPS QCDR (accountability program: payment)
  - The Physical Therapy Provider Network (accountability program: state-level payment)
  - Therapy Partners (TPI) (accountability program: state-level payment; quality improvement program)
- This measure is not publicly reported at this time. However, CMS does plan to make all measures under the MIPS quality performance category available for public reporting on Physician Compare, as feasible.

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others [vetting]

- The developers describe feedback by expert therapists on the Neck FS PROM.
- The developers describe information provided to providers; this includes patient-level information and benchmarked performance results on a 3-month basis. They also describe orientation and training opportunities through a variety of modalities.
- Feedback (e.g., value of the Neck FS PROM; desire for additional risk-adjustment) is obtained from providers using the Neck FS PROM through several mechanisms, including e-mail, phone, and web-conferencing.

Additional Feedback:

- This measure was considered by the MAP Clinician Workgroup for the MIPS program in the 2018-2019 pre-rulemaking process. The MAP conditionally supported this measure pending NQF endorsement. MAP recognized the value of including patient-reported outcome measures in MIPS; however, MAP highlighted the importance that the proprietary survey tool remain freely available to providers.

Questions for the Committee:

- *How has the measure been vetted in real-world settings by those being measured or others?*
- *Do you have any concerns regarding the ability to provide feedback on the measure?*
- *Do you have any concerns regarding the developers' use of feedback when considering modifications to the measure?*

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results

- Developer provided year-over-year data in section 1b.2.
- To assess performance over time, developer compared the mean performance between the two-yearly periods using a paired sample t-test only for providers that had data for both years (2016 and 2017), showing improvement over the two years.
- Developer interprets the result as follows: The wide range of performance scores by deciles of average risk-adjusted residuals at the clinic (- 6.5 to 6.4) and clinician (-7.4 to 7.7) levels demonstrate the presence of notable gaps in provider performance as measured by NQF 3461. Although this is a new measure with data collection starting only in the year 2016, the statistically significant improvement of mean residual scores over these two year periods at the clinic and clinician levels demonstrate preliminary data supporting that improvement over time is feasible.

4b2. Benefits vs. harms**.** Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Unexpected findings (positive or negative) during implementation

- The developer noted no unexpected findings resulting from implementation of this measure.

Potential harms

- The developer did not indicate any potential harms or unexpected benefits from this measure.

***Questions for the Committee**:*

- *Would you like information on potential for improvement in case gained through the use of the measure in the Physical Therapy Provider Network or Therapy Partners programs, if available?*
- *Do you have any concerns regarding the usability of the measure?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for Usability: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

RATIONALE: Staff rated this measure as HIGH, given the lack of harms identified from implementation of this measure. Given that this measure is in use, improvement results should continue to be presented.

Committee Pre-evaluation Comments:
Criteria 4: Usability and Use

4a. Use: *4a1. Use - Accountability and Transparency: How is the measure being publicly reported? Are the performance results disclosed and available outside of the organizations or practices whose performance is measured? For maintenance measures - which accountability applications is the measure being used for? For new measures - if not in use at the time of initial endorsement, is a credible plan for implementation provided? 4a2. Use - Feedback on the measure: Have those being measured been given performance results or data, as well as assistance with interpreting the measure results and data? Have those being measured or other users been given an opportunity to provide feedback on the measure performance or implementation? Has this feedback has been considered when changes are incorporated into the measure?*

- Not currently reported publicly, but will be in the future. Feedback on the measure has been provided by users (clinicians) and patients did provide feedback on the usefulness of questions.
- I would not recommend that this measure be used for the MIPS program as it would be better used at a group, rather than individual clinician, level. More data collection and feedback on the measure need to occur before it is tied to payment or publicly reported. And the feedback being sought now on the users of the measures seems be be pretty high level - whether they like it and can use it or not vs. whether or not they believe that they can use it to improve care for their patients and that their patients are able to see meaningful improvements in their health because of the QI that the measure would hopefully lead to.

it is also not clear if/how the developers are using the feedback they are receiving to improve the measure if needed.

4b. Usability: *4b1. Usability – Improvement: How can the performance results be used to further the goal of high-quality, efficient healthcare? If not in use for performance improvement at the time of initial endorsement, is a credible rationale provided that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations? 4b2. Usability – Benefits vs. harms: Describe any actual unintended consequences and note how you think the benefits of the measure outweigh them.*

- Benefits outweigh the harm; the measure has high usability and the results could be used to provide improved treatment plans in the future.
- It would be helpful to receive information on potential improvement gained by the use of the measure in PT programs. While the measure is promising, it is also not clear if it will ultimately lead to further high-quality, efficient healthcare. It is also very narrowly focused on neck pain - it would seem to benefit from being focused on a broader set of conditions in order to better facilitate change.

Criterion 5: Related and Competing Measures

Competing measure

- 0428: Functional status change for patients with general orthopedic impairments

Harmonization

- NQF may ask the Committee to discuss need for the neck measure and/or make recommendations for harmonizing measures.

Committee Pre-evaluation Comments: Criterion 5: Related and Competing Measures

Related and Competing: *Are there any related and competing measures? If so, are any specifications that are not harmonized? Are there any additional steps needed for the measures to be harmonized?*

- 0428 Functional Status Change for patients with general orthopedic impairments; does not focus specifically on neck pain making it a related but not competing measure. No additional steps needed.
- Yes, there is a competing measure. I need to review that to see if it would be a preferred measure or not. The alternative measure is focused on a broader set of impairments.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: June/13/2019

While this is not an eCQM, we would encourage the measure steward to use a standard terminology such as LOINC for encoding the FIM instrument in their measure. Without this level of standardization, interoperability will be a perpetual challenge, and impact the ability to measure a patient's functional status across the continuum of care.

1. Evidence and Performance Gap – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

NQF_evidence_attachment_Sep2017_Importance_tab_1a_NEW-636915401712959943.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

1a. Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 3461 – previously submitted, not endorsed

Measure Title: Functional Status Change for Patients with Neck Impairments

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: NA

Date of Submission: 4/1/2019

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

Outcome

☐ Outcome:

☒Patient-reported outcome (PRO): functional status

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)

☐ Intermediate clinical outcome (*e.g., lab value*):

☐ Process:

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

Step #1: Patient with neck impairment arrives at an outpatient clinic for initial evaluation by the treating clinician.

Step #2: Patient completes an intake survey including patient characteristics needed for risk adjustment, and the Neck Functional Status (FS) Patient Reported Outcome Measure (PROM).

Step #3: A patient-specific report is produced that describes the data entered, the Neck FS PROM score and its corresponding functional stage, the predicted discharge PROM score derived from the risk-adjusted model, the corresponding predicted discharge functional stage, the minimal detectable change, and the minimal clinically important improvement to assist clinical interpretation of the PROM. (These terms are described in detail within the Measure Testing form in the Scientific Acceptability section).

Step #4: Clinician completes a comprehensive examination and evaluation that includes interpretation of the outcomes data described in Step 3. The data from Step 3 is also factored into the clinician's decision-making and patient communication for establishing individual patient-focused goals and a plan of care. Clinician establishes a plan of care and begins treatment that is tailored to the patient's functional goals as identified in Step 3.

Step #5: The patient is re-evaluated throughout the episode of care. The Neck FS PROM and other components of Step 3 are re-administered and re-calculated periodically as components of the re-evaluations. The timing of re-evaluations is at the discretion of the clinician.

Step #6: Step #5 continues until a decision to end the episode of care (discharge) is reached. The process to end the episode of care includes completing a FOTO Staff Discharge which includes information on number of visits and duration of the care episode.

1a.3 Value and Meaningfulness: IF this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

We surveyed a convenience sample of adults who were currently experiencing neck pain or had experienced neck pain in the past year. For the primary aim of the survey, participants reviewed the 28 items (physical activity questions) of the Neck FS PROM item bank and then responded to the following: *How valuable and meaningful to you are the physical activity questions with regards to your neck pain?*

The survey also asked 4 additional questions:

1. Have you received any treatment for your neck pain during the past year?
2. Are you currently experiencing neck pain?
3. What is your age?
4. What is your sex?

We hypothesized that most or all participants would find at least some of the Neck FS PROM items to be valuable and meaningful to them. Since the Neck FS PROM questions represent a continuum of low to high levels of physical activity (functional status), individuals may vary in how many of the questions they find valuable and meaningful. For example, an individual with high (good) functional status might only find the few most high functioning items to be valuable and meaningful because they have no difficulty with the rest of the items.

We also hypothesized that individuals currently experiencing neck pain and those who had received treatment for neck pain might be more likely to find more of the items to be valuable and meaningful. Differences in rate of response categories by participant characteristic was tested using Pearson Chi-squared.

Table 1 describes the characteristics of the sample of 48 individuals who reported experiencing neck pain at some point in the past year. Sixty-nine percent were female, 73% were between the ages of 35-64, 71% were currently experiencing neck pain, and 44% had previously received treatment for neck pain.

Table 1: Sample Characteristics (N=48)

		Number	**percent**
Sex	Female	33	69
	Male	14	29
	Other	1	2
Age	18-24	3	6
	25-34	7	15
	35-44	16	33
	45-54	11	23
	55-64	8	17
	65+	3	6
	65+	3	6
Current Neck Pain	No	14	29
	Yes	34	71
Previous Treatment	No	27	56
	Yes	21	44

Table 2 shows the results from the primary survey question, *"How valuable and meaningful to you are the survey questions with regards to your neck pain?"* Nearly all participants (92%) felt that some, most, or all of the questions were valuable and meaningful, with 52% responding that most or all questions were valuable and meaningful. These findings were consistent with our hypotheses that the questions would be meaningful to individuals with neck pain and that some degree of variability was expected.

Table 2 Results from primary survey question

	Number	**percent**
No questions are valuable and meaningful	1	2
Few questions are valuable and meaningful	3	6
Some questions are valuable and meaningful	19	40
Most or all questions are valuable and meaningful	25	52

Respondents who were age 45-54 and 55-64 found the questions to be more valuable and meaningful than younger respondents (P value = 0.023), with 73% and 88% feeling that most or all questions we valuable and meaningful, respectivly. We found no differences in degree of value/meaningfulness by sex, having had previous treatment, and having current neck pain (P-value>0.05) .

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

Empirical evidence demonstrating the relationship between the Neck PRO-PM (NQF measure 3461) and the clinical process of administering interim PRO-PM assessments during the episode of care

Background:

We assessed the relationship between NQF measure 3461 (the outcome) to the clinical process of administering interim patient-reported outcome measure (PROM) assessments during the first 2 weeks of the episode of care. We define an ***interim PROM*** as a PROM administered during the patient's episode of care in addition to the intake and discharge PROMs. Interim PROM assessment(s) may be beneficial by providing a treating clinician with immediate patient feedback regarding a patient's functional status, possibly in response to the interventions prescribed during the episode of care. Thereby the clinician can continue or modify the intervention depending on how the patient reports he or she is progressing. Therefore, we consider the administration of interim PROM assessments as a clinical process that, if found to be positively associated with the outcome, could be used by clinicians to improve their patient-reported outcome performance measure (PRO-PM) scores. We hypothesized that clinicians with high rates of ***early interim PROM assessments*** (one or more interims with a first interim within two weeks from admission) would demonstrate significantly better outcomes compared to clinicians with lower rates of interim PROM assessments.

Method: Our hypothesis was tested using several stages. **First**, we identified patients that were administered one or more early interim assessments with the first one administered during the first two weeks from admission. **Second**, we identified all patients that had completed the PROM at admission and discharge only, i.e., had no interim assessments. **Third**, to control for patient baseline characteristics that are associated with the outcome of interest (functional status at discharge), for each patient with an early interim assessment, we matched 1 patient without an interim assessment. Matching was done on all variables used in the Neck Functional Status PROM risk-adjusted model (details on the risk adjusted model are provided within the scientific acceptability testing form). In addition, patient matching was also done for the duration of the episode of care and the number of visits, both of which may be important confounders of the potential for the administration of interim assessments. Only episodes with a treatment-duration of 7 to 180 days and number of treatment visits of 3 to 25, representing the 5th to 95th percentiles, were included. We considered treatment-duration and number of visits for patients being treated in rehabilitation therapy with neck impairments above these thresholds as outliers and below these thresholds as not appropriate for interim PROM administration.

Patient matching was done using a ***propensity score matching*** (PSM) approach using the nearest neighbor method with a caliper of 0.01 on the propensity score.[1, 2] To ensure that the PSM approach matched patients on all risk-adjusted variables successfully, as well as on the episode duration and number of visits, we compared means or rates of all included variables between patients with or without an early interim assessment. For these analyses we considered only clinicians with at least 10 complete episodes in the year 2016, with ***complete episode*** defined as a patient care episode in which a PROM assessment was administered, at minimum, at admission and discharge. **Finally**, data were aggregated at the clinician level to enable the assessment of the relationship between early interim assessments and functional status outcomes at the provider score level. For each clinician, a rate (in percent) of early interim PROM administration were calculated. Then, clinicians were categorized into two groups above or below the median rate of early interim use. ***High interim rate clinicians*** would be those clinicians with a higher percentage of patients with an early interim PROM. ***Low interim rate clinicians*** would be those with a lower percentage of patients with an early interim PROM. We then compared the mean outcome (functional status at discharge) of the two clinician groups using a two-sample t-test.

Higher outcomes for the high interim rate clinician group would provide empirical evidence that there is something that a clinician can do i.e., administer a first interim PROM within 2 weeks after admission, to try to improve their score level outcomes using NQF measure 3461.

Results:

Patients with early interim PROMs (n=6295) were each matched with one patient that had no interim assessment, selected from all available patients that had no interim assessment (n=25,889). The means for continuous variables and rates (%) for categorical variables of the matched samples are provided in Table 1.

Table 1: Comparison of patient baseline characteristics, episode duration, and number of visits, between patients with early interim assessments and their matched sample with no interim assessment

	Early interim	**Matched patients with no interim**
	n=6,295	**n=6,295**
Patient characteristics used for risk-adjustment		
Functional status at admission	51.5	51.5
Age	54.1	54.0
Female	65.0%	65.5%
Days from onset to admission (acuity)		
0-7 days	4.5%	5.1%
8-14 days	8.2%	8.4%
15-21 days	10.7%	9.7%
22-90 days	26.7%	25.4%
91 days to 6 months	13.3%	13.4%
Over 6 months	36.6%	37.9%
Payer		
Indemnity insurance	1.7%	2.0%
Medicaid	2.6%	2.8%
Medicare B Age 65 or above	18.3%	18.4%
Medicare B Under Age 65	3.8%	3.5%
No fault, Auto insurance	4.4%	3.8%
Workers compensation	6.3%	5.9%
health maintenance organization, Preferred Provider	52.0%	56.0%
Surgical history		
No related surgery	89.0%	88.7%
1 related surgery	8.1%	8.3%
2 related surgeries	1.8%	1.8%
3 or more related surgeries	1.1%	1.3%
Exercise history		
At least 3x/week	36.0%	35.8%
1-2x/week	27.5%	27.5%
Seldom or Never	36.5%	36.7%
Medication use at intake	54.9%	54.7%
Received Previous treatment	37.9%	39.2%
Post-surgical: Neck Fusion	1.3%	1.3%
Specific comorbidities:		
Anxiety	19.4%	20.0%
Arthritis	42.8%	43.4%
Back Pain	78.2%	79.3%
Depression	18.2%	19.2%
Gastro-intestinal	19.3%	19.6%

	Early interim	**Matched patients with no interim**
	n=6,295	**n=6,295**
Headache	47.5%	48.4%
Kidney, Bladder, Prostate or Urination	8.8%	9.2%
Obesity	36.8%	35.7%
Osteoporosis	8.9%	8.4%
Previous accidents	14.5%	15.1%
Sleep dysfunction	21.9%	22.4%
Additional confounders		
Number of visits	12.8	12.8
Duration of episode in days	40.0	41.6

Patients treated by clinicians with high rates of early interim assessment (n=1,078) had on average 3 additional functional status points at discharge, compared to those treated by clinicians with low rates of early interim assessment (n=1,212 clinicians). These differences were highly significant (P<0.001) and are described in table 2.

Table 2: Clinician level outcomes by rates of early interim assessments

	Clinicians	**Mean rate of early interim assessments**	**Mean FS at discharge (95% confidence interval)**
Low rates of early interim assessment	1,212	4.6%	62.6 (62.1-63.2)
High rates of early interim assessment	1,078	80.8%	65.6 (65.0-66.2)

Interpretation: The differences in outcomes between clinicians with high or low rates of early interim assessments reported above provide empirical evidence supporting our hypothesis that administering a first interim during the first 2 weeks of the episode of care is an important and feasible clinical process associated with higher patient outcomes as assessed using NQF Measure 3461.

References:

1. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Stat Med*. 2014;33:1057-1069.

2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41-55.

1a.3. SYSTEMATIC REVIEW(S) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

NA - This is not an intermediate outcome, process, or structure performance measure and we responded to **1a.2**

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review 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 a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐ Other

NA

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	
Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	
Provide all other grades and definitions from the recommendation grading system	
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	
Estimates of benefit and consistency across studies	
What harms were identified?	
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

1a.4 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.

NA due to response to **1a.2**

1a.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

1a.4.2 What process was used to identify the evidence?

1a.4.3. Provide the citation(s) for the evidence.

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

Neck pain is recognized as a global healthcare burden.1,2 Prevalence estimates from epidemiologic studies on neck pain, defined as pain in the neck with or without pain referred into one or both upper limbs that lasts for at least 1 day, have a mean 1-year prevalence range from 23% 4 to 37% 3 and mean lifetime prevalence of 49%.3 The use of patient-reported outcome measures (PROMs) for assessing functional status in patients with neck pain is an essential step in addressing this burden, provided 1) scores can be interpreted in clinically useful ways to inform patient-centered clinical decision making, 2) performance across providers can be reliably and validly assessed, and 3) a performance gap between providers can be identified setting an opportunity for improvement over time. For example, results (Measure Testing form FIGURE 2b4ii-iii) demonstrate performance gaps that may form a basis for improvements in quality envisioned for this measure; we expect providers ranked as having low quality (mean residuals below zero) to improve over time to average or high quality (mean residuals zero or above) Evidence supporting that NQF measure 3461 can successfully address these 3 elements is outlined above and is described in the Measure Testing form in the Scientific Acceptability section.

PROMs are increasingly advocated as necessary components of an overall strategy to improve healthcare5,6 and are advocated for use in clinical decision making in clinical practice guidelines.7-11 However, prior to the development of the Neck Functional Status (FS) PROM, the literature lacked a neck-specific functional status PROM based on modern scientific measurement methods like item response theory (IRT) and computer adaptive testing (CAT).12-16 Further, when modern measurement theory approaches were applied to previously existing measures, psychometric limitations were discovered including floor and ceiling effects, invalid assumptions of interval scaling, and multi-dimensionality. The Neck FS PROM is free of such limitations.17-21 Furthermore, using the Neck FS PROM reduces patient burden by minimizing the number of functional questions the patient must respond to in order to obtain a precise estimate of the patient's functional ability level.22-23

When combined with robust risk adjustment ,24 the IRT-based Neck FS PROM forms the basis for a valuable patient reported outcome performance measure (PRO-PM). Placing risk-adjusted Neck FS PROM data directly into the hands of the provider embodies the definition of patient-centered healthcare and is consistent with National Quality Forum's vision to achieve performance improvement and accountability through patient-reported outcomes.25 This approach improves quality of care by promoting improved communication between provider and patient and enhancing the provider's understanding of the patient's perception of functional status. The Neck FS PROM and PRO-PM results can even be shared with the patient to further promote patient engagement; as one example, the FOTO Outcomes Measurement system provides a visually pleasing, patient-focused real-time report of the patient's (risk-adjusted) PRO-PM results.

1. Hoy D, March L, Woolf A, et al. The global burden of neck pain: estimates from the global burden of disease 2010 study. Ann Rheum Dis. 2014;73:1309-1315.

2. Hurwitz EL, Randhawa K, Yu H, Cote P, Haldeman S. The Global Spine Care Initiative: a summary of the global burden of low back and neck pain studies. Eur Spine J. 2018;27:796-801.

3. Fejer R, Kyvik KO, Hartvigsen J. The prevalence of neck pain in the world population: a systematic critical review of the literature. Eur Spine J. 2006;15:834-848.

4. Hoy DG, Protani M, De R, Buchbinder R. The epidemiology of neck pain. Best Pract Res Clin Rheumatol. 2010;24:783-792.

5. Black N. Patient reported outcome measures could help transform healthcare. BMJ. 2013;346:f167.

6. Griggs CL, Schneider JC, Kazis LE, Ryan CM. Patient-reported outcome measures: a stethoscope for the patient history. Ann Surg. 2017;265:1066-1069.

7. Blanpied PT, Gross AR, Elliott JM, et al. Neck Pain: Revision 2017 Clinical Practice Guidelines Linked to the International Classification of Functioning, Disability and Health from the Orthopaedic Section of the American Physical Therapy Association. J Orthop Sports Phys Ther. 2017;47(7):A1-A83. doi:10.2519/jospt.2017.0302.

8. Childs JD, Cleland JA, Elliott JM, Teyhen DS et al. Neck Pain: Clinical Practice Guidelines Linked to the International Classification of Functioning, Disability and Health from the Orthopaedic Section of the American Physical Therapy Association. J Orthop Sports Phys Ther. 2008;38:A1-A34.

9. Baisden J, Easa J, Fernand R, Lamer T, et al. North American Spine Society Evidence-Based Clinical Guidelines for Multidisciplinary Spine Care Diagnosis and Treatment of Cervical Radiculopathy from Degenerative Disorders. North American Spine Society 2010.

10. Bono CM, Ghiselli G, Gilbert TJ, Kreiner S et al. An evidence-based clinical guideline for the diagnosis and treatment of cervical radiculopathy from degenerative disorders. Spine J 2011;11:64-72.

11. Bier JD, Scholten-Peeters WGM, Staal JB, Pool J et al. Clinical Practice Guideline for Physical Therapy Assessment and Treatment in Patients With Nonspecific Neck Pain. Phys Ther 2018:98:162-171.

12. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther. 1991;14:409-415.

13. Jordan A, Manniche C, Mosdal C, Hindsberger C. The Copenhagen Neck Functional Disability Scale: a study of reliability and validity. J Ma¬nipulative Physiol Ther. 1998;21:520-527.

14. Wheeler AH, Goolkasian P, Baird AC, Darden BV, 2nd. Development of the Neck Pain and Disability Scale. Item analysis, face, and criterion-related validity. Spine (Phila Pa 1976). 1999;24:1290-1294.

15. BenDebba M, Heller J, Ducker TB, Eisinger JM. Cervical Spine Outcomes Questionnaire: its de¬velopment and psychometric properties. Spine (Phila Pa 1976). 2002;27:2116-2123; discussion 2124.

16. Leak AM, Cooper J, Dyer S, Williams KA, Turner-Stokes L, Frank AO. The Northwick Park Neck Pain Questionnaire, devised to measure neck pain and disability. Br J Rheumatol. 1994;33:469-474.

17. Hung M, Cheng C, Hon SD, Franklin JD, Lawrence BD, Neese A, Grover CB, Brodke DS. Challenging the norm: further psychometric investigation of the Neck Disability Index. The Spine Journal. 2015;15(11):2440 – 2445.

18. Van der Velde G, Beaton D, Hogg-Johnston S, Hurwitz E, Tennant A. Rash analysis provides new insights into the measurement properties of the Neck Disability Index. Arthritis & Rheumatism. 2009;61(4):544-551.

19. Ailliet L, Knol DL, Rubinstein SM, de Vet HC, van Tulder MW, Terwee CB. Definition of the construct to be measured is a prerequisite for the assessment of validity. The Neck Disability Index as an example. J Clin Epidemiol. 2013; 66(7): 775-782

20. Wang YC, Cook KF, Deutscher D, Werneke MW, Hayes D, Mioduski JE. The Development and Psychometric Properties of the Patient Self-Report Neck Functional Status Questionnaire (NFSQ). J Orthop Sports Phys Ther. 2015;45:683-692.

21. Deutscher D, Cook KF, Kallen MA, et al. Clinical Interpretation of the Neck Functional Status Computer Adaptive Test. J Orthop Sports Phys Ther. 2019;Accepted:

22. Swaminathan H, Hambleton R. Fundamentals of item response theory. Newbury Park [CA]: Sage Publications; 1991. In: Stephanie Nikolaus, JMIR Human Factors. 2014;1(1):e4.

23. Sands WA, Waters BK, McBride JR, editors. Computerized adaptive testing: from inquiry to operation. Washington, DC: American Psychological Association; 1997.

24. Deutscher D, Werneke MW, Hayes D, et al. Impact of Risk Adjustment on Provider Ranking for Patients with Low Back Pain Receiving Physical Therapy. J Orthop Sports Phys Ther. 2018;48(8):637-648.

25. National Quality Forum. Patient Reported Outcomes (PROs) in Performance Measurement. January 10, 2013. https://www.qualityforum.org/publications/2012/12/patient-reported_outcomes_in_performance_measurement.aspx Accessed March 21, 2019.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

Method:

Performance scores were assessed at the provider level based on clinics' and clinicians' average residual scores derived from the Neck Functional Status PROM risk-adjusted model (details on the risk adjusted model are provided within the scientific acceptability testing form). Data from all patients that had completed the Neck Functional Status PROM at intake and discharge from January 1st 2016 to December 31st 2017 were considered (Sample details are described within the scientific acceptability testing -TABLE 1.6.IV). For these analyses only providers who met the threshold used for all other provider-level testing were included [i.e., clinicians with 10+ patients per calendar year for the clinician level, and clinics with 10+ patients per clinician per calendar year for small clinics (up to 4 clinicians) or 40+ patients per calendar year for large clinics (5 or more clinicians) for the clinic level]. Mean performance and the associated standard deviation, inter-quartile range, minimum and maximum residuals we calculated for the overall sample of clinics and clinicians and are also presented by provider decile ranking. To assess performance over time, we compared the mean performance between the two-yearly periods using a paired sample t-test only for providers that had data for both years (2016 and 2017).

Results:

Number of patients, clinics, clinicians and states for the analyses at the clinic and clinician levels:

	Clinic level	**Clinician level**
Patients	123,194	112,178
Clinicians	7,025	4,537
Clinics	1,378	1,913
States	49	50

Performance scores by decile ranking at the clinic level:

Deciles	N	mean	Standard deviation	Inter Quartile Range	Minimum	Maximum
1	138	-6.5	1.9	1.8	-14.0	-4.7
2	138	-3.9	0.3	0.6	-4.7	-3.4
3	138	-2.8	0.3	0.4	-3.4	-2.4
4	138	-1.9	0.3	0.5	-2.4	-1.4
5	137	-1.1	0.2	0.4	-1.4	-0.7
6	138	-0.3	0.2	0.5	-0.7	0.1
7	138	0.5	0.3	0.4	0.1	1.0
8	138	1.6	0.3	0.6	1.0	2.2
9	138	2.9	0.5	0.8	2.2	3.8
10	137	6.4	2.9	2.6	3.8	22.1
Total	1378	-0.5	3.6	4.3	-14.0	22.1

Performance scores by decile ranking at the clinician level:

Deciles	N	mean	Standard deviation	Inter Quartile Range	Minimum	Maximum
1	454	-7.4	1.6	2.0	-14.1	-5.5
2	454	-4.7	0.4	0.7	-5.5	-4.0
3	454	-3.3	0.4	0.7	-4.0	-2.7
4	453	-2.2	0.3	0.5	-2.7	-1.6
5	454	-1.1	0.3	0.5	-1.6	-0.6
6	454	-0.1	0.3	0.5	-0.6	0.4
7	453	0.9	0.3	0.5	0.4	1.5
8	454	2.2	0.4	0.8	1.5	3.0
9	454	3.9	0.6	1.0	3.0	5.0
10	453	7.7	2.7	2.6	5.0	22.1
Total	4537	-0.4	4.3	5.5	-14.1	22.1

Comparison of performance scores by year for clinics contributing data during 2016 and 2017 (P-value <0.001):

Year	Number of clinics	Mean	Standard error	Standard deviation	95% Confidence lower level	95% Confidence upper level
2016	680	-0.32	0.14	3.68	-0.60	-0.05
2017	680	0.46	0.14	3.66	0.19	0.74

Comparison of performance scores by year for clinicians contributing data during 2016 and 2017 (P-value <0.001):

Year	Number of clinics	Mean	Standard error	Standard deviation	95% Confidence lower level	95% Confidence upper level
2016	1686	-0.16	0.11	4.51	-0.37	0.06
2017	1686	0.47	0.11	4.50	0.26	0.69

Interpretation:

The wide range of performance scores by deciles of average risk-adjusted residuals at the clinic (- 6.5 to 6.4) and clinician (-7.4 to 7.7) levels demonstrate the presence of notable gaps in provider performance as measured by NQF 3461. Although this is a new measure with data collection starting only in the year 2016, the statistically significant improvement of mean residual scores over these two year periods at the clinic and clinician levels demonstrate preliminary data supporting that improvement over time is feasible.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

NA. Adequate performance data for 1b2 was available.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

Method:

Disparities were assessed using three patient characteristics included as independent variables within the Neck Functional Status PROM risk-adjusted model predicting functional status at discharge (details on the risk adjusted model are provided within the scientific acceptability testing form). We compared unstandardized beta coefficients and their 95% confidence intervals between 5 age groups, sex (male or female), and 10 categories of insurance status, after controlling for all other variable included in the risk-adjusted model (i.e., functional status at admission; acuity levels; surgical and exercise history; medication use at intake, receiving previous treatment for the same condition; having a neck fusion surgery; and specific comorbidities). Data from all patients that had completed the Neck Functional Status PROM at intake and discharge from January 1st 2016 to December 31st 2017 were included in this analyses (Sample details are described within the scientific acceptability testing -TABLE 1.6.IV). To assess whether group differences differed over time, we also assessed the beta coefficients separately for each year of data collection (2016 and 2017).

Results:

Unstandardized beta coefficients (95% confidence level) predicting functional status at discharge for the overall data collection period (years 2016-2017) and by year.

Variable	**2016-2017 (n=169,039)**	**2016 (n=79,616)**	**2017 (n=89,423)**
Age groups			
14 to <18	8.4 (7.6 to 9.1)	8.6 (7.5 to 9.8)	8.2 (7.2 to 9.2)
18 to <45	5.7 (5.2 to 6.3)	5.9 (5.1 to 6.7)	5.6 (4.9 to 6.4)
45 to <65	3.1 (2.6 to 3.7)	3.4 (2.7 to 4.2)	2.9 (2.2 to 3.6)
65 to <85	2.4 (1.9 to 2.9)	2.5 (1.8 to 3.2)	2.4 (1.7 to 3.1)
85 or more (reference)			
Gender			
Female	-1.1 (-1.2 to -1.0)	-1.1 (-1.3 to -1.0)	-1.0 (-1.2 to -0.9)
Male (Reference)			

Variable	**2016-2017 (n=169,039)**	**2016 (n=79,616)**	**2017 (n=89,423)**
Insurance status			
Indemnity insurance	-2.7 (-3.0 to -2.3)	-2.2 (-2.7 to -1.7)	-3.1 (-3.6 to -2.6)
Medicaid	-3.4 (-3.7 to -3.1)	-3.2 (-3.6 to -2.8)	-3.6 (-4.0 to -3.2)
Medicare A	-1.5 (-2.0 to -0.9)	-1.0 (-1.8 to -0.2)	-1.8 (-2.6 to -1.1)
Patient	-0.2 (-0.9 to 0.5)	-0.3 (-1.4 to 0.8)	-0.1 (-1.1 to 0.9)
Workers compensation	-5.4 (-5.7 to -5.1)	-5.5 (-5.9 to -5.1)	-5.2 (-5.6 to -4.8)
Other (Litigation, Medicare C, School, No charge, Early Intervention, Commercial Insurance)	-0.6 (-0.7 to -0.4)	-0.6 (-0.9 to -0.3)	-0.6 (-0.8 to -0.3)
No fault, Auto insurance	-2.6 (-2.9 to -2.3)	-2.4 (-2.8 to -2.0)	-2.7 (-3.1 to -2.3)
Medicare B under age 65	-2.5 (-2.8 to -2.2)	-2.5 (-3.0 to -2.1)	-2.5 (-2.9 to -2.0)
Medicare B age 65 or above	-0.2 (-0.4 to 0.0)	0.0 (-0.3 to 0.3)	-0.3 (-0.6 to 0.0)

Health Maintenance Organization; Preferred Provider Organization (reference)

Interpretation:

These results demonstrate significant differences in risk-adjusted outcomes by age groups, with lower outcomes achieved by older patients. Females had about 1 less functional status points at discharge compared to males. Compared to patients insured by a health maintenance organization or other preferred providers, those with worker's compensation and Medicaid payers had the lowest functional status at discharge. These disparities were stable over the two-year period analyzed here. These results are also consistent with results from other spine-related risk-adjustment models.1 Studies on whether these disparities can be decreased using specific treatment approaches for specific patient groups are warranted.

1. Deutscher D, Werneke MW, Hayes D, et al. Impact of Risk Adjustment on Provider Ranking for Patients With Low Back Pain Receiving Physical Therapy. J Orthop Sports Phys Ther. 2018;48:637-648.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4

NA. Adequate data was available for 1b.4

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. Subject/Topic Area *(check all the areas that apply):*

De.6. Non-Condition Specific*(check all the areas that apply):*

De.7. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

S.1. Measure-specific Web Page *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

https://www.fotoinc.com/science-of-foto/nqfneck

S.2a. If this is an eMeasure, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)

This is not an eMeasure **Attachment:**

S.2b. Data Dictionary, Code Table, or Value Sets *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*

Attachment **Attachment:** Neck_PRO_PM_CodeBook_-_RA_Coefficients_20180731_ICD_10_codes.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

Attachment **Attachment:** Item_bank_for_the_Neck_FS_CAT.xlsx

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

Patient

S.3.1. For maintenance of endorsement: Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.

S.3.2. For maintenance of endorsement, please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

NA

S.4. Numerator Statement *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*

IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

The numerator is based on residual scores (actual change scores - predicted change after risk adjustment) of patients receiving care for neck impairments and who: a) completed the Neck PRO-PM at admission and at the end of the episode of care; and b) were discharged from care.

S.5. Numerator Details *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*

IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

Patient Level: The residual functional status score for the individual patient (the residual score is the actual change score - predicted change after risk adjustment).

Clinician Level: The average of residuals in functional status scores in patients who were treated by the clinician in a 12 month period.

Clinic Level: The average of residuals in functional status scores in patients who were treated by the clinic in a 12 month period.

Further details are provided in the Measure Testing Form

S.6. Denominator Statement *(Brief, narrative description of the target population being measured)*

All patients 14 years and older with a neck impairment who have initiated an episode of care and completed the neck functional status PROM at admission and discharge.

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*

IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

All patients 14 years and older with a neck impairment who have an episode of care and completed the neck functional status PROM at admission and discharge.

An episode is considered completed and the patient discharged when the clinician ceases to provide care for the neck impairment as signified by a discharge from that care. For clinicians who use the FOTO system, the completion of an episode is formally signified when the clinician or clinician's representative completes a short process called a FOTO Staff Discharge which includes completing data fields for the date of the last care visit and the total number of visits used in the episode of care.

The ICD-10-CM codes relevant for this measure are included below.

G54.2; G54.8; G55; G89.29; M05.69; M05.79; M05.89; M06.08; M06.28; M06.38; M06.88; M08.08; M08.1; M08.28; M08.48; M08.88; M08.98; M11.08; M11.18; M11.28; M11.88; M12.08; M12.18; M12.28; M12.48; M12.58; M12.88; M13.0; M13.88; M14.68; M14.88; M15.0; M15.3; M15.4; M15.8; M15.9; M19.90; M19.91; M19.92; M19.93; M24.08; M24.10; M24.28; M24.80; M24.9; M25.28; M25.30; M25.50; M25.60; M25.78; M25.80; M25.9; M32.10; M32.19; M32.8; M32.9; M40.03; M40.12; M40.13; M40.202; M40.203; M40.292; M40.293; M41.112; M41.113; M41.122; M41.123; M41.22; M41.23; M41.41; M41.42; M41.43; M41.52; M41.53; M41.82; M41.83; M42.01; M42.02; M42.03; M42.11; M42.12; M42.13; M43.01; M43.02; M43.03; M43.11; M43.12; M43.13; M43.21; M43.22; M43.23; M43.3; M43.4; M43.5X2; M43.5X3; M43.6; M43.8X1; M43.8X2; M43.8X3; M45.1; M45.2; M45.3; M46.01; M46.02; M46.03; M46.21; M46.22; M46.23; M46.31; M46.32; M46.33; M46.41; M46.42; M46.43; M46.51; M46.52; M46.53; M46.81; M46.82; M46.83; M46.91; M46.92; M46.93; M47.11; M47.12; M47.13; M47.21; M47.22; M47.23; M47.811; M47.812; M47.813; M47.891; M47.892; M47.893; M48.01; M48.02; M48.03; M48.11; M48.12; M48.13; M48.21; M48.22; M48.23; M48.31; M48.32; M48.33; M48.41; M48.42; M48.43; M48.51; M48.52; M48.53; M48.8X1; M48.8X2; M48.8X3; M49.81; M49.82; M49.83; M50.00; M50.01; M50.020; M50.021; M50.022; M50.023; M50.03; M50.10; M50.11; M50.120; M50.121; M50.122; M50.123; M50.13; M50.20; M50.21; M50.220; M50.221; M50.222; M50.223; M50.23; M50.30; M50.31; M50.320; M50.321; M50.322; M50.323; M50.33; M50.80; M50.81; M50.820; M50.821; M50.822; M50.823; M50.83; M50.90; M50.91; M50.920; M50.921; M50.922; M50.923; M50.93; M53.0; M53.1; M53.2X1; M53.2X2; M53.2X3; M53.81; M53.82; M53.83; M54.11; M54.12; M54.13; M54.2; M54.81; M54.89; M54.9; M62.830; M62.838; M62.89; M63.88; M65.28; M65.88; M66.18; M70.88; M70.98; M71.48; M71.58; M71.88; M79.12; M79.7; M80.08; M80.88; M81.0; M81.6; M81.8; M85.88; M89.8X8; M93.28; M93.88; M93.98; M95.3; M96.1; M99.01; M99.11; M99.21; M99.31; M99.41; M99.51; M99.61; M99.71; M99.81; Q76.1; Q76.2; Q76.3; Q76.411; Q76.412; Q76.413; Q76.49; R25.2; R29.3; R29.898; R29.91; R51; S12.000; S12.001; S12.01; S12.02; S12.030; S12.031; S12.040; S12.041; S12.090; S12.091; S12.100; S12.101; S12.110; S12.111; S12.112; S12.120; S12.121; S12.130; S12.131; S12.14; S12.150; S12.151; S12.190; S12.191; S12.200; S12.201; S12.230; S12.231; S12.24; S12.250; S12.251; S12.290; S12.291; S12.300; S12.301; S12.330; S12.331; S12.34; S12.350; S12.351; S12.390; S12.391; S12.400; S12.401; S12.430; S12.431; S12.44; S12.450; S12.451; S12.490; S12.491; S12.500; S12.501; S12.530; S12.531; S12.54; S12.550; S12.551; S12.590; S12.591; S12.600; S12.601; S12.630; S12.631; S12.64; S12.650; S12.651; S12.690; S12.691; S12.8; S12.9; S13.0; S13.100; S13.101; S13.110; S13.111; S13.120; S13.121; S13.130; S13.131; S13.140; S13.141; S13.150; S13.151; S13.160; S13.161; S13.170; S13.171; S13.180; S13.181; S13.20; S13.29; S13.4; S13.5; S13.8;

S13.9; S14.2; S14.8; S14.9; S16.1; S16.2; S16.8; S16.9; S19.80; S19.89; T85.850; Z82.61 FOR ICD-10 CODES WITH DESCRIPTORS PLEASE SEE CODE BOOK ATTACHED IN SECTION S2b

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*

Patients who are not being treated for a neck impairment. Patients who are less than 14 years of age.

S.9. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*

NA

S.10. Stratification Information *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*

The methods used to develop the FOTO risk-adjustment neck model were the same as the methods described in detail in a recent publication by Deutscher et at, 2018 [Deutscher D, Werneke MW, Hayes D, et al. Impact of Risk-Adjustment on Provider Ranking for Patients With Low Back Pain Receiving Physical Therapy. J Orthop Sports Phys Ther. 2018;1-35.] Briefly, we used data from adult patients with neck pain treated in outpatient physical therapy clinics during 2016, that had complete outcomes data at admission and discharge, to develop the risk-adjustment model. The data included the following patient factors that could be evaluated for inclusion in a model for risk-adjustment: FS at admission (continuous); age (continuous); sex (male/female); acuity as number of days from onset of the treated condition (6 categories); type of payer (10 categories); number of related surgeries (4 categories); exercise history (3 categories); use of medication at intake for the treatment of LBP (yes/no); previous treatment for LBP (yes/no); treatment post-surgery (lumbar fusion, laminectomy or other); and 31 comorbidities.

Please see Measure Testing Form section 2b3 for more details.

S.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)

Statistical risk model

If other:

S.12. Type of score:

Continuous variable, e.g. average

If other:

S.13. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*

Better quality = Higher score

S.14. Calculation Algorithm/Measure Logic (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*

A Residual score is defined as an actual change score minus the risk-adjusted predicted change score. The Residual(s) are calculated at three levels:

- Patient Level: The residual Neck FS Change score for the individual patient.
- Individual Clinician Level: The average of residuals for change in Neck FS scores in patients who were treated by a clinician in a 12-month time period.

- Clinic Level: The average of residuals for change in Neck FS scores in patients who were treated within a clinic in a 12-month time period.

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

From the reliability at the provider level analysis, the minimum sample sizes needed to achieve a minimum reliability threshold of 0.7 are as follows:

Reliability results are presented by groups of providers based on their number of patients with complete episodes per year, i.e., completed the PRO-PM at admission and discharge.

Average reliability, as well as minimum and maximum reliability coefficients and the proportion of providers that have reliability coefficients >0.7 are detailed in the testing form (TABLE 2a2iii). In summary, the average reliability of clinics meeting the FOTO unique threshold of number of patients per clinic for quality reporting was 0.79. At the clinician level, average reliability for clinicians with 10 or more, or 20 or more patients per year was 0.64 and 0.76, respectively, suggesting a minimum of 20 patients per 12-month period is preferred.

For patients who are unable to respond to questions independently, the FOTO system allows for both Proxy and Recorder modes of administration. Below are the descriptions and data entry fields as seen by providers in the FOTO system:

A PROXY should be used if someone else will be answering the questions on the patient's behalf for any of the following reasons (select all that apply):

- Cognitive Issues (i.e., pt. cannot give accurate answers about their health or cannot answer reliably. For example, the patient has dementia or had a stroke that caused cognitive problems.)
- Age less than 8 years old
- Patient is > 8 yrs old but is uncomfortable responding independently
- If a proxy was used, please indicate if the proxy was:
- spouse
- parent
- child over 8
- other family member
- friend or companion, not family member
- caregiver
- office staff
- clinician (not recommended unless no other option is available)

Does proxy live with the patient?

- Yes
- No

A RECORDER should be used if the patient provides all of the answers independently, but someone else will enter the responses for any of the following reasons (select all that apply):

- Language Barrier (Patient cannot read English or other language that the surveys are in)
- Difficulty Reading (Patient has trouble reading but can answer reliably)
- Motor Impairment (Patient cannot enter their own responses due to problems with their hand, arm, or etc.)

- Visual Impairment (Patient cannot enter their own responses due to difficulty seeing)
- Patient uncomfortable using computer technology
- Telephone survey (i.e., the survey was administered over the phone)

If a recorder was used, please indicate if the recorder was:

- spouse
- parent
- child over 8
- other family member
- friend or companion, not family member
- caregiver
- office staff
- clinician (not recommended unless no other option is available.)

Proxy use was very rare within our data (0.02%). Thus, we did not assess proxy data separately in our analyses.

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*

Specify calculation of response rates to be reported with performance measure results.

Minimum response rate is NA because the computer adaptive test will continue to ask questions until a low level of standard error is achieved. On the short form version, all 10 items must be responded to.

Patient instructions are:

The following assessment will ask you about difficulties you may have with certain activities.

It's an important part of your evaluation. It will help us:

- understand how your condition is affecting your activities, and
- develop treatment goals with you.

Please answer the questions with respect to the problem for which we are seeing you. Respond based on how you have been over the past few days.

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.18.

Instrument-Based Data

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*

IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.

The data source is the Focus on Therapeutic Outcomes measurement and reporting system. The instruments are the Neck FS PROM and risk adjustment questions (as described in the Measure Testing Form) which are administered via computer administration.

S.19. Data Source or Collection Instrument *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

Available at measure-specific web page URL identified in S.1

S.20. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*

Clinician : Group/Practice, Clinician : Individual

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

Outpatient Services

If other:

S.22. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

na

2. Validity – See attached Measure Testing Submission Form

Validity_of_performance_score_level__3461_additional_information_April_5.docx,Neck_Testing_Form_Jan_6_2019_alt_text_added_April_21_2019.docx

2.1 For maintenance of endorsement

Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.2 For maintenance of endorsement

Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.3 For maintenance of endorsement

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:**
Measure Title: Functional Status Change for Patients with Neck Impairments
Date of Submission: 1/7/2019

Type of Measure:

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

TABLE OF CONTENTS

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.

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.17*)**	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ claims	☐ claims
☐ registry	☐ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: Clinical Database	☒ other: Clinical Database

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).

NA

1.3. What are the dates of the data used in testing?

Different aspects of testing utilized different years of data and samples. See TABLE 1.5

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.20*)**	**Measure Tested at Level of:**
☒ individual clinician	☒ individual clinician
☒ group/practice	☒ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☐ health plan	☐ health plan
☒ other: individual patient level	☒ other: individual patient level

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)

See Table 1.5 below

TABLE 1.5: Measured Entities by level of analysis and Data Source

Analysis	Data source (years)	Entities tested			
		Patients	Clinicians	Clinics	States
2a2. RELIABILITY TESTING					
2a2i. **Data elements (patient) level:** Internal consistency (Using both Cronbach's alpha & IRT person reliability)	Wang et al 2015:[1] (2007- 2012)	439	NR+	56	18
2a2ii. **Data elements (patient) level:** Reliability of point estimates and change scores, using computerized adaptive test data.	FOTO internal analysis (2016-2017)	169,039	15,524	3,578	50+DC
2a2iii. **Clinician performance score level:** at different sample thresholds per clinician per calendar year*	FOTO internal analysis (2016-2017)	112,178	4,711	1,913	50
2a2iv. **Clinic performance score level:** at different sample thresholds per clinic per calendar year**	FOTO internal analysis (2016-2017)	123,194	7,025	1,378	49
2b1. VALIDITY TESTING					
2b1i. **Data elements (patient) level:** Content validity (Do test items appear to be measuring the construct of interest?); Structural validity (uni-dimensionality, local independence and item fit); Differential Item Functioning	Wang et al 2015:[1] (2007- 2012)	439	NR+	56	18
2b1ii. **Data elements (patient) level:** Known groups construct validity; Sensitivity to change; Functional staging	FOTO internal analysis (2016-2017)	169,039	15,524	3,578	50+DC
2b1iii. **Data elements (patient) level:** Clinically important improvement	FOTO internal analysis (2016-2017)	126,026	13,402	3,281	50+DC
2b1iv. **Clinician performance score level:** Validity of performance Classification*	FOTO internal analysis (2016-2017)	112,178	4,711	1,913	50
2b1v. **Clinic performance score level:** Validity of performance Classification**	FOTO internal analysis (2016-2017)	123,194	7,025	1,378	49
2b2. EXCLUSIONS ANALYSIS					
2b2. Age exclusion	FOTO internal analysis (2016-2017)	169,039	15,524	3,578	50+DC
2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES					
2b3.Risk adjustment model development	FOTO internal analysis (2016)	77,277	10,348	2,886	50+DC

Analysis	Data source (years)	Entities tested			
		Patients	Clinicians	Clinics	States
2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE					
2b4i. Performance patient level	FOTO risk-adjustment development dataset (2016)	77,277	10,348	2,886	50+DC
2b4ii. Performance individual clinician level*	FOTO internal analysis (2016-2017)	112,178	4,711	1,913	50
2b4iii. Performance clinic/group practice level **	FOTO internal analysis (2016-2017)	123,194	7,025	1,378	49
2b4iv. Validity of Provider Classification – clinician level*	FOTO internal analysis (2016-2017)	112,178	4,711	1,913	50
2b4v. Validity of Provider Classification – clinic level**	FOTO internal analysis (2016-2017)	123,194	7,025	1,378	49
2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS					
2b6i. Comparing patients with or without complete outcomes	FOTO internal analysis (2016-2017)	250,741	17,110	3691	50+DC
2b6ii. Correlations between clinician residuals and their completion rates for clinicians participating in the performance analyses	FOTO internal analysis (2016-2017)	112,178	4,711	1,913	50
2b6iii. Correlations between clinic residuals and completion rates for clinics participating in the performance analyses	FOTO internal analysis (2016-2017)	123,194	7,025	1,378	49
2b6iv. Average residuals at the clinician level by completion rate categories with or without the use of Inverse Probability Weighting*	FOTO internal analysis (2016-2017)	112,178	4,711	1,913	50
2b6v. Average residuals at the clinic level by completion rate categories with or without the use of Inverse Probability Weighting**	FOTO internal analysis (2016-2017)	123,194	7,025	1,378	49
+ NR=not reported **Clinicians with 10+ patients per calendar year with FS measures at intake & discharge.* ***Clinics with 10+ patient per clinician per calendar year for small clinics (up to 4 clinicians) or 40+ patients per calendar year for large clinics (5 or more clinicians), with FS measures at intake & discharge* ***Abbreviations:*** *FS = functional status*					

1.6. How many and which patients 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)

TABLE 1.6: Patient sample by level of analysis and Data Source

Analysis	Data source (years)	Sample selection criteria	TABLE number
2a2. RELIABILITY TESTING			
2a2i. **Data elements (patient) level:** Internal consistency (Using both Cronbach's alpha & IRT person reliability)	Wang et al 2015:[1] (2007- 2012)	Patients responding to the full item bank considered for the measure development	TABLE 1.6.I
2a2ii. **Data elements (patient) level:** Reliability of point estimates and change scores, using computerized adaptive test data.	FOTO internal analysis (2016-2017)	Patients with FS scores at intake & discharge	TABLE 1.6.IV
2a2iii. **Clinician performance score level:** at different sample thresholds per clinician per calendar year*	FOTO internal analysis (2016-2017)	Patients treated by clinicians with 10+ patients per calendar year with FS scores at intake & discharge	TABLE 1.6.II
2a2iv. **Clinic performance score level:** at different sample thresholds per clinic per calendar year**	FOTO internal analysis (2016-2017)	Patients treated in clinics with 10+ patient per calendar year per clinician for small clinics (up to 4 clinicians) or 40+ patients for large clinics (5 or more clinicians), with FS measures at intake & discharge	TABLE 1.6.III
2b1. VALIDITY TESTING			
2b1i. **Data elements (patient) level:** Content validity (Do test items appear to be measuring the construct of interest?); Structural validity (uni-dimensionality, local independence and item fit); Differential Item Functioning	Wang et al 2015:[1] (2007- 2012)	Patients responding to the full item bank considered for the measure development	TABLE 1.6.I
2b1ii. **Data elements (patient) level:** Known groups construct validity; Sensitivity to change; Functional staging	FOTO internal analysis (2016-2017)	Patients with FS measures at intake & discharge	TABLE 1.6.IV
2b1iii. **Data elements (patient) level:** Clinically important improvement	FOTO internal analysis (2016-2017)	Patients with FS measures at intake & discharge who also complete the patient global rating of change at discharge	TABLE 1.6.V
2b1iv. **Clinician performance score level:** Validity of performance Classification*	FOTO internal analysis (2016-2017)	Patients treated by clinicians with 10+ patients per calendar year with FS measures at intake & discharge	TABLE 1.6.II
2b1v. **Clinic performance score level:** Validity of performance Classification**	FOTO internal analysis (2016-2017)	Patients treated in clinics with 10+ patient per calendar year per clinician for small clinics (up to 4 clinicians) or 40+ patients for large clinics (5 or more clinicians), with FS measures at intake & discharge	TABLE 1.6.III
2b2. EXCLUSIONS ANALYSIS			
2b2. Age exclusion	FOTO internal analysis (2016-2017)	Patients with FS measures at intake & discharge	TABLE 1.6.IV

Analysis	Data source (years)	Sample selection criteria	TABLE number
2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES			
2b3.Risk adjustment model development	FOTO internal analysis (2016)	Patients with FS measures at intake & discharge	TABLE 1.6.VII
2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE			
2b4i. Performance patient level	FOTO risk-adjustment development dataset (2016)	Patients with FS measures at intake & discharge	TABLE 1.6.VII
2b4ii. Performance individual clinician level	FOTO internal analysis (2016-2017)	Patients treated by clinicians with 10+ patients per calendar year with FS measures at intake & discharge	TABLE 1.6.II
2b4iii. Performance clinic/group practice level	FOTO internal analysis (2016-2017)	Patients treated in clinics with 10+ patient per calendar year per clinician for small clinics (up to 4 clinicians) or 40+ patients for large clinics (5 or more clinicians), with FS measures at intake & discharge	TABLE 1.6.III
2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS			
2b6i. Comparing patients with or without complete outcomes	FOTO internal analysis (2016-2017)	Patients with FS measures at intake	TABLE 1.6.VI
2b6ii. Correlations between clinician residuals and their completion rates for clinicians participating in the performance analyses	FOTO internal analysis (2016-2017)	Patients treated by clinicians with 10+ patients per calendar year with FS measures at intake & discharge	TABLE 1.6.II
2b6iii. Correlations between clinic residuals and completion rates for clinics participating in the performance analyses	FOTO internal analysis (2016-2017)	Patients treated in clinics with 10+ patient per calendar year per clinician for small clinics (up to 4 clinicians) or 40+ patients for large clinics (5 or more clinicians), with FS measures at intake & discharge	TABLE 1.6.III
2b6iv. Average residuals at the clinician level by completion rate categories with or without the use of Inverse Probability Weighting	FOTO internal analysis (2016-2017)	Patients treated by clinicians with 10+ patients per calendar year with FS measures at intake & discharge	TABLE 1.6.II
2b6v. Average residuals at the clinic level by completion rate categories with or without the use of Inverse Probability Weighting	FOTO internal analysis (2016-2017)	Patients treated in clinics with 10+ patient per calendar year per clinician for small clinics (up to 4 clinicians) or 40+ patients for large clinics (5 or more clinicians), with FS measures at intake & discharge	TABLE 1.6.III

Abbreviations: *FS = functional status*

TABLE 1.6.I: Patient Characteristics at Intake; original development sample (n = 439 patients)

Characteristic	Total
Age (mean±SD)	48.4±13.8
Sex (female)	59
Acuity of Symptoms	
Acute (0 to 21 d)	19
Sub-acute (22-90 d)	28
Chronic (>90 d)	52
Surgical History	
None	81
1	11
2	5
3 or more	3
Number of Comorbid Conditions	
None	17
1 or 2	24
3 or 4	27
5 or more	32
Exercise History	
At least 3 times/wk	42
1 to 2 times/wk	26
Seldom or Never	31
Missing	1
Payer Source	
Indemnity Insurance	1
Medicaid	3
Medicare A	1
Medicare B	8
Patient	1
HMO	6
PPO	29
Workers' compensation	11
Auto Insurance	1
Other	39

Abbreviations:

HMO=health maintenance organization;

PPO=preferred provider organization.

*Values are percent unless otherwise indicated.

TABLE 1.6.II: Patient Characteristics at Intake; Clinician level (n = 112,178 patients)

Characteristic	Total
Age **(mean±SD)**	**54.3±16.1**
Sex **(female)**	**65.7**
Acuity of Symptoms	
0-7 days	**4.0**
8-14 days	**6.9**
15-21 days	**8.5**
22-90 days	**26.7**
91 days to 6 months	**14.0**
Over 6 months	**39.8**
Surgical History	
None	**87.7**
1	**9.1**
2	**2.1**
3 or more	**1.1**
Number of Comorbid Conditions	
None	**3.7**
1	**6.1**
2	**11.6**
3 or more	**78.6**
Exercise History	
At least 3 times/wk	**38.0**
1 to 2 times/wk	**26.5**
Seldom or Never	**35.5**
Payer Source	
Indemnity Insurance	**2.0**
Medicaid	**5.0**
Medicare A	**1.1**
Medicare B under 65	**3.6**
Medicare B 65 or above	**19.4**
Patient	**0.6**
Workers' compensation	**4.5**
HMO /PPO	**49.1**
No Fault, Auto insurance	**4.5**
Other	**10.3**
Medication use at intake	**50.9**
Previous treatment	**40.5**

Abbreviations:

HMO=health maintenance organization;

PPO=preferred provider organization.

* Values are percent unless otherwise indicated, and sum to 99.9-100.1 due to rounding.

TABLE 1.6.III: Patient Characteristics at Intake; Clinic level (n = 123,194 patients)

Characteristic	Total
Age (mean±SD)	54.4±16.1
Sex (female)	65.7
Acuity of Symptoms	
0-7 days	4.0
8-14 days	6.8
15-21 days	8.5
22-90 days	26.9
91 days to 6 months	14.1
Over 6 months	39.7
Surgical History	
None	87.7
1	9.0
2	2.1
3 or more	1.2
Number of Comorbid Conditions	
None	3.7
1	6.2
2	11.5
3 or more	78.6
Exercise History	
At least 3 times/wk	38.1
1 to 2 times/wk	26.6
Seldom or Never	35.3
Payer Source	
Indemnity Insurance	1.8
Medicaid	5.2
Medicare A	1.1
Medicare B under 65	3.7
Medicare B 65 or above	19.6
Patient	0.6
Workers' compensation	4.2
HMO /PPO	49.4
No Fault, Auto insurance	4.5
Other	10.0
Medication use at intake	50.8
Previous treatment	40.3

Abbreviations:

HMO=health maintenance organization;

PPO=preferred provider organization.

*Values are percent unless otherwise indicated, and sum to 99.9-100.1 due to rounding.

TABLE 1.6.IV: Patient Characteristics at Intake for Episodes with Complete Outcomes (n = 169,039 patients)

Characteristic	Total
Age (mean±SD)	54.6±16.2
Sex (female)	65.4
Acuity of Symptoms	
0-7 days	4.0
8-14 days	6.9
15-21 days	8.6
22-90 days	27.1
91 days to 6 months	14.1
Over 6 months	39.4
Surgical History	
None	87.6
1	9.2
2	2.1
3 or more	1.2
Number of Comorbid Conditions	
None	3.8
1	6.3
2	11.6
3 or more	78.3
Exercise History	
At least 3 times/wk	38.5
1 to 2 times/wk	26.5
Seldom or Never	35.0
Payer Source	
Indemnity Insurance	2.9
Medicaid	4.9
Medicare A	1.3
Medicare B under 65	3.8
Medicare B 65 or above	20.6
Patient	0.6
Workers' compensation	4.6
HMO /PPO	46.3
No Fault, Auto insurance	4.6
Other	10.4
Medication use at intake	50.3
Previous treatment	40.3

Abbreviations:

HMO=health maintenance organization;

PPO=preferred provider organization.

*Values are percent unless otherwise indicated, and sum to 99.9-100.1 due to rounding.

TABLE 1.6.V: Patient Characteristics at Intake for Episodes with Complete Outcomes & global rating of change data (n = 126,026 patients)

Characteristic	Total
Age (mean±SD)	54.6±16.2
Sex (female)	65.4
Acuity of Symptoms	
0-7 days	4.0
8-14 days	6.9
15-21 days	8.5
22-90 days	27.1
91 days to 6 months	14.1
Over 6 months	39.4
Surgical History	
None	87.5
1	9.2
2	2.1
3 or more	1.2
Number of Comorbid Conditions	
None	3.6
1	6.3
2	11.5
3 or more	78.6
Exercise History	
At least 3 times/wk	38.8
1 to 2 times/wk	26.4
Seldom or Never	34.8
Payer Source	
Indemnity Insurance	2.7
Medicaid	5.0
Medicare A	1.3
Medicare B under 65	3.8
Medicare B 65 or above	20.6
Patient	0.6
Workers' compensation	4.5
HMO /PPO	46.5
No Fault, Auto insurance	4.6
Other	10.4
Medication use at intake	50.4
Previous treatment	40.3

Abbreviations:

HMO=health maintenance organization;

PPO=preferred provider organization.

*Values are percent unless otherwise indicated, and sum to 99.9-100.1 due to rounding.

TABLE 1.6.VI: Patient Characteristics at Intake for Episodes with Complete and Incomplete Outcomes (n = 250,741 patients)

Characteristic	Total
Age (mean±SD)	53.6±16.0
Sex (female)	65.3
Acuity of Symptoms	
0-7 days	4.0
8-14 days	6.7
15-21 days	8.4
22-90 days	26.8
91 days to 6 months	14.0
Over 6 months	40.1
Surgical History	
None	87.7
1	9.0
2	2.1
3 or more	1.2
Number of Comorbid Conditions	
None	4.0
1	6.5
2	11.8
3 or more	77.7
Exercise History	
At least 3 times/wk	38.1
1 to 2 times/wk	26.4
Seldom or Never	35.4
Payer Source	
Indemnity Insurance	3.5
Medicaid	5.7
Medicare A	1.2
Medicare B under 65	4.0
Medicare B 65 or above	18.1
Patient	0.7
Workers' compensation	4.2
HMO /PPO	47.6
No Fault, Auto insurance	4.3
Other	10.7
Medication use at intake	50.9
Previous treatment	40.0

Abbreviations:

HMO=health maintenance organization;

PPO=preferred provider organization.

*Values are percent unless otherwise indicated, and sum to 99.9-100.1 due to rounding.

TABLE 1.6.VII: Patient Characteristics at Intake for Episodes with Complete Outcomes for risk-adjustment modeling (n = 77,227 patients)

Characteristic	Total
Age (mean±SD)	55.1±15.8
Sex (female)	65.5
Acuity of Symptoms	
0-7 days	3.9
8-14 days	6.7
15-21 days	8.4
22-90 days	27.3
91 days to 6 months	14.4
Over 6 months	39.4
Surgical History	
None	87.7
1	9.1
2	2.1
3 or more	1.2
Number of Comorbid Conditions	
None	3.2
1	7.2
2	11.4
3 or more	78.3
Exercise History	
At least 3 times/wk	38.4
1 to 2 times/wk	26.3
Seldom or Never	35.3
Payer Source	
Indemnity Insurance	3.0
Medicaid	4.8
Medicare A	1.2
Medicare B under 65	4.1
Medicare B 65 or above	21.2
Patient	0.6
Workers' compensation	4.8
HMO /PPO	46.3
No Fault, Auto insurance	4.6
Other	9.6
Medication use at intake	51.0
Previous treatment	40.5

Abbreviations:

HMO=health maintenance organization;

PPO=preferred provider organization.

*Values are percent unless otherwise indicated, and sum to 99.9-100.1 due to rounding.

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.

As described in this application, the Neck Functional Status (FS) Patient-Reported Outcome Measure (PROM) has undergone extensive testing. Different aspects of testing utilized different years of data and samples as described in TABLE 1.5. The specific data and samples used for each analysis are presented in detail in section 1.6.

1.8 What were the social risk factors that were available and analyzed?

For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.

We posit that the traits of having Medicaid or Medicare B under age 65 (e.g., recipients of disability benefits under Social Security) serve as proxy variables for socioeconomic factors. These variables were accounted for in the risk adjustment model – please see section 2b3.

A standard data point to ask all respondents their level of education was recently added for a limited period of time to the FOTO system for data collection. Because this was a standard question asked of all patients, we acquired a large sample size for this variable and will include in the next round of risk adjustment testing.

____

2a2. RELIABILITY TESTING

Note: 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.

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)
☒ **Performance measure score** (e.g., signal-to-noise analysis)

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)

Methods: Critical data elements used in the measure

2a2i: Internal consistency:

Reliability-based estimates of internal consistency were calculated using data collected at admission from the measure development sample of patients answering all items.[1] Reliability was examined through classical analyses of internal consistency, as well as through item response theory (IRT) person reliability analysis, which is equivalent to the calculation of Cronbach's alpha.[2]

2a2ii: Reliability of point estimates and change scores.

Reliability of individual scores (point estimates) were based on the standard error of measurement (SEM) associated with the final estimate of ability obtained during the computerized adaptive test (CAT) administration. This approach to estimating reliability is more conservative than estimates based on administration of the full bank of items.

The scale-level reliability of the CAT was summarized as:

[1-$SEM^2_{baseline}$/$SD^2_{baseline}$]

where, $SEM^2_{baseline}$ is the median SEM for the Neck FS-CAT and the $SD^2_{baseline}$ is the standard deviation of FS scores at admission.[3] To assess reliability at different levels of scores, we calculated median SEM of individual sores by quartiles of FS estimates at admission. The preferred level of confidence in individual point estimates is a subjective choice that considers the desired probability that the true population mean falls within score intervals drawn from multiple samples.[4,5] Therefore, we used several levels of confidence to calculate confidence intervals (CI), including 68% CI which is equivalent to 1 SEM, and 80% CI, 90% CI and 95% CI. CIs were computed by multiplying the SEM by the corresponding Z-value from the standard normal deviate associated with the desired confidence level. For example, for 95% CI, the SEM was multiplied by 1.96. To test if CIs for point estimates differed at different scale ranges, we calculated CIs for the full range of scores and by quartiles of FS scores at admission.

In addition to the interpretation of a point estimate, clinicians are faced with the need to interpret change in scores during treatment. In most studies, thresholds are estimated for minimally detectable change (MDC), which requires a two-tailed hypothesis test (change for the better and change for the worse).[6-10] However, since the expectation for patients with neck pain conditions is that most patients will get better following treatment, the interpretation of score *improvement* rather than score *change* seems more appropriate. Thus, we calculated one-tailed CIs at 90% and 95% levels of confidence, which are equivalent to 80% and 90% two-sided hypothesis tests, respectively. We refer to the resulting CIs as the minimal detectable improvement (MDI) at different levels of confidence, i.e., MDI_{90}, MDI_{95}, $MDI_{97.5}$ (corresponding to two-tailed MDCs of 80%, 90%, and 95% Cis). Since change involves at least two measure points, a factor of two comes into play, therefore reliability-based estimates of MDI were calculated by multiplying the SEM of the difference ($SEM_{difference}$=SEM * square-root of 2), by the appropriate Z-value.[11] MDIs were calculated for the full range of scores and by quartiles of FS scores at admission.

Methods: Performance measure score (e.g., *signal-to-noise analysis*)

2a2iii-iv: Reliability of providers at the clinician and clinic levels (*signal-to-noise analysis*):

Individual provider reliability was calculated based on Adams' 2009 formula reproduced below.[12] In this calculation, provider-to-provider variance is divided by total variance defined as the sum of provider-to-provider variance plus provider-specific error variance.

$$Reliabilty = \frac{\sigma^2_{provider-to-provider}}{\sigma^2_{provider-to-provider} + \sigma^2_{provider-specific-error}}$$

where provider-specific-error variance is adjusted for the number of patient scores (*'n'* named 'items' in this formula):

$$\sigma^2_{provider-specific-error} = \frac{\sigma^2_{average-item-error}}{n}$$

<u>The variance between all provider groups (signal)</u> was estimated using a mixed-effects hierarchical linear model (HLM) with patients nested within the provider. The dependent variable was functional status change at discharge, adjusting for all variables used by FOTO for risk adjustment (See details in the risk-adjustment section 2b3). The HLM subtracts measurement error variance from overall variance in provider scores to estimate the variance among providers (provider-to-provider variance). The variance component associated with the provider level represents the variance between all provider groups.

<u>The variance within each provider (noise/error)</u> was calculated using the square of the standard deviation of the residual scores, divided by the number of patients (n) for the provider assessed.

We then calculated the average reliability for all providers and the percent of providers passing the recommended 0.7 threshold.[12]

Only providers that passed the threshold for inclusion in the FOTO benchmarking process were included in this calculation (for the clinic level, 10+ patients per clinician per clinic per 12-months period for small clinics, and 40+ patients per clinic per year for larger clinics with 5 or more clinicians. For the clinician level, at least 10 patients per clinician per 12-months period). However, when the average reliability for all providers did not meet the minimum criteria of 0.7, we tested a more conservative threshold by increasing the number of patients per provider until the minimum recommended reliability level was met.

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)

Results: Data elements (patient) level:

2a2i: Internal consistency:

Internal consistency indices as calculated using response from the full item bank were a Cronbach's alpha of 0.98 and an IRT-based person reliability of 0.96.

2a2ii: Reliability of point estimates and change scores:

The scale-level reliability of the CAT scores was 0.91.

Median SEMs of individual scores are reported in TABLE 2a2ii1 both for the entire score range and by quartile. Also included for the entire score range and by quartile is the width of the CI for confidence levels of 80%, 90%, and 95%. Estimates were based on scores obtained at admission. SEMs were stable across the measurement continuum ranging from 3.7 to 3.9 FS points, which corresponds to 7.2 to 7.6 FS points at the 95% confidence level.

TABLE 2a2ii1: Reliability of point estimates* for baseline FS scores

		Width of Confidence Interval		
Baseline FS score	**SEM**	**80%CI**	**90%CI**	**95%CI**
Overall score range	**3.7**	**4.8**	**6.1**	**7.3**
1st quartile (FS 0-43.8)	3.8	4.8	6.2	7.4
2nd quartile (FS>43.8-51.9)	3.7	4.7	6.0	7.2
3rd quartile (FS>51.9-59.3)	3.7	4.7	6.0	7.2
4th quartile (FS>59.3-100)	3.9	4.9	6.4	7.6

Abbreviations: FS, Functional Status; SEM, median standard error from the computerized adaptive test surveys; CI, Confidence Interval

**Confidence in point estimates for the overall score range or by quartiles of functional status scores at admission (n=169,039)*

The MDIs at different levels of confidence (MDI_{90}, MDI_{95}, $MDI_{97.5}$) for the full range of scores and by quartiles at admission are presented in TABLE 2a2ii2. Because the interest was in minimum levels of *improvement*, we used z-scores associated with one-tail of the distribution (positive changes). As an example of how these data should be interpreted, a patient with an admission score of 40 (1st quartile), at the 90% level of confidence, needs to improve by 6.8 FS points to exceed measurement error.

TABLE 2a2ii2: Reliability of improvement scores*

Baseline FS score	**MDI$_{90}$**	**MDI$_{95}$**	**MDI$_{97.5}$**
Overall score range	6.8	8.7	10.4
1st quartile (FS 0-43.8)	6.8	8.8	10.4
2nd quartile (FS>43.8-51.9)	6.6	8.5	10.1
3rd quartile (FS>51.9-59.3)	6.6	8.5	10.2
4th quartile (FS>59.3-100)	7.0	9.0	10.7

Abbreviations: FS, Functional Status; SEM, median standard error from the computerized adaptive test surveys; CI, Confidence Interval; MDI$_{90/95/97.5}$, minimal detectable improvement (one tailed) at 90/95/97.5%CI;

**Confidence in improvement scores for the overall score range or by quartiles of functional status scores at admission (n=169,039)*

Results: Performance score level (*signal-to-noise analysis*)

2a2iii-iv: Reliability of providers at the clinician and clinic levels:

Because the number of providers in the FOTO database is so large, we present reliability statistics by groups of providers based on their number of patients with complete episodes per calendar year, i.e., completed the PRO-PM at admission and discharge (TABLE 2a2iii). Average reliability, as well as minimum and maximum reliability coefficients and the proportion of providers that have reliability coefficients ≥0.7 are shown in the table below.

In summary, the average reliability of clinics meeting the FOTO unique threshold of number of patients per clinic for quality reporting was 0.79. At the clinician level, average reliability for clinicians with 10 or more, or 20 or more patients per calendar year was 0.64 and 0.76, respectively.

TABLE 2a2iii: Reliability (R) at the provider level

Reliability (R) at the provider level: 2016-2017

	Number of patients with complete episodes per clinician per calendar year	**Variance explained (%) by the provider level**	**N providers**	**Average R**	**Min R**	**Max R**	**N if R≥0.7**	**% if R≥0.7**
Clinic	*FOTO	6.4	1378	**0.79**	0.19	0.99	1078	78.2
	20+	6.2	1225	**0.81**	0.26	0.99	1036	84.6
	30+	6.1	1082	**0.83**	0.53	0.99	968	89.5
	40+	5.8	950	**0.84**	0.58	0.99	878	92.4
Clinician	10+	7.7	4537	**0.64**	0.20	0.96	1827	40.3
	20+	8.8	1432	**0.76**	0.35	0.97	1040	72.6
	30+	9.4	456	**0.81**	0.59	0.97	437	95.8
	40+	10.5	141	**0.86**	0.72	0.97	141	100.0

**10+ per clinician for small clinics (1-3 clinicians), 40+ per clinic for large clinics (4 or more clinicians)*

Acceptable levels of reliability are marked in green

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?)

Interpretation: Data elements (patient) level

The results suggest that scores on the Neck FS PROM have excellent internal consistency (>0.95) and CAT-based reliability (0.91). Reliability of point estimates and improvement scores are stable across the measurement continuum.

Interpretation: Performance score level

Based on these findings and using the minimum threshold of a reliability of ≥0.7, we believe that clinic level PRO-PM scores are reliable when used for both small and large clinics using the threshold for inclusion in the FOTO benchmarking process [10+ per clinician for small clinics, 40+ per clinic for large clinics (4 or more clinicians)]. However, findings suggest that the threshold of 10 patients for the clinician level PRO-PM may be insufficient to reliably differentiate between levels of clinicians. Thus, FOTO will establish new thresholds for benchmarking at the clinician level of 20 patients per clinician per calendar year. FOTO will reevaluate reliability periodically, as the database grows, given that this is a new measure and that the calculation of reliability coefficients is influenced by sample size. The variance explained by the provider level from the overall variance in risk-adjusted outcomes in the mixed-effects model is consistent with what we are used to seeing, i.e. values in the range of 5-10%. The fact that the majority of providers had a reliability estimate of 0.7 or more supports an adequate reliability signal, when using the thresholds of number of patients per provider described above.

2b1. VALIDITY TESTING

2b1.1. 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**

☐ **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*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

2b1.2. 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)

Methods: Data elements (patient) level:

2b1i: Content validity (Do test items appear to be measuring the construct of interest?)

The item development process was conducted in 2006 to create a new candidate item pool for individuals with functional problems related to neck impairments. The pool consisted of, 54 items. Items were developed based on review of existing measures in the literature, FOTO's General Orthopedic CAT item bank,[13,14] and input from physical therapists with clinical experiences treating patients with neck-related disorders. Items were worded to represent tasks with different levels of difficulty. To evaluate the clinical relevance of candidate items, an expert panel of 8 physical therapists experienced in treating patients with neck impairments (3 women, 5 men; mean +/- SD years of clinical experience, 16 +/- 9 years) was assembled.

Therapists rated the clinical relevance of the 54 items as (1) highly relevant; (2) partially relevant, beneficial to ask; (3) neutral, not certain; and (4) not relevant at all. Items were considered for inclusion if they were rated as “highly relevant” by at least half the therapists.

2b1i: Structural validity (uni-dimensionality, local independence and item fit)

We assessed responses to the candidate items for uni-dimensionality and local independence, critical assumptions of IRT models. Responses to items of a scale are unidimensional if a single construct (level of the trait being measured) drives how people respond to those items.[2] We conducted exploratory factor analyses of latent trait variables, followed by confirmatory factor analyses on all item responses. Items were considered for removal if factor loadings were below 0.40.[15] Local independence requires that, after taking into account patient ability (in this case, cervical function), item responses are statistically independent of each other. After accounting for the level of the trait being measured, item responses should be uncorrelated. This was tested by evaluating the residual correlation matrix and magnitude of standardized coefficients. Residual correlations greater than 0.20 were flagged for potential problematic local dependency.[16] Model fit was evaluated using the comparative fit index (CFI), Tucker-Lewis index (TLI), and root-mean-square error of approximation (RMSEA). On the CFI and TLI, values greater than 0.90 are indicative of good model fit, and RMSEA values of less than 0.08 suggest adequate fit.[17] We eliminated 1 item from each pair of items with a residual correlation of 0.20 or more. Items that had a higher number of residual correlations with other items were inspected and removed if necessary to improve model fit.

Fit was evaluated using infit and outfit statistics that estimate the ratio of observed variance to expected variance. A recommended criterion for reasonable fit for clinical rating-scale data is infit and outfit values of 1.4 or smaller.[18] Items whose infit and outfit values were greater than 1.4 were dropped.

2b1i: Differential Item Functioning

Differential item functioning (DIF) analyses evaluate whether the difficulty of items is different in different groups (e.g., male versus female). Though different groups may vary in how much they have of the trait being measured, the difficulty of the items should not vary by group membership. That is, when level of neck function is constant, there should be no differences in how subgroups of patients answer particular items. Differential item functioning was evaluated by age (44 years or younger, 45-64 years, and 65 years or older), sex (female and male), and acuity (acute, subacute, and chronic). We compared the item difficulty hierarchy using intraclass correlation coefficients (2-way random model with measures of absolute agreement). We also defined a trivial impact as a difference in item calibrations from the 2 analyses between subgroups of less than 0.5 logits.

2b1ii: Known groups construct validity

We used known group construct validity methods to assess the ability of the CAT generated FS scores to discriminate among groups of patients expected to have different levels of neck function. The independent variables assessed included intake FS, age, gender, symptom acuity, exercise history, surgical history, number of comorbidities, and medication use at admission. We used one-way ANCOVAs with functional status change as the dependent variable, intake FS as the covariate, with one ANCOVA for each risk-adjustment variable as the independent variable. Statistically significant results support the ability of the FS scores to distinguish known groups.

2b1ii: Sensitivity to change

Sensitivity to change was assessed using a distribution-based approach. Effect size statistics were estimated as follows: (discharge FS minus intake FS)/(intake FS standard deviation).

2b1ii: Functional staging

Meaningful clinical interpretation of the FS scores supports their validity. Functional staging is used to describe clinical meaningfulness of the quantitative scores provided by a measure. We developed a functional staging model using methods described previously.[7-10,19] Score-based functional abilities are described for patients at different score levels.[19] We graphically displayed the most likely responses for each item across all measured levels of function. We reviewed the output and reached consensus on expected performance of patients at 5 hierarchical stages of neck function and on the 4 cut scores that defined the 5 stages. Based on the agreed upon structure, we constructed a staging chart that portrays expected responses to each item at each functional stage. The Neck-CAT PRO-PM items are listed in a descending order of difficulty, along with a short item description, in the APPENDIX.

To further assess the functional staging model's responsiveness, we tested the rates of functional staging changes during treatment. Large rates of change during treatment would support the model's responsiveness.

2b1iii: Clinically important improvement

Clinically important improvement was assessed using an anchor-based approach by calculating the proportion of patients whose FS change was greater than minimal clinically important improvement (MCII), which is improvement considered important to the patient. To incorporate the patient's perspective on the clinical importance of FS score change, we used a global rating of change (GROC) as the external anchor.[20] The GROC used includes one question with a 15-point scale for the degree of change (-7 to +7), with zero representing no change. Data from patients who completed both the FS and the GROC at discharge were used for this analysis. We assessed meaningful change thresholds of MCII by dichotomizing patients into those that improved (GROC ≥ 3) or did not improve (GROC < 3). We chose a threshold of 3 or more (3= "somewhat better") because previous studies showed that this cut-score provided adequate assessment of important improvement.[21-23] Because of the large body of evidence that MCII levels are dependent on baseline FS,[6-10,21-31] we also estimated MCII by quartiles of baseline FS. Using receiver operating characteristic (ROC) analyses, MCII cut points were identified by selecting the FS change score with the largest average specificity and sensitivity values. Percent of improved patients, MCIIs and their 95% CI, areas under the receiver operator curve (AUC) and their 95% CI, and percentage of patients whose FS change was equal to or greater than MCII.

Methods: Performance score level

2b1iv-v: Clinician & Clinic Performance Score Level:

Provider (clinic and clinicians) performance as determined by the average residual was validated against an external marker using each provider's rate of patients achieving at least the minimal clinically important improvement (MCII). MCII was calculated as described above.

We used two methods for categorizing provider's into performance levels. First, providers were categorized into 3 quality levels (low, average, high) based on uncertainty assessments. This method allows to establish statistically significant differences between performance levels. Second providers were categorized into 10 quality levels based on percentile ranking that allows to create evenly distributed performance groups, although they may not represent statistically distinct quality levels. Additionally, percentile ranking represents a categorization that is easy to interpret and meaningful to clinicians, managers, and payers.

Performance based on uncertainty assessments:

We calculated patient level residual scores (residual = actual change – predicted change) after risk adjustment modeling and aggregated scores by individual clinician or clinic. At the clinic level, performance was evaluated only for large clinics (4 or more clinicians) that had a minimum of 40 patients per calendar year, and small clinics (1-3 clinicians) that had a minimum of 10 patients per clinician per calendar year. At the individual clinician level, performance was evaluated only for clinicians that had a minimum of 10 patients per calendar year. To examine statistical differences between entities (individual clinics or clinicians) performance scores, we plotted each entity's average aggregated patient residual scores (with their 95% confidence intervals) to examine whether or not there were statistically significant differences between clinics/clinicians, or between each clinic/clinician and the national average. Since the mean residual score is hypothetically centered at zero, each entity can be compared to that standard which is the predicted clinic aggregated outcome. When the

95% CI for a clinic/clinician crosses zero, the performance for that year is determined to be no different (statistically) than the predicted national average. If 95% CIs are below or above zero, the performance for that year is determined to be worse or better than the predicted national average, respectively. Thus, provider performance scores with 95% CIs were classified into three groups: low performance (clinics with 95% CI of residual scores below 0), average performance (clinics with 95% CI of residual scores crossing 0), and high performance (clinics with 95% CI of residual scores above 0).

Performance based on percentile ranking:

Providers were divided into 10 performance groups by deciles of their average residuals.

For both methods described above, A one-way ANOVA was conducted to determine if the rate of MCII achievement at the provider level was different by the clinic's assigned performance group as expected, i.e., higher rates of MCII achievement for higher performance.

2b1.3. What were the statistical results from validity testing?

(e.g., correlation; t-test)

Results: Validity of Data elements (patient) level:

2b1i: Content validity (Do test items appear to be measuring the construct of interest?)

After the expert panel consultation, 19 items were removed, and the remaining 35 items were evaluated in subsequent analyses. Neck function items were presented by asking the patient, “Because of your neck, how much difficulty do you have.…” This was followed by activities such as “turning to look behind you” or “placing a 25-lb box on a shelf overhead.” Five rating-scale response categories were used: (1) extreme difficulty or unable to perform, (2) quite a bit of difficulty, (3) moderate difficulty, (4) a little bit of difficulty, and (5) no difficulty.

2b1i: Structural validity (uni-dimensionality, local independence and item fit)

Of the 35 remaining candidate items, 3 items (sleeping more than 4 hours, sleeping more than 6 hours, and rolling over in bed) were removed because of their high negative residual correlations with several other items. The exploratory factor analyses supported a 1-factor structure of the remaining 32 items. The first factor accounted for 65% of the total variance; all items had factor loadings above 0.40. The CFI and TLI fit statistics were 0.84 and 0.98, respectively. RMSEA was 0.16. Four items (sleeping more than 1 hour, sleeping through the night, lying flat on your back for 30 minutes, running a block) were removed due to high infit and outfit statistics (fit statistics greater than 1.4). The remaining 28 items were accepted and comprised the final item bank for the CAT.

2b1i: Differential Item Functioning

Item difficulty hierarchical structures were highly consistent across age group, sex, and acuity (all intraclass correlation coefficients greater than 0.94). Out of 196 item-to-item comparisons of item difficulty parameters across age group, sex, and acuity, 190 (96.9%) comparisons had a

difference in logits of less than 0.50, indicating negligible DIF.

2b1ii: Known groups construct validity

Results supported known group construct validity of the FS measures estimated and discriminated groups of patients with expected patterns (TABLE 2b1iia). Statistically better FS change was obtained for patients who were younger, male, had more acute symptoms, exercised before initiating the episode of care, had no or fewer surgeries related to their neck pain, had no or fewer comorbid conditions, and did not use medications for the neck pain at the start of the episode compared their reference groups.

TABLE 2b1iia: Known-Groups Construct Validity (n=169,039)

Patient characteristic		Model (ANCOVA)			Marginal means*		
Variable	**Groups**	**N**	**%**	**Prob>F**	**b**	**95% CI**	
Age	14 to <18	1,988	1.2%	P<0.001	**18.5**	17.9	19.1
	18 to <45	42,755	25.3%		**14.7**	14.6	14.9
	45 to <65	71,820	42.5%		**11.3**	11.2	11.4
	65 to <85	50,115	29.6%		**11.4**	11.2	11.5
	85 or more	2,361	1.4%		**9.3**	8.8	9.9
Gender	(1)Male	58,530	34.6%	P<0.001	**13.0**	12.9	13.1
	(2)Female	110,509	65.4%		**11.8**	11.8	11.9
Acuity	0-7 days	6,763	4.0%	P<0.001	**20.7**	20.4	21.0
	8-14 days	11,612	6.9%		**17.6**	17.4	17.8
	15-21 days	14,511	8.6%		**15.8**	15.6	16.0
	22-90 days	45,844	27.1%		**13.5**	13.4	13.6
	91 days to 6 months	23,772	14.1%		**11.3**	11.1	11.4
	Over 6 months	66,537	39.4%		**9.1**	9.0	9.2
Exercise history	At least 3x/week	65,145	38.5%	P<0.001	**12.7**	12.6	12.8
	1-2x/week	44,751	26.5%		**12.5**	12.4	12.6
	Seldom or Never	59,143	35.0%		**11.5**	11.4	11.6
Surgical history	No related surgeries	148,026	87.6%	P<0.001	**12.8**	12.8	12.9
	1 related surgery	15,484	9.2%		**8.6**	8.4	8.8
	2 or more related surgeries	5,529	3.3%		**6.1**	5.8	6.4
Number of comorbidities	0	6,410	3.8%	P<0.001	**16.3**	16.0	16.6
	1	10,661	6.3%		**15.7**	15.5	16.0
	2	19,558	11.6%		**14.8**	14.6	14.9
	3 or more comorbidities	132,410	78.3%		**11.4**	11.3	11.4
Medication use at intake	No(0)	84,089	49.7%	P<0.001	**12.6**	12.5	12.7
	Yes(1)	84,950	50.3%		**11.9**	11.8	11.9
***Marginal means at mean FS at intake of 51.9**							

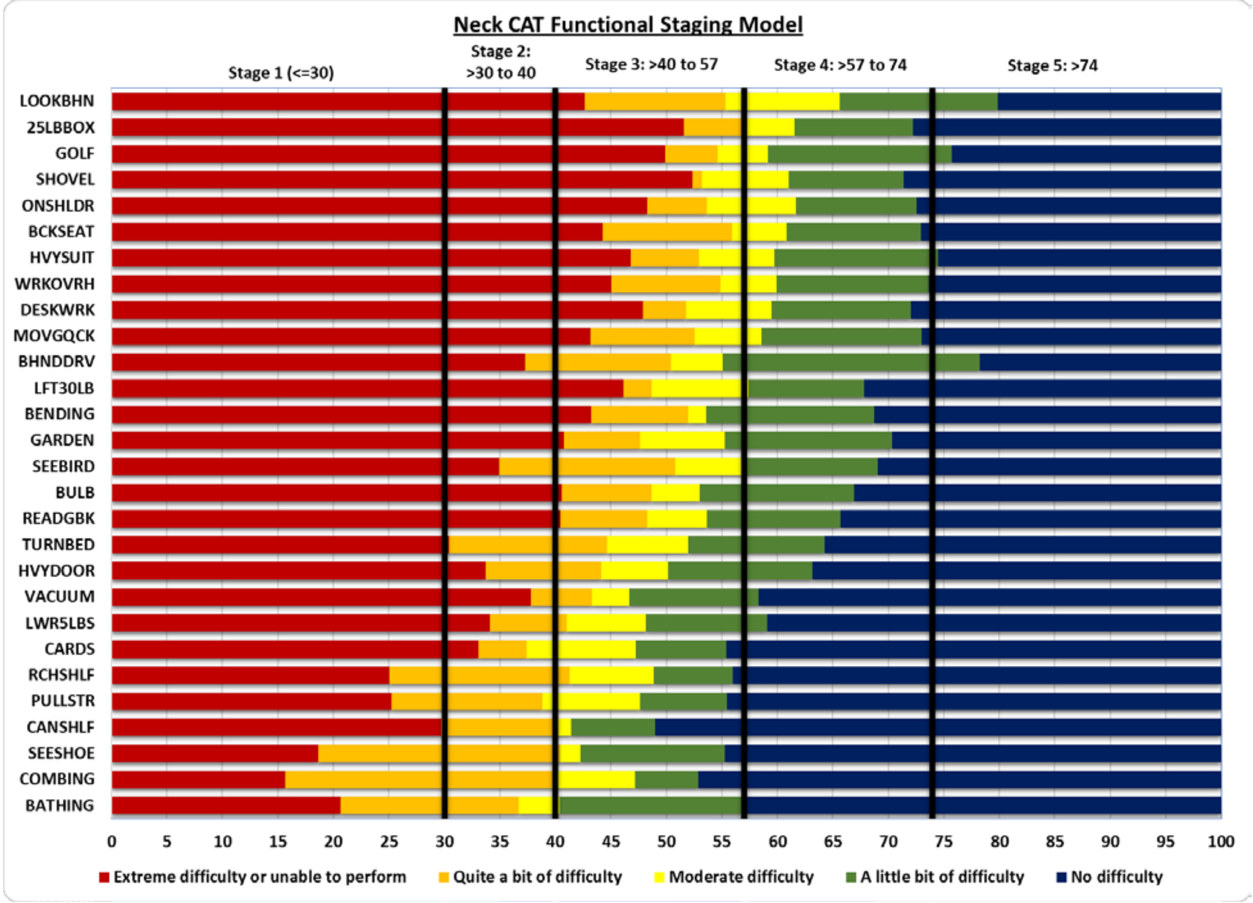
2b1ii: Sensitivity to change

Results supported the sensitivity of FS scores to change; the mean of intake FS scores was 51.9 (SD 12.3) and 64.1 (SD 14.8) at discharge. Mean FS change scores was 12.2 (SD 13.8), which produces an effect size of 0.99, which is considered large.

2b1ii: Functional staging

The staging chart shows expected responses to each item at each functional stage (FIGURE 2b1ii) The functional staging model's operational definitions are presented in TABLE 2b1iia. Percentages of functional staging change from admission to discharge is presented in TABLE 2b1iib, demonstrating large rates of functional staging change during treatment, with 61% of patients demonstrating a functional staging change (56% improved and 5% worsened).

FIGURE 2b1ii: Functional Staging chart



Accompanying table for Figure 2b1ii

(Range of functional status scores for each response category. Row data adds up to the maximum score of 100):

	Extreme difficulty or unable to perform	Quite a bit of difficulty	Moderate difficulty	A little bit of difficulty	No difficulty
LOOKBHN	42.64	12.66	10.32	14.24	20.14
25LBBOX	51.6	5.2	4.73	10.71	27.76
GOLF	49.88	4.75	4.53	16.58	24.26
SHOVEL	52.35	0.87	7.8	10.34	28.64
ONSHLDR	48.26	5.41	8.02	10.84	27.47
BCKSEAT	44.24	11.66	4.9	12.14	27.06
HVYSUIT	46.76	6.15	6.8	14.8	25.49
WRKOVRH	45.05	9.8	5.05	13.75	26.35
DESKWRK	47.87	3.9	7.69	12.55	27.99
MOVGQCK	43.15	9.39	6.03	14.4	27.03
BHNDDRV	37.29	13.15	4.65	23.14	21.77
LFT30LB	46.14	2.55	8.69	10.44	32.18
BENDING	43.26	8.71	1.61	15.14	31.28
GARDEN	40.75	6.87	7.67	15.05	29.66
SEEBIRD	34.96	15.84	6.07	12.19	30.94
BULB	40.57	8.1	4.34	13.9	33.09

	Extreme difficulty or unable to perform	Quite a bit of difficulty	Moderate difficulty	A little bit of difficulty	No difficulty
READGBK	40.43	7.84	5.37	12.05	34.31
TURNBED	30.43	14.21	7.3	12.31	35.75
HVYDOOR	33.71	10.41	6.02	12.99	36.87
VACUUM	37.77	5.51	3.36	11.67	41.69
LWR5LBS	34.12	6.92	7.12	10.91	40.93
CARDS	33.08	4.31	9.86	8.13	44.62
RCHSHLF	25.06	16.25	7.56	7.14	43.99
PULLSTR	25.25	13.56	8.82	7.85	44.52
CANSHLF	29.7	10.23	1.48	7.57	51.02
SEESHOE	18.65	21.79	1.82	13.01	44.73
COMBING	15.65	24.46	7.05	5.73	47.11
BATHING	20.67	16.02	3.71	16.5	43.1

TABLE 2b1iia: Functional staging model operational definitions

Stage # (score range)	Title	Operational Definition
Stage 1 (0-30)	Limited Self-Care	Exceedingly limited in neck motion, basic self-care tasks, or reaching.
Stage 2 (>30 to 40)	Light Activity	Able to perform neck motion, basic self-care tasks, or reaching with difficulty.
Stage 3 (>40 to 57)	Moderate activity	Able to move light to medium weight objects, perform neck motions or move in bed with minimal to moderate difficulty. Able to perform basic self-care tasks with minimal to no difficulty.
Stage 4 (>57 to 74)	High activity	Able to perform high-level activities with minimal to moderate difficulty or neck motions with minimal to no difficulty.
Stage 5 (>74-100)	Vigorous activity	Able to perform vigorous work/occupation, sports, recreation, heavy household tasks/yard work, handling heavy objects overhead with minimal to no difficulty.

TABLE 2b1iib Functional staging change from admission to discharge*

		Functional stage at discharge					
		1	2	3	4	5	Total
Functional stage at admission	1	0.4	0.7	1.2	0.9	0.3	**3.5**
	2	0.3	1.6	5.4	4.1	1.0	**12.4**
	3	0.1	1.3	16.2	24.0	7.2	**48.9**
	4	0.0	0.1	2.8	18.4	11.0	**32.2**
	5	0.0	0.0	0.0	0.7	2.2	**3.0**
	Total	**0.8**	**3.7**	**25.7**	**48.1**	**21.7**	**100.0**

** Rates (%) of functional staging change from admission to discharge (n=169,039), with 61% of patients demonstrating a functional staging change (56% improved and 5% worsened).*

2b1iii: Clinically important improvement

Spearman's rank correlation between FS scores and GROC ratings was 0.52, which is considered adequate for acceptance of MCII (> 0.3). SEM, percent of improved patients (GROC ≥ 3), MCII estimates, AUC, and percentage of patients whose FS change was equal to or greater than MCII are presented in TABLE 2b1iii for the overall score and by baseline FS. MCII estimates ranged from 15 to 4 FS points from 1st to 4th quartile of baseline FS scores, respectively. Thus, more FS change points were needed to achieve minimal clinically important improvement for patients with lower baseline FS, supporting previous results described above.

TABLE 2b1iii: Anchor-based estimate of minimal clinically important improvement*

Baseline FS score	SEM	% improved (GROC≥3)	MCII / ROC cut point	MCII 95%CI	AUC	AUC 95%CI	% ≥ MCII
Overall score range	3.7	82.4%	8.1	7.1, 9.0	.75	.74, .75	57.5%
1st quartile (FS 0-43.8)	3.8	76.7%	15.2	13.4, 17.0	.79	.78, .80	55.7%
2nd quartile (FS>43.8-51.9)	3.7	82.5%	10.2	9.8, 10.6	.80	.79, .81	56.1%
3rd quartile (FS>51.9-59.3)	3.7	85.0%	7.1	6.0, 8.3	.79	.78, .80	54.8%
4th quartile (FS>59.3-100)	3.9	85.7%	3.7	2.6, 4.7	.75	.74, .76	56.8%

Abbreviations: FS, Functional Status; SEM, median standard error from the computerized adaptive test surveys; CI, Confidence Interval; MCII, minimal clinically important improvement; ROC, receiver operating characteristic analysis; AUC, area under the ROC curve

** Estimate of minimal clinically important improvement based on a global rating of change cut score of 3 or more (n=126,026)*

Results: Validity of Performance score level

2b1iv: Clinician Performance Score Level:

A higher proportion of patient episodes managed by higher performing providers experienced change equal to or greater than the MCII as compared to lower performing providers. This pattern was observed using both methods of provider performance ranking; uncertainty assessments (3 levels) and percentile ranking (10 levels).

Method 1: Validity of clinician performance based on uncertainty assessments (3 levels):

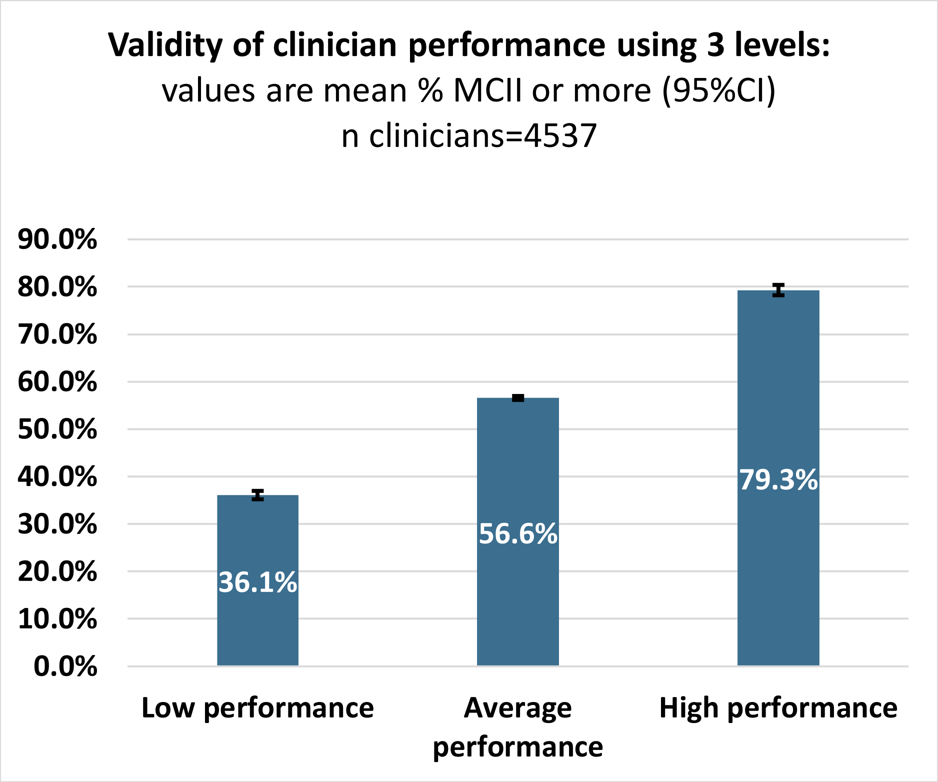
The three performance levels had statistically significant differences between groups as determined by one-way ANOVA (F(2,4534) = 1852.1, p < 0.001) with a monotonic increase in rates of MCII achievement (TABLE 2b1iv-a).

TABLE 2b1iv-a: Validity of performance at the clinician level

Performance level	N Clinicians (%)	% MCII or more
Low	735 (16.2)	36.1%
Average	3321 (73.2)	56.6%
High	481(10.6)	79.3%

A Tukey post-hoc test revealed that all groups were significantly different from one another (p<0.001) (FIGURE 2b1iv-a).

FIGURE 2b1iv-a: Validity of performance at the clinician level



Method 2: Validity of clinician performance based on percentile ranking (10 levels):

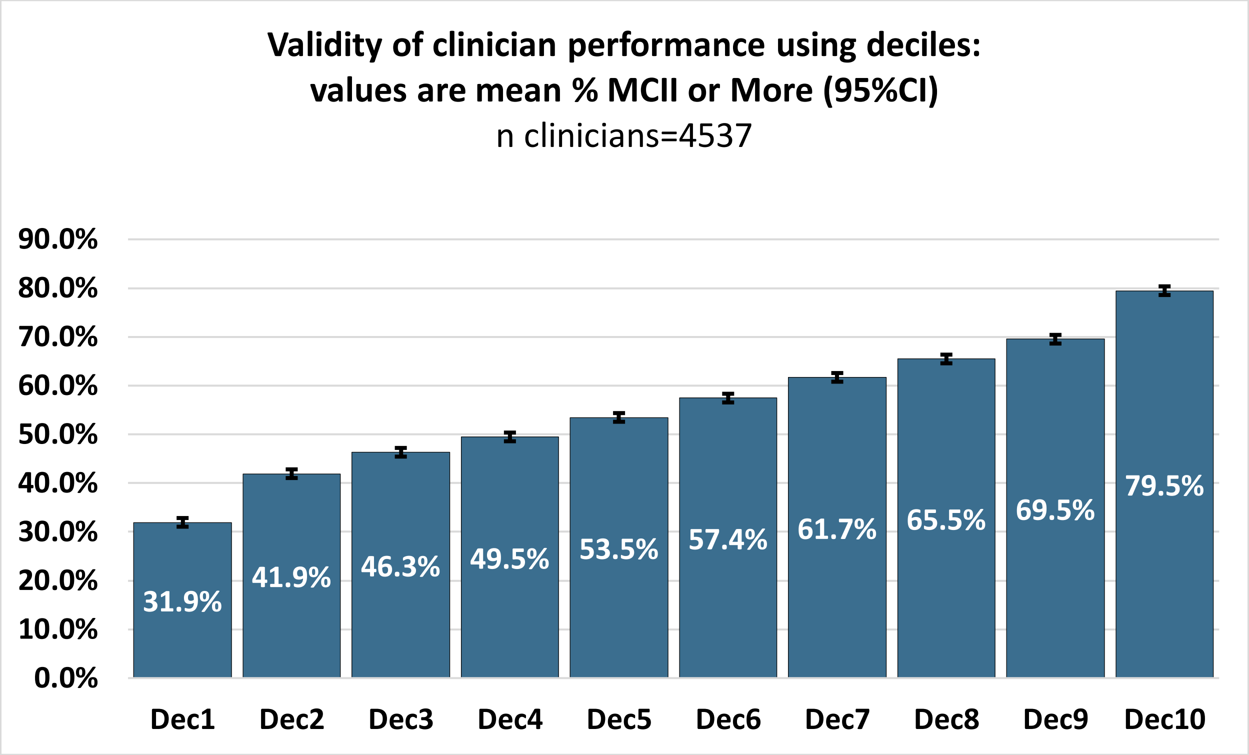
The ten performance levels had statistically significant differences between groups as determined by one-way ANOVA (F(9,4527) = 955.1, p < 0.001), with a monotonic increase in rates of MCII achievement (TABLE 2b1iv-b).

TABLE 2b1iv-b: Validity of performance at the clinician level

Performance level	N Clinicians (%)	% MCII or more
Decile 1	454	31.9%
Decile 2	454	41.9%
Decile 3	454	46.3%
Decile 4	453	49.5%
Decile 5	454	53.5%
Decile 6	454	57.4%
Decile 7	453	61.7%
Decile 8	454	65.5%
Decile 9	454	69.5%
Decile 10	453	79.5%

A Tukey post-hoc test revealed that all groups were significantly different from one another (p <0.001) (FIGURE 2b1iv-b).

FIGURE 2b1iv-b: Validity of performance at the clinician level



2b1v: Clinic Performance Score Level:

A higher proportion of patient episodes managed by the higher performing providers experienced change equal to or greater than the MCII as compared to lower performing providers. This pattern was observed using both methods of provider performance ranking; uncertainty assessments (3 levels) and percentile ranking (10 levels).

Method 1: Validity of clinic performance based on uncertainty assessments (3 levels):

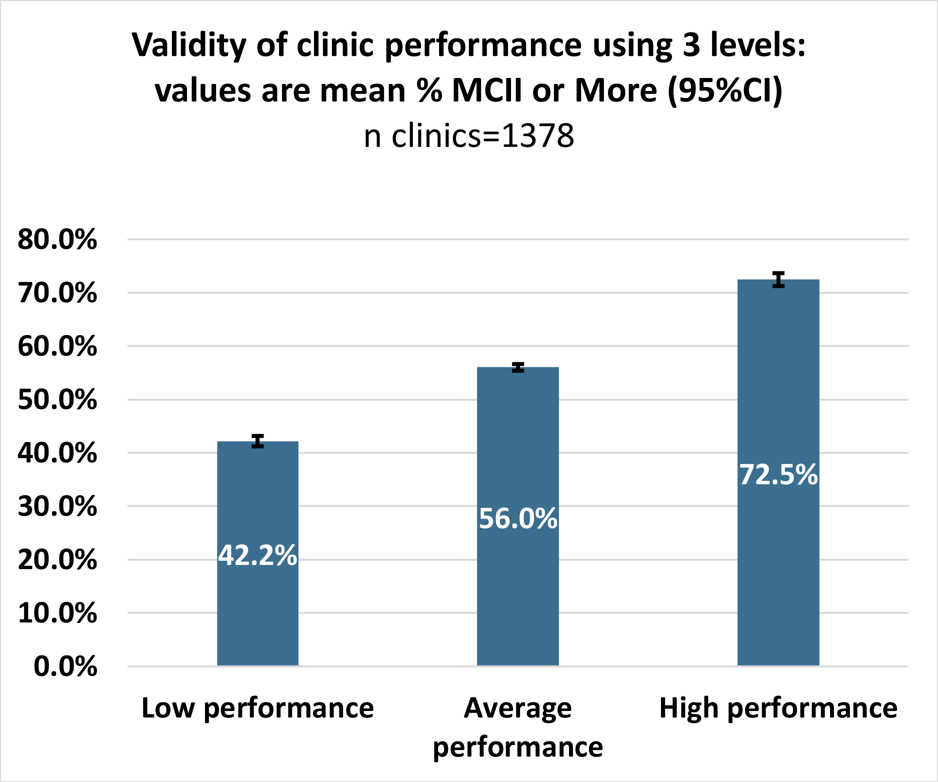
The three performance levels had statistically significant differences between groups as determined by one-way ANOVA (F(2,1375) = 738.2, p < 0.001) with a monotonic increase in rates of MCII achievement (TABLE 2b1v-a).

TABLE 2b1v-a: Performance at the clinic level

Performance level	N Clinics (%)	% MCII or more
Low	334 (24.2)	42.2%
Average	827(60.0)	56.0%
High	217(15.7)	72.5%

A Tukey post-hoc test revealed that all groups were significantly different from one another (p<0.001) (FIGURE 2b1v-a).

FIGURE 2b1v-a: Validity of performance at the clinic level



Method 2: Validity of clinic performance based on percentile ranking (10 levels):

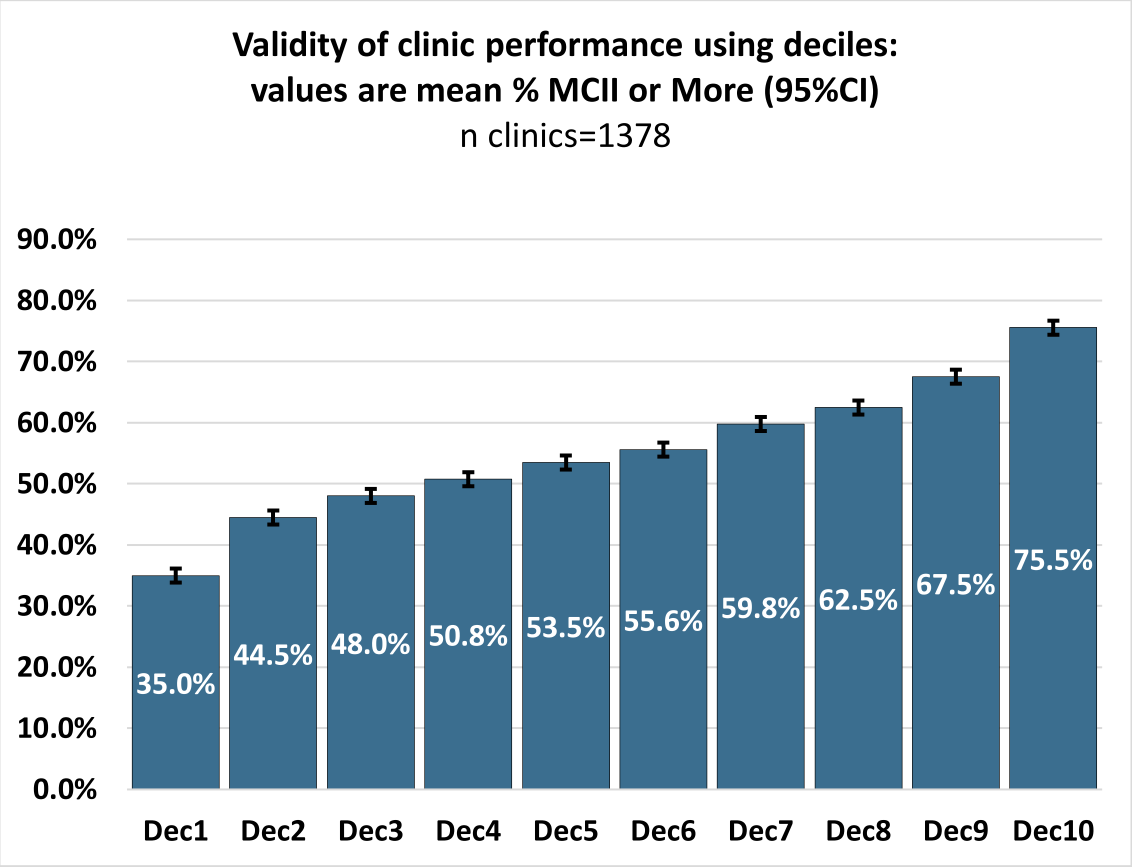
The ten performance levels had statistically significant differences between groups as determined by one-way ANOVA (F(9,1368) = 400.9, p < 0.001), with a monotonic increase in rates of MCII achievement (TABLE 2b1v-b).

TABLE 2b1v-b: Validity of performance at the clinic level

Performance level	N Clinics (%)	% MCII or more
Decile 1	138	35.0%
Decile 2	138	44.5%
Decile 3	138	48.0%
Decile 4	138	50.8%
Decile 5	137	53.5%
Decile 6	138	55.6%
Decile 7	138	59.8%
Decile 8	138	62.5%
Decile 9	138	67.5%
Decile 10	137	75.5%

A Tukey post-hoc test revealed that all groups were significantly different from one another (p=0.039 to <0.001), except for decile 5 vs decile 6 (P=0.243) (FIGURE 2b1v-b).

FIGURE 2b1v-b: Validity of performance at the clinic level



2b1.4. 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?)

Interpretation: Validity of data elements (patient) level:

Results produced multiple levels of validity evidence including content and structural validity. Items were supported by the expert panel; retained items demonstrated essential unidimensionality, local independence and item fit. Known group construct validity of the FS scores was supported with FS scores discriminating groups of patients in clinically known and logical ways. Strong evidence for the sensitivity to change and responsiveness was obtained, with a majority of patients achieving a minimal clinically important improvement. The functional staging model improves score interpretation and demonstrates responsiveness to FS change during treatment.

Interpretation: Validity of Clinician & Clinic Performance Score:

Validity of performance levels identified using either 3 levels based on uncertainty assessments, or 10 levels based on deciles of average residuals, was supported by demonstrating increased rates of patients achieving the MCII at higher performance levels. This pattern was observed both in the clinician and clinic levels. Additionally, rates of MCII increases monotonically between consecutive performance levels, supporting clinically logical expectations.

Overall, this supports the validity of provider performance measures based on the neck-CAT PRO-PM risk-adjusted residuals, at both the clinician and clinic levels.

2b2. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section* 2b4

2b2.1. 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)

Age exclusion: The FS measures were designed and tested for adult patients aged 14 years or older. However, the risk-adjustment (RA) model was developed using data from patients aged 18 or above. This raised the question of whether residuals derived from the current RA model for patients aged 14 to 17 would differ from those derived from a model specific to this younger age range. Therefore, first, we calculated residuals for patients aged 14 to 17 using the current FOTO RA model (Model 1). Second, we calculated for the same patient group a separate set of residuals from a model adapted to this patient population (Model 2), using a backwards stepwise regression that allowed only significant variables to enter the model (P-entry=0.05, P-removal=0.1).), as done for the original RA modeling.[32] Finally, we conducted a sensitivity analysis by comparing these two sets of residuals. Comparisons were done using a pairwise Pearson correlation (r), and an interclass correlation coefficient (ICC(2,1)) to confirm that a high correlation would not result from a correlation with a constant offset. A high correlation between the two sets of residuals would support the validity of the current FOTO risk-adjustment model for the neck PRO-PM for patients aged 14 to 17.

2b2.2. 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)

The correlation between the two sets of residuals, those derived from the current FOTO RA model (model 1) and those derived from the model adapted to patients aged 14-17 (model 2) was 0.98 (P<0.001), with an ICC(2,1) of 0.97 (P<0.001). FIGURE 2b2.2 plots the association between these two sets of residuals.

FIGURE 2b2.2: Age exclusion sensitivity analysis

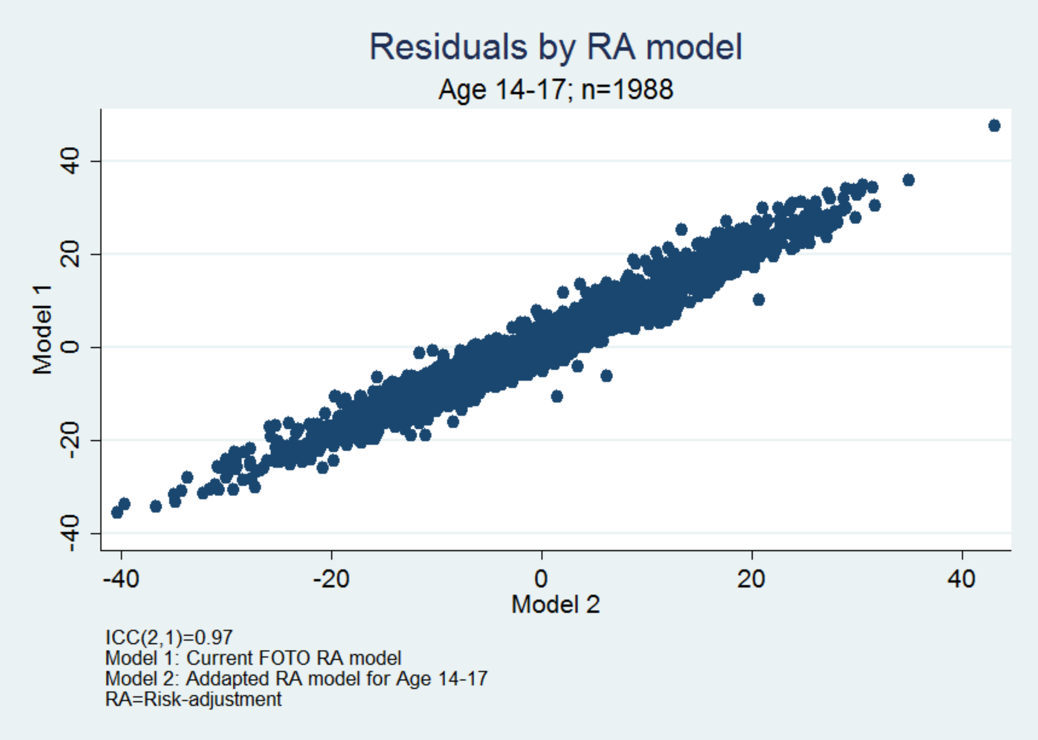


TABLE 2b2.2 compares the coefficients from model 1 & 2. As described above, only significant coefficients were allowed to enter model 2.

TABLE 2b2.2: Risk-adjusted models for calculating residuals used to test exclusion criteria for age

Dependent Variable: FS at discharge	Model 1: All ages (14-89)	Model 2: Age 14-17
N	169,039	1,988
Adjusted R-squared	32.9%	23.4%

Independent variables	Beta	95% CI		Beta	95% CI	
		Lower	Upper		Lower	Upper
Constant	46.6	46.1	47.1	51.1	47.4	54.7
Intake FS	0.5	0.5	0.5	0.4	0.4	0.5
Age (continuous)	-0.1	-0.1	-0.1			
Gender: Female	-1.1	-1.3	-1.0	-2.6	-3.8	-1.3
Acuity						
0-7 days	9.3	9.0	9.6	13.6	11.3	15.9
8-14 days	6.7	6.4	6.9	8.5	6.3	10.7
15-21 days	5.1	4.9	5.3	5.7	3.7	7.6
22-90 days	3.2	3.1	3.4	3.7	2.2	5.3
91 days to 6 months	1.3	1.1	1.5	1.4	-0.5	3.3
Over 6 months (Ref)						
Payer						
Indemnity	-2.6	-3.0	-2.3	-4.4	-7.1	-1.6
Medicaid	-3.4	-3.7	-3.1			
Medicare B, age 65 or above	0.8	0.6	1.0			
Medicare B, under age 65	-2.4	-2.7	-2.1			
No fault, Auto	-2.5	-2.8	-2.2	-2.2	-4.6	0.2
Workers' compensation	-5.4	-5.7	-5.1			
HMO, preferred provider (Ref)						
Surgical history						
1 related surgeries	-2.5	-2.7	-2.2			
2 related surgeries	-3.8	-4.2	-3.4			
3 or more related surgeries	-4.3	-4.8	-3.7			
No related surgery (Ref)						
Exercise history						
At least 3x/week	0.6	0.4	0.7			
1-2x/week	0.4	0.2	0.5			
Seldom or Never (Ref)						
Medication use at intake	-0.9	-1.0	-0.8			
Previous treatment	-1.7	-1.8	-1.6	-3.2	-4.5	-1.9
Neck surgical code: Fusion	1.1	0.7	1.6	-12.7	-22.2	-3.2

Independent variables	Beta	95% CI		Beta	95% CI	
		Lower	Upper		Lower	Upper
Comorbidities						
Anxiety	**-0.8**	-1.0	-0.6			
Arthritis	**-0.8**	-1.0	-0.7			
Back pain	**-0.9**	-1.1	-0.8	**-1.6**	-2.8	-0.3
Depression	**-0.8**	-1.0	-0.6	**-1.6**	-3.3	0.1
Gastro	**-0.4**	-0.6	-0.3			
Headaches	**-1.0**	-1.1	-0.9			
Kidney, Bladder, Prostate or Urination	**-0.6**	-0.8	-0.4			
Obesity (BMI>=30)	**0.5**	0.4	0.6			
Osteoporosis	**-0.5**	-0.7	-0.3			
Previous accidents	**-0.8**	-0.9	-0.6			
Sleep dysfunction	**-1.2**	-1.4	-1.1	**-2.4**	-4.3	-0.6

Beta coefficient indicating the amount of expected change in discharge FS given a 1-unit change in the value of the variable, given that all other variables in the model are held constant.
Abbreviations: BMI, body mass index (kg/m^2); FS, functional status; HMO, health maintenance organization.
Ref, Reference group

2b2.3. 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)

The extremely high correlation between the two sets of residuals (ICC(2,1) of 0.97) suggests no practical impact of the model selected on performance score level results for the younger age group of 14-17. Additionally, the comparison of the two models used to calculate the two sets of residuals shows that, except for neck fusion post-surgical status, all other significant coefficients had similar trends and direction. Variables not significant in the younger age group seemed clinically logical given this young and small age range (e.g., age, older population payer categories, number of related surgeries, comorbidities). Overall, we interpret these results as supporting the validity of the current FOTO risk-adjustment model for the neck PRO-PM for patients aged 14 to 17.

2b3. 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.

2b3.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 11 risk factors**

☐ **Stratification by _risk categories**

☐ **Other,**

2b3.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.

The methods used to develop the FOTO risk-adjustment neck model were the same as the methods described in detail in a recent publication by Deutscher et at, 2018.[32] Briefly, we used data from adult patients with neck pain treated in outpatient physical therapy clinics during 2016, that had complete outcomes data at admission and discharge to develop the risk-adjustment model. The data included the following patient factors that could be evaluated for inclusion in a model for risk-adjustment: FS at admission (continuous); age (continuous); sex (male/female); acuity as number of days from onset of the treated condition (6 categories); type of payer (10 categories); number of related surgeries (4 categories); exercise history (3 categories); use of medication at intake for the treatment of LBP (yes/no); previous treatment for LBP (yes/no); treatment post-surgery (lumbar fusion, laminectomy or other); and 31 comorbidities.

The risk-adjustment model was constructed and assessed for predictive validity in several steps. We used a backward stepwise linear ordinary-least-square (OLS) regression to identify patient factors that significantly contributed to the prediction of FS outcomes at discharge. The backward stepwise procedure allows variables to be removed and entered in a sequential manner to create the most parsimonious final model. Variables were entered if significance of their T value was less than 0.05 (entry level) and removed if significance was greater than 0.1 (removal level). Categorical variables were tested in comparison to a reference category represented by the largest category for nominal data, e.g., payer categories, or the largest of the extreme (minimal or maximal) category for ordinal variables, e.g., acuity. Multiple regression models in general, and stepwise procedures specifically, have a risk of over-interpretation based on the particular characteristics of the sample at hand, a phenomenon known as overfitting.[33] Because of the large sample size examined and the generous ratio of cases per number of predictors tested, we expected the risk of overfitting to be minimal, even when adopting strict criteria for the ratio between sample size and number of predictors.[34] Nonetheless, assessing for model overfitting, i.e., yielding findings that will not replicate in a different sample, is necessary (see section 2b3.5 below for the additional risk-adjustment model development steps).

2b3.2. If an outcome or resource use component 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.

NA

2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk 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) **Also discuss any "ordering" of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

The methods used to develop the FOTO risk-adjustment models are described in detail in a recent publication by Deutscher et at, 2018.[32]

Patient factors

We selected and examined the patient factors available to us and known to be associated with FS outcomes to establish an optimal risk adjustment model for our data set.[13,35,36] We selected non-modifiable patient factors to avoid misclassification of provider performance and control for their relationships with outcomes of interest.

Social factors

We posit that the traits of having Medicaid or Medicare B under age 65 (e.g., recipients of disability benefits under Social Security) serve as proxy variables for socioeconomic factors. These variables were accounted for in the risk adjustment model.

A standard data point to ask all respondents their level of education was recently added for a limited period of time to the FOTO system. Because this was a standard question asked of all patients, we acquired a large sample size for this variable and will include in the next round of risk adjustment testing.

2b3.3b. How was the conceptual model of how social risk impacts this outcome developed?

Please check all that apply:

☒ **Published literature**

☒ **Internal data analysis**

☐ **Other (please describe)**

2b3.4a. What were the statistical results of the analyses used to select risk factors?

The adjusted R-squared was 33.3%.

TABLE 2b3.4a: Risk-adjusted model:

Associations between patient characteristics at admission and FS at discharge.

Significant Predictors of FS at Discharge (Reference group for categorical variables)	**B**	**T**	**P-value**
(Constant)	45.5	122.91	<0.001
Intake FS	0.5	132.84	<0.001
Age	-0.1	-29.26	<0.001
Female	-1.2	-12.09	<0.001
Acuity:			
0-7 days	9.6	40.44	<0.001
8-14 days	6.7	35.47	<0.001
15-21 days	4.9	28.91	<0.001
22-90 days	3.2	28.40	<0.001
91 days to 6 months	1.3	9.24	<0.001
Over 6 months (Ref)			
Payer:			
Indemnity	-2.1	-8.17	<0.001
Medicaid	-3.3	-15.71	<0.001
Medicare B, age 65 or above	0.8	5.67	<0.001
Medicare B, Under age 65	-2.4	-10.49	<0.001
No fault, Auto	-2.2	-10.46	<0.001
Workers' compensation	-5.6	-26.52	<0.001
HMO, preferred provider (Ref)			
Surgical history:			
1 related surgery	-2.6	-15.73	<0.001
2 related surgeries	-3.4	-10.89	<0.001
3 or more related surgeries	-4.0	-9.67	<0.001
No related surgeries (Ref)			

Significant Predictors of FS at Discharge (Reference group for categorical variables)	**B**	**T**	**P-value**
Exercise history:			
At least 3x/week	0.6	5.99	<0.001
1-2x/week	0.5	4.65	<0.001
Seldom or Never (Ref)			
Medication use at intake	-0.8	-8.68	<0.001
Previous treatment	-1.7	-18.10	<0.001
Post-surgical: Fusion	1.5	4.41	<0.001
Comorbidities			
Anxiety	-0.6	-4.72	<0.001
Arthritis	-0.7	-7.17	<0.001
Back pain	-1.0	-8.85	<0.001
Depression	-0.9	-6.87	<0.001
Gastro	-0.4	-3.13	0.002
Headaches	-0.8	-8.81	<0.001
Kidney, Bladder, Prostate or Urination	-0.6	-3.62	<0.001
Obesity (BMI>=30)	0.5	5.72	<0.001
Osteoporosis	-0.6	-3.51	<0.001
Previous accidents	-0.9	-7.11	<0.001
Sleep dysfunction	-1.3	-12.03	<0.001

B-coefficient indicating the amount of expected change in discharge FS given a 1-unit change in the value of the variable, given that all other variables in the model are held constant.

T values indicate the importance of each independent variable for predicting discharge FS

Abbreviations: BMI, body mass index (kg/m2); FS, functional status; HMO, health maintenance organization.

2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors

(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects.) Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.

Given the results presented above in TABLE 2b3.4a, it would appear that the variables for Medicaid or Medicare B under age 65 have a notable influence toward predicting poorer outcomes of functional status change. While these variables may represent aspects of social risk, it would be illogical to remove them and test the model separately without them because their primary purpose is to provide a complete list of payer categories. Presently, we have collected a data set pertaining to the construct of educational level and plan to include in the next round of risk adjustment analyses.

2b3.5. 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)

To assess for overfitting, we examined results from three cross-validation analyses using two randomly and evenly split samples: a development and a test sample. We fit the stepwise regression model separately for the development and test samples. Variables that were significant in both samples were identified as being 'stable' and tested in the final model. Next, we calculated R-squared shrinkage [33] and the predictive ratio[13]. R-squared shrinkage was assessed using several approaches. We compared the adjusted R-squared to the unadjusted R-squared results from the stepwise regression. The adjusted R-squared is an estimate of what the fit of the regression model would be if it were fitted against a new data set, assuming all the degrees of freedom have been accounted for.[33] Then, we used the development sample to estimate the predicted FS at discharge for the full sample (development and test samples). The predicted estimate was then fitted against the FS scores at discharge using only the test sample. We compared the predictive power (R-squared) of the test sample using a prediction model created using the development sample, to the R-squared of the development sample. Shrinkage is defined as the decrease in R-squared between the development sample and the test sample. Although there are no clear standards for acceptable levels of shrinkage, we considered shrinkage of less than 10% to be sufficient to support the generalizability of the model's coefficients. As a confirmation analysis, a previously recommended bootstrap procedure [37] was applied using the 'regvalidate' STATA program.[38] To estimate the predictive ratio, the mean predicted discharge FS scores of the test sample, estimated using the development sample, was divided by mean actual discharge FS scores obtained from the test sample.[39] When the average predicted discharge FS was close to the average actual discharge FS, i.e., the predictive ratio is close to 1, the predictive validity of the regression model was considered to be supported.[13,39]

Additionally, the final model's error terms (residuals) for the test sample were visually inspected to assess for normality and homoscedasticity (i.e., deviations of the residuals are constant across the predicted outcome). Normality and homoscedasticity are assumptions of linear regression. The residual was the difference between the actual and predicted outcome, with positive and negative residuals representing higher and lower outcomes, respectively. We preferred the visual inspection over statistical testing because large datasets tend to have substantial power and can yield statistically significant results when there are only trivial deviations from normality and homoscedasticity. Normality was inspected by plotting a normal distribution line against the distribution of the residuals. Homoscedasticity was inspected by fitting a regression line to the squared residuals across the predicted outcome. A horizontal fitted line supports homoscedasticity.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.

If stratified, skip to 2b3.9

2b3.6. Statistical Risk Model Discrimination Statistics

(e.g., c-statistic, R-squared)**:**

The model identified 11 constructs that explained 33.3% of the variance in discharge FS, with FS at admission, acuity, payer type and age being the most important predictors.

R-squared shrinkage was less than 1% for both methods used to assess shrinkage.

The average predicted discharge FS of the test sample (n/2= 38,614), estimated using the development sample, was practically identical to the average actual discharge FS obtained by the test sample (63.767 and 63.769, respectively) resulting in a predictive ratio of 1.0.

Plots of the model's residuals for normality and homoscedasticity are presented in FIGUREs 2b3.6i-ii, respectively. The results supported normality with only slight deviations. Residuals were consistent across the predicted FS scores, supporting homoscedasticity.

FIGURE 2b3.6i: Visual inspection of normality of residuals

Distribution of the error term (residuals) from the risk-adjusted model, compared to the normal distribution. A distribution of residuals that is close to normal supports the normality assumption of linear regression.

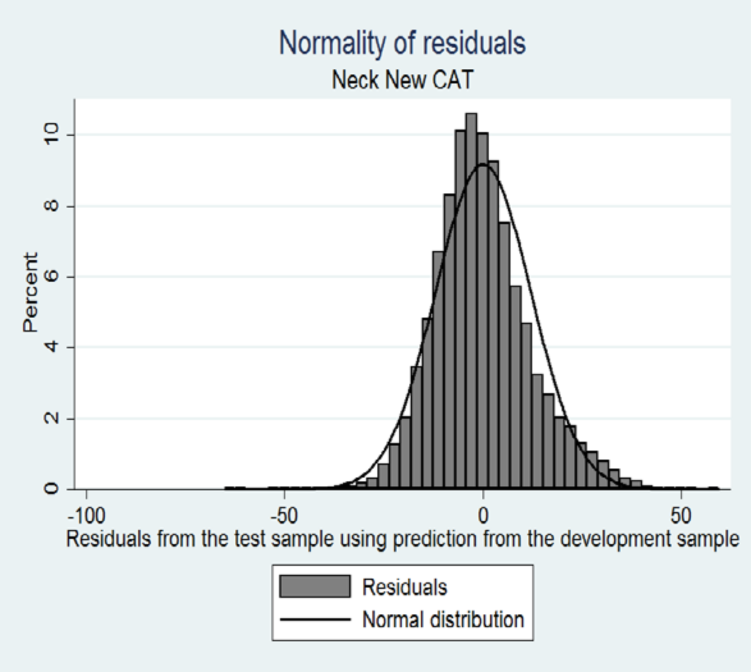
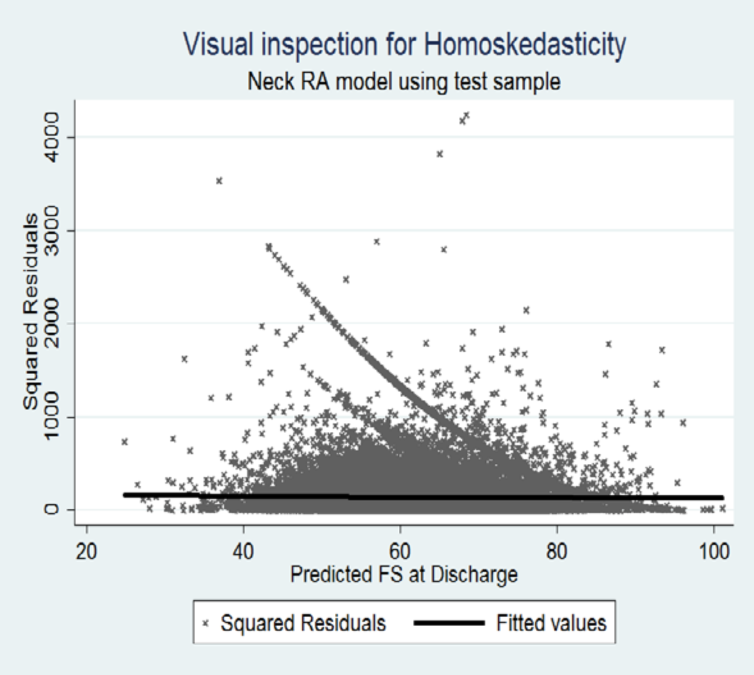


FIGURE 2b3.6ii: Visual inspection of homoscedasticity

Distribution of residuals (squared) across the range of the predicted FS scores at discharge. The fitted line represents fitted values for the squared residuals. A horizontal fitted line supports the homoscedasticity assumption of linear regression; that is, deviations of residuals are constant across the predicted outcome.



2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b3.9. Results of Risk Stratification Analysis:

2b3.10. 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)

We are not aware of an agreed upon value for an acceptable level of shrinkage. However, we considered a shrinkage of less than 1% to strongly support the model's external validity. Along with the predictive ration of 1, we interpret these results providing strong support for the predictive validity of the final risk-adjusted model. Additional support for the model's validity was provided by the support of the normality and homoscedasticity assumptions of linear regressions.

2b3.11. 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)

To assess the potential for patient selection bias and the impact of our selection criteria, we assessed the impact of adjusting for patient censoring using inverse probability weighting (IPW) on our results.[40] In this method, complete cases are weighted by the inverse of their probability of being a complete case.[41] Hence, patients less likely to have complete FS data were given more weight in the risk-adjusted model than those who were likely to have complete data.[40] We compared the coefficients created by the un-weighted and weighted models.

All coefficients were practically identical when using un-weighted and weighted models, supporting missing data are mostly missing at random.

2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b4.1. 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)

2b4i. Data elements (patient) level:

Performance of the PROM at the patient level was assessed by calculating the patient's risk-adjusted residual score after modeling as described above in section 2b3. We calculated residual scores for each patient, which we interpret as the amount of FS change beyond the predicted value, given their independent patient characteristics adjusted within the model. If the residual score is greater than zero the patient changed more than expected, and if less than zero the patient changed less than expected.

2b4ii-iii. Clinician & Clinic Performance Score Level:

We calculated patient level residual scores (residual = actual change – predicted change) after risk adjustment modeling and aggregated scores by individual clinician or clinic. Performance of providers was evaluated using uncertainty assessments and percentile ranking as described in section 2b1iv-v above (Validity of Clinician & Clinic Performance Score Level).

2b4.2. 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)

2b4i. Performance patient level

The mean residual score was 0, sd = 12.1, and range -65.1 to 59.5.

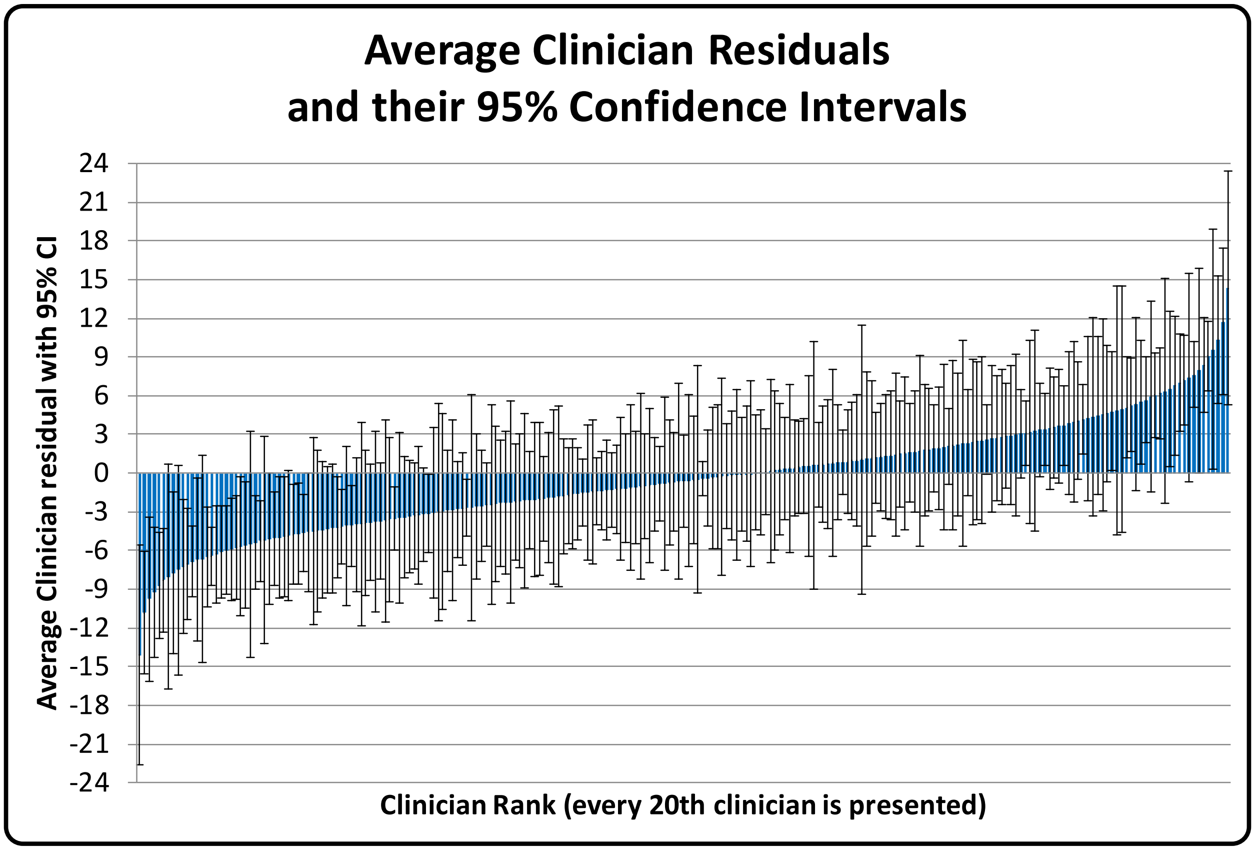
2b4ii. Performance individual clinician level

The distribution of clinician performance is presented in the Validity Testing section above.

Clinician performance based on uncertainty assessments is summarized in TABLE 2b1iv-a above (Validity Testing section), and illustrated in FIGURE 2b4ii below, with 16%, 73% and 11% of clinicians achieving low, average and high performance, respectively.

Clinician performance based on percentile ranking (deciles) is summarized in TABLE 2b1iv-b above (Validity Testing section) and illustrated in FIGURE 2b1iv-b above (Validity Testing section), showing monotonic increase between ranks of rates of patients achieving the minimal clinically important improvement (MCII), which were also statistically different from one another. Also, clinicians at the highest performance rank had on average only 80% of patients achieving the MCII, leaving room for further improvement.

FIGURE 2b4ii: Average clinician residual (95%CI)



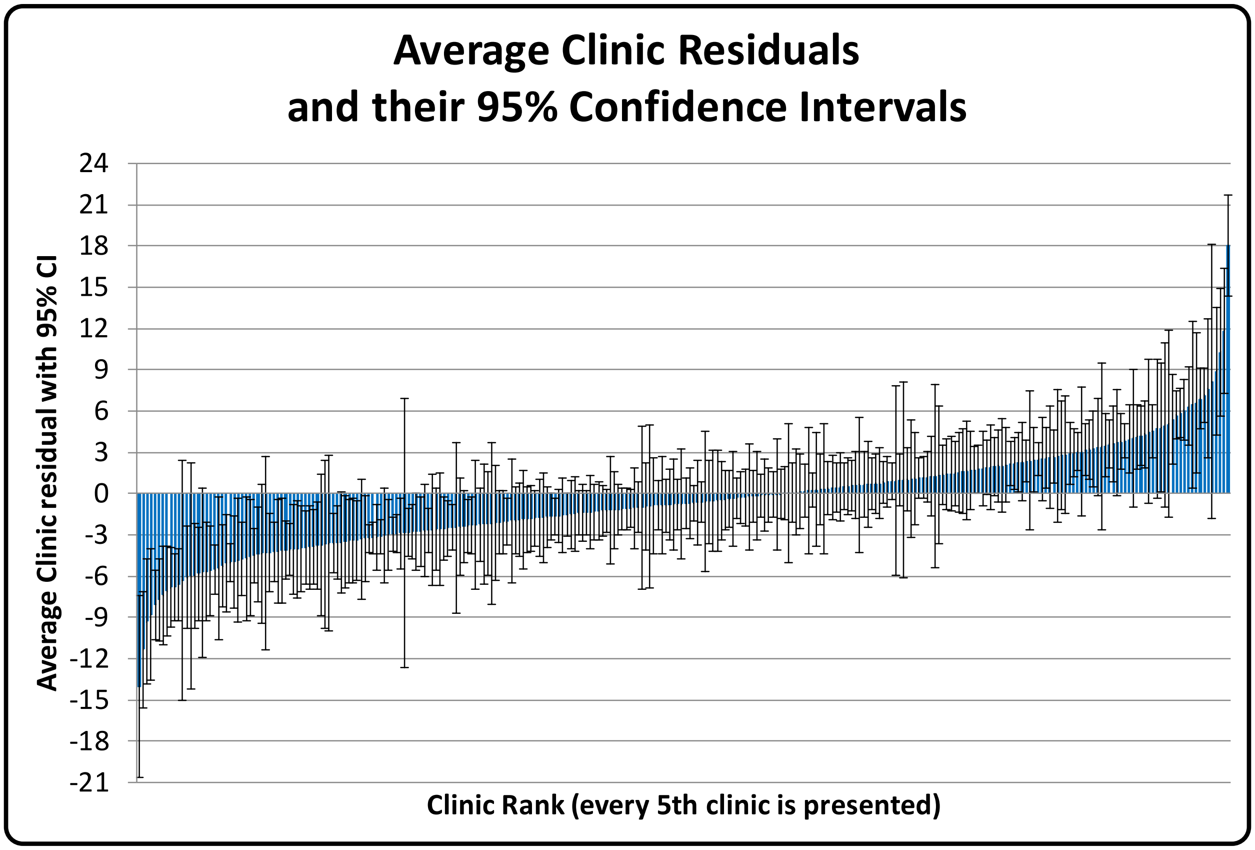
2b4iii. Performance clinic/group practice level

The distribution of clinic performance is presented in the Validity Testing section above.

Clinic performance based on uncertainty assessments summarized in TABLE 2b1v-a above (Validity Testing section), and illustrated in FIGURE 2b4iii below, with 24%, 60%, and 16% of clinic achieving low, average, and high performance, respectively.

Clinic performance based on percentile ranking (deciles) is summarized in TABLE 2b1v-b above (Validity Testing section) and illustrated in FIGURE 2b1v-b above (Validity Testing section), showing monotonic increase between ranks of rates of patients achieving the minimal clinically important improvement (MCII), which were also statistically different from one another, except for deciles 5 & 6. Also, clinics at the highest performance rank had on average only 76% of patients achieving the MCII, leaving room for further improvement.

FIGURE 2b4iii: Average clinic residual (95%CI)



2b4.3. 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?)

These results support the ability of the Neck PRO-PM scores to identify statistically significant and clinically important differences in performance levels across patients and measured entities. Also, these results suggest the measure is not “topped out”; that is, there is additional room for clinically important improvement at high performance levels.

However, when estimating a level of performance based on an aggregate of patients, there is always a chance that different distributions of patient-level scores might exist between providers ranked at the same level of performance, and vice versa. This is an inherent limitation of using an average score representing aggregated data at the provider level. This concern is ameliorated by the use of lower threshold of number of patients seen by providers to ensure adequate reliability at the provider score level.

2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

NA

Note: *This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *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).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with***

more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.

2b5.1. Describe the method of testing conducted to compare 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*)

2b5.2. 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*)

2b5.3. What is your interpretation of the results in terms of the differences in 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*)

____

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. 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)

2b61i. Comparing patients with or without complete outcomes

Patient selection bias related to missing data could occur if patients with better outcomes were encouraged to report their outcomes and those with worse outcomes were discouraged from reporting. In this hypothetical scenario, a provider could potentially bias their data by not recording complete episodes (patient with admission and discharge outcomes data) for more 'difficult' patients that they perceive as having a potential of lowering their overall adjusted scores. This selection bias might occur even if it is not logical to do so from a statistical stand-point, since the measure is risk adjusted. This could lead to a less representative sample of those treated by the provider, with a potential to impact their performance scores. One common way to assess whether missing data is largely missing at random is to compare patients included to those excluded due to missing outcomes data at discharge to identify characteristics known to be associated with outcomes. If no specific trends are identified, the assumption of missing data largely at random is supported, reducing concern that systematic patient selection bias exists. We also tested the impact of a weighted adjustment for missing data using inverse probability weighting (IPW). [40] In this method, complete cases are weighted by the inverse of their probability of being a complete case.[41] Hence, patients less likely to have complete FS data were given more weight in the analyses of interest than those who were likely to have complete data.[40] See more details in section 2b61iv-v below.

Historically, FOTO has assessed the potential impact of missing data by comparing the characteristics of patients with and without complete FS data at discharge. Of interest were specific patient characteristics known to be predictive of outcomes. If a systematic patient selection bias at discharge existed, we expected that patients with complete PROMs data would have higher values or frequencies of characteristics associated with better outcomes (i.e., better FS) compared to those with incomplete PROMs data (e.g., younger, more acute conditions, more active exercise history). We compared characteristics of patients with incomplete (admission only) and complete (admission and discharge) PROMs data using t-tests or chi-square as appropriate (See TABLE 2b62i below).

The following patient characteristics (and their known associations with outcomes) were used to compare those with complete and incomplete outcomes data. We evaluated FS scores at admission because they are known to be the strongest positively associated predictor of outcomes, i.e., higher FS at admission is associated with higher FS at discharge. Other continuous variables studied were age and number of comorbidities, both of which are negatively associated with outcomes. Categorical variables and their known association with outcomes included: sex (lower outcomes for females); acuity as number of days from onset of the treated condition (6 categories) with more chronic conditions associated with lower outcomes; type of payer (10 categories) with most categories associated with lower outcomes compared to Health Maintenance Organization (HMO) and Preferred Provider Organization (PPO), except for Medicare B aged 65 or above; surgical history as number of related surgeries (4 categories) with no surgical history associated with higher outcomes; exercise history (3 categories) with higher levels of exercise history associated with higher outcomes; use of medication at intake for the treatment of the neck pain (yes/no); and having received previous treatment for neck pain (yes/no), both associated with lower outcomes.[13,32,35,42]

2b61ii-iii. Correlations between clinician and clinic residuals and their completion rates

We assessed whether missing data was a source of systematic bias by testing associations between clinician and clinic completion rates and clinician and clinic quality (as measured by clinic average residual scores after risk adjustment modeling) for clinicians and clinics included in the performance analysis. Residual scores are the difference between predicted functional outcomes (given risk adjustment factors) and the actual outcomes. Existence of systematic bias was assumed to result in some associations between completion rates and quality, with possibly higher quality for providers with lower completion rates, if patient with higher outcomes were systematically selected to complete more surveys at discharge compared with those thought to have lower outcomes. We examined Pearson Correlations between clinician and clinic completion rate and their average residual scores.

2b61iv-v. Average residuals at the clinician or clinic levels by completion rate categories with or without the use of inverse probability weighting

To further examine whether there was an underlying pattern to the relationship between clinic completion rate and risk adjusted residual scores aggregated at the clinician and clinic levels, and the impact on such relationship when adjusting for missing data using inverse probability weighting (IPW) described above, we grouped clinicians and clinics into 10 completion rate categories.

2b6.2. 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)

2b62i. Comparing patients with or without complete outcomes

The comparison of patients with complete and incomplete FS outcomes data is presented in TABLE 2b62i. No differences between groups were identified for admission FS and sex. Statistically significant but trivial differences were identified for number of comorbidities, acuity, exercise and surgical history, and receiving previous treatment for neck pain. Patients with complete outcomes data were 3 years older (not supporting potential for selection bias), had a higher rate of Medicare Part B for ages 65 or above (supporting potential for selection bias), and used less medications related to their neck pain at admission (supporting potential for selection bias), compared to those with incomplete outcomes data. Overall, these analyses were inconclusive and did not support a systematic patient selection bias.

TABLE 2b62i: Health and Demographic Patient Characteristics of those with complete or incomplete FS outcomes data*

Patient characteristics	Total (N= 250,741)	Complete (n= 169,039; 67%)	Incomplete (n= 81,702; 33%)	p-value[†]
FS score at admission: Mean ± SD (Min to Max)	51.9±12.6 (3-97)	51.9±12.3 (3-96)	51.8±13.1 (3-97)	0.205[‡]
Age (years): Mean ± SD (Min to Max)	53.6±16.0 (14-89)	54.6±16.2 (14-89)	51.5±15.5 (14-89)	<0.001[‡]
Number of comorbidities: Mean ± SD (Median, IQR §)	5.3±3.1 (5,)	5.3±3.1 (5,)	5.3±3.2 (5,)	0.004[‡]
Sex: Female	65.3	65.2	65.4	0.482
Acuity:				<0.001
0-7 days	4.0	4.0	3.9	
8-14 days	6.7	6.9	6.3	
15-21 days	8.4	8.6	8.0	
22-90 days	26.8	27.1	26.1	
91 days to 6 months	14.0	14.1	14.0	
Over 6 months	40.1	39.4	41.7	
Payer:				<0.001
Indemnity insurance	3.5	2.9	4.6	
Medicaid	5.7	4.9	7.3	
Medicare A	1.2	1.3	1.0	
Medicare B Under Age 65	4.0	3.8	4.4	
Medicare B Age 65 or above	18.1	20.6	13.0	
Patient	0.7	0.6	0.8	
Workers compensation	4.2	4.6	3.5	
Other (Litigation, Medicare C, School, No charge, Early Intervention, Commercial Insurance)	10.7	10.4	11.4	
No fault, Auto insurance	4.3	4.6	3.5	
HMO, PPO	47.6	46.3	50.4	
Surgical history:				<0.001
No related surgery	87.7	87.6	87.9	
1 related surgery	9.0	9.2	8.6	
2 related surgeries	2.1	2.1	2.2	
3 or more related surgeries	1.2	1.2	1.3	
Exercise history:				<0.001
At least 3x/week	38.1	38.5	37.2	
1-2x/week	26.4	26.5	26.4	
Seldom or Never	35.4	35.0	36.4	
Medication use at intake	50.9	50.3	52.3	<0.001
Previous treatment	40.0	40.3	39.4	<0.001

Difference not supporting potential for selection bias

Differences supporting potential for selection bias

No differences or differences interpreted as not clinically important

Abbreviations: FS, functional status; N/A, Not Available; IQR, inter quartile range; HMO, health maintenance organization; PPO, preferred provider organization

Patient characteristics for all included patients (Total), patient with functional status data at admission and discharge (Complete) and patient with functional status data at admission only (Incomplete).

Values are percent unless otherwise indicated.

†*P-values are a result of chi-square tests unless otherwise indicated.*

‡*P values are a result of t tests.*

§Median and IQR are reported for number of comorbidities due to the skewed distribution

2b61ii-iii. Correlations between clinician and clinic residuals and their completion rates

No correlations were found between completion rates and residual scores. At the clinician and clinic levels, correlations were 0.008 and 0.007, respectively.

2b62iv-v. Average residuals at the clinician and clinic levels by completion rate categories with or without the use of Inverse Probability Weighting

Results shown below suggest that the relationship between completion rate and aggregated residual scores is not linear and has no strong pattern, with no impact of IPW on the results.

TABLE 2b62iv: Average residuals at the clinician level by completion rate categories with or without the use of inverse probability weighting

Completion rate categories (%)	N patients	N clinics	Residual without IPW	Residual with IPW
20-30	94	6	-0.14	-0.12
30-40	884	50	0.53	0.52
40-50	4166	245	0.05	0.05
50-60	11685	573	0.01	0.01
60-70	21269	946	-0.08	-0.08
70-80	33569	1337	0.01	0.01
80-90	29179	1053	0.12	0.13
90-100	11332	501	0.42	0.42
Total	**112,178**	**4,711**	**0.1**	**0.1**

TABLE 2b62v: Average residuals at the clinic level by completion rate categories with or without the use of inverse probability weighting

Completion rate categories (%)	N patients	N clinics	Residual without IPW	Residual with IPW
30-40	523	11	0.21	0.28
40-50	4506	72	-0.03	-0.05
50-60	14967	185	-0.05	-0.05
60-70	30715	345	0.01	0.01
70-80	41542	423	0.02	0.03
80-90	27015	262	0.14	0.13
90-100	3926	80	0.36	0.35
Total	**123,194**	**1,378**	**0.0**	**0.0**

2b6.3. 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)

Overall, the comparisons of characteristics of patients with and without complete outcomes data show no systematic pattern suggesting a selection bias in the collection of discharge Neck FS PROM data.

The lack of correlations between completion rates and residual scores strengthens the conclusion of no systematic patient selection bias. Finally, the lack of a linear association between completion rate categories and average residuals at the clinician and clinic levels, with no impact of adjustment for missing data using IPW, supports that missing data were mostly missing at random.

REFERENCES

1. Wang YC, Cook KF, Deutscher D, Werneke MW, Hayes D, Mioduski JE. The Development and Psychometric Properties of the Patient Self-Report Neck Functional Status Questionnaire (NFSQ). *J Orthop Sports Phys Ther.* 2015;45(9):683-692.
2. Lord FM, ed *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Associated; 1980.
3. Pilkonis PA, Yu L, Dodds NE, Johnston KL, Maihoefer CC, Lawrence SM. Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS) in a three-month observational study. *J Psychiatr Res.* 2014;56:112-119.
4. Neyman J. Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philosophical Transactions of the Royal Society of London, Ser A 767.* 1937;236:333-380.
5. O'Brien SF, Yi QL. How do I interpret a confidence interval? *Transfusion.* 2016;56(7):1680-1683.
6. Wang YC, Hart DL, Cook KF, Mioduski JE. Translating shoulder computerized adaptive testing generated outcome measures into clinical practice. *J Hand Ther.* 2010;23(4):372-382; quiz 383.
7. Wang YC, Hart DL, Stratford PW, Mioduski JE. Clinical interpretation of computerized adaptive test outcome measures in patients with foot/ankle impairments. *J Orthop Sports Phys Ther.* 2009;39(10):753-764.
8. Wang YC, Hart DL, Stratford PW, Mioduski JE. Clinical interpretation of computerized adaptive test-generated outcome measures in patients with knee impairments. *Arch Phys Med Rehabil.* 2009;90(8):1340-1348.
9. Wang YC, Hart DL, Stratford PW, Mioduski JE. Clinical interpretation of a lower-extremity functional scale-derived computerized adaptive test. *Phys Ther.* 2009;89(9):957-968.
10. Wang YC, Hart DL, Werneke M, Stratford PW, Mioduski JE. Clinical interpretation of outcome measures generated from a lumbar computerized adaptive test. *Phys Ther.* 2010;90(9):1323-1335.
11. Stratford PW. Getting more from the literature: Estimating the standard error of measurement from reliability studies. *Physiother Can.* 2004;56:27-30.
12. Adams JL. The Reliability of Provider Profiling: A Tutorial. In: RAND Corporation; 2009.
13. Hart DL, Connolly JB. Pay-for-Performance for Physical Therapy and Occupational Therapy: Medicare Part B Services. Grant #18-P-93066/9-01. In: Health & Human Services/Centers for Medicare & Medicaid Services.; 2006.

14. Hart DL, Deutscher D, Werneke MW, Holder J, Wang YC. Implementing computerized adaptive tests in routine clinical practice: experience implementing CATs. *Journal of applied measurement.* 2010;11(3):288-303.

15. Nunnally JC. *Psychometric theory.* New York,: McGraw-Hill; 1967.

16. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care.* 2007;45(5 Suppl 1):S22-31.

17. Hu LT, Bentler P. Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling.* 1999;6:1-55.

18. Wright BD, Linacre JM. Reasonable meansquare fit values. *Rasch Meas Trans.* 1994;8:370.

19. Jette AM, Tao W, Norweg A, Haley S. Interpreting rehabilitation outcome measurements. *J Rehabil Med.* 2007;39(8):585-590.

20. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials.* 1989;10(4):407-415.

21. Hart DL, Wang YC, Stratford PW, Mioduski JE. A computerized adaptive test for patients with hip impairments produced valid and responsive measures of function. *Arch Phys Med Rehabil.* 2008;89(11):2129-2139.

22. Hart DL, Wang YC, Stratford PW, Mioduski JE. Computerized adaptive test for patients with foot or ankle impairments produced valid and responsive measures of function. *Qual Life Res.* 2008;17(8):1081-1091.

23. Hart DL, Wang YC, Stratford PW, Mioduski JE. Computerized adaptive test for patients with knee impairments produced valid and responsive measures of function. *J Clin Epidemiol.* 2008;61(11):1113-1124.

24. Wang YC, Hart DL, Stratford PW, Mioduski JE. Baseline dependency of minimal clinically important improvement. *Phys Ther.* 2011;91(5):675-688.

25. Stratford PW, Binkley J, Solomon P, Finch E, Gill C, Moreland J. Defining the minimum level of detectable change for the Roland-Morris questionnaire. *Phys Ther.* 1996;76(4):359-365; discussion 366-358.

26. Riddle DL, Stratford PW, Binkley JM. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 2. *Phys Ther.* 1998;78(11):1197-1207.

27. Ostelo RW, Deyo RA, Stratford P, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. *Spine.* 2008;33(1):90-94.

28. Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L, Grunnet-Nilsson N. Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. *BMC Musculoskelet Disord.* 2006;7:82.

29. Hart DL, Werneke MW, Wang YC, Stratford PW, Mioduski JE. Computerized adaptive test for patients with lumbar spine impairments produced valid and responsive measures of function. *Spine (Phila Pa 1976).* 2010;35(24):2157-2164.

30. Hart DL, Wang YC, Cook KF, Mioduski JE. A computerized adaptive test for patients with shoulder impairments produced responsive measures of function. *Phys Ther.* 2010;90(6):928-938.

31. Goldsmith CH, Boers M, Bombardier C, Tugwell P. Criteria for clinically important changes in outcomes: development, scoring and evaluation of rheumatoid arthritis patient and trial profiles. OMERACT Committee. *J Rheumatol.* 1993;20(3):561-565.

32. Deutscher D, Werneke MW, Hayes D, et al. Impact of Risk-Adjustment on Provider Ranking for Patients With Low Back Pain Receiving Physical Therapy. *J Orthop Sports Phys Ther.* 2018:1-35.

33. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med.* 2004;66(3):411-421.

34. Maxwell SE. Sample size and multiple regression analysis. *Psychol Methods.* 2000;5(4):434-458.

35. Deutscher D, Horn SD, Dickstein R, et al. Associations between treatment processes, patient characteristics, and outcomes in outpatient physical therapy practice. *Arch Phys Med Rehabil.* 2009;90(8):1349-1363.

36. Gozalo PL, Resnik LJ, Silver B. Benchmarking Outpatient Rehabilitation Clinics Using Functional Status Outcomes. *Health Serv Res.* 2016;51(2):768-789.

37. Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol.* 2001;54(8):774-781.

38. Ender PB. regvalidate. 2010; http://www.philender.com/courses/linearmodels/notes2/cross.html.

39. Kautter J, Ingber M, Pope GC, Freeman S. Improvements in Medicare Part D risk adjustment: beneficiary access and payment accuracy. *Med Care.* 2012;50(12):1102-1108.

40. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11(5):550-560.

41. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res.* 2013;22(3):278-295.

42. Deutscher D, Hart DL, Stratford PW, Dickstein R. Construct validation of a knee-specific functional status measure: a comparative study between the United States and Israel. *Phys Ther.* 2011;91(7):1072-1084.

Appendix: Item descriptions

#	Name	Short description
1	LOOKBHN	Turning to look behind you
2	25LBBOX	Placing a 25lbbox on a shelf overhead
3	GOLF	Performing forceful recreational activities
4	SHOVEL	Using a shovel to dig a hole in the dirt
5	ONSHLDR	Carrying objects on your shoulders
6	BCKSEAT	Touching an object on the back seat of a car
7	HVYSUIT	Lifting and carrying a heavy suitcase
8	WRKOVRH	Work overhead for more than 2 minutes
9	DESKWRK	Light desk work for 8 hours
10	MOVGQCK	Moving your head quickly
11	BHNDDRV	Turning to look behind you to drive a car
12	LFT30LB	Lifting medium weights (20-30 lb.) from the floor
13	BENDING	Bending over to clean a bathtub
14	GARDEN	Performing garden or yard work
15	SEEBIRD	Looking up to see a bird
16	BULB	Changing a light bulb overhead
17	READGBK	Sitting and reading a book for 1 hour
18	TURNBED	Turning over in bed
19	HVYDOOR	Pulling or pushing a heavy door
20	VACUUM	Using a vacuum cleaner
21	LWR5LBS	Lowering a light-weight object (1-5 lb.) from a top shelf
22	CARDS	Performing low effort recreational activities
23	RCHSHLF	Reaching a shelf that is at shoulder height
24	PULLSTR	Reaching and pulling a fan string
25	CANSHLF	Placing a can of soup on a shelf overhead
26	SEESHOE	Looking down to see your shoes
27	COMBING	Combing or brushing your hair
28	BATHING	Performing personal care activities

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. Data Elements Generated as Byproduct of Care Processes.

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Other

If other: Additionally, computer-administration to collect the patient-reported components. This clarification also applies to our response in 3b.1 below. Furthermore, the NQF Feasibility Score Card is NA because this is not an eMeasure.

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. To what extent are the specified data elements available electronically in defined fields (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.

Patient/family reported information (may be electronic or paper)

3b.2. If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources. For maintenance of endorsement, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

All the data elements are from electronic sources with the exception of the provider having the option to print the short form for manual administration and scoring.

3b.3. If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.

IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

DATA COLLECTION

For patients (i.e., those providing the data), patients respond to, on average, 5 questions from the Neck FS PROM CAT followed by 10 questions pertaining to risk adjustment. The typical amount of time needed to complete the PROM and risk adjustment questions is 5 minutes. (N=167,488, year 2018, excluded top and bottom 10th percentiles due to clinical environment issues such as interrupted assessments with the system left idle).

For patients who have difficulty responding independently to computer-administered questions, the FOTO system allows for both Proxy and Recorder modes of administration. Please see Specifications tab, S.15. Sampling, for further details about Proxy and Recorder modes of administration.

For providers (i.e., those being measured), a few minutes of set-up time, usually by front office staff, is required to input certain details such as patient name, age, and payer source. This set up time is eliminated for many providers with an electronic health record (EHR) that has written to FOTO's applied interface programming (API). Presently, 14 EHR companies are integrated with the FOTO API for the sake of eliminating double entry for the provider, that is, the provider only needs to enter standard medical record-type data points in the EHR, and the needed data points for FOTO are automatically pulled from the EHR into the FOTO system. The current 14 EHR integrations benefit 1136 clinics that subscribe to the FOTO system. We expect these numbers to continue to grow.

AVAILABILITY OF DATA

For patients: all data points requested for entry by patients are of the patient-self report nature and thus readily available

For providers: any data points requested for entry by providers are also readily available in that they already have or need the data points as part of the standard medical record.

MISSING DATA

For patients – Missing data on the patient level is relevant in that the PROM and related results are meaningful in the context of patient-provider communication and clinical decision-making in the context of the individual patient episode that is being managed at the time. FOTO provides clinical education about using patient-reported outcome data in clinical care.

For providers – Providers insure that clinic operational processes support strong rates of completed episodes. That is, insuring that each patient completes an assessment at Intake (admission) and at least one additional time at or near the time of discharge from the episode of care. Furthermore, providers must officially close the episode of care (discharge) by providing the number of visits incurred and date of last visit (for duration) in the FOTO system; alternatively, this can be accomplished automatically by discharging the patient in the EHR only, with the needed data points sent automatically from the EHR to the FOTO system.

TIMING AND FREQUENCY OF DATA COLLECTIONN

The assessments are to be completed, at a minimum, at the time of Intake (admission) and at least one additional time at or near the time of discharge from the episode of care. Furthermore, providers must officially close the episode of care (discharge) by providing the number of visits incurred and date of last visit (for duration) in the FOTO system; alternatively, this can be accomplished automatically by discharging the patient in the EHR only, with the needed data points sent automatically from the EHR to the FOTO system.

SAMPLING

Sampling is NA. All patients with neck impairments are included.

PATIENT CONFIDENTIALITY

The FOTO system follows all requirements of the Healthcare Insurance Portability and Accountability Act (HIPAA) to protect the confidentiality, integrity and availability of patient data. FOTO uses an Information Security Management System, and policies for all relevant areas of HIPAA are maintained and reviewed on an annual basis. Strong encryption is used for all data in transit and at rest. The application is scanned weekly for vulnerabilities, with reports issued to the development and IT teams to address any findings. Infrastructure is hosted by a third-party datacenter which undergoes a Service Organization Control 2 Type II audit on an annual basis and employs redundant mechanisms and channels to keep data highly available. A Business Continuity/Disaster Recovery plan is in place to ensure there is no data loss if the primary site is inoperable. Risk management is performed on an annual basis to identify and plan for any potential risks from an application and corporate level. Business Associate Agreements are executed with all customers and contain specific details about FOTO’s responsibilities hosting the provider’s data.

TIME AND COST OF DATA COLLECTION

The information provided below in section 3c.2. regarding fees and licensing is most relevant in addition to the information provided above under Data Collection.

3c.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified *(e.g., value/code set, risk model, programming code, algorithm)*.

Providers have 3 options for use of the Neck PRO-PM:

1. Free public access
   a. The components needed to calculate the reportable scores are available free for use by providers at https://www.fotoinc.com/science-of-foto/nqfneck
2. FOTO Outcomes Manager (OM) Lite services
   a. Provides the minimal level of services required for providers' regulatory and compliance needs such as the Merit-based Incentive Program (MIPS).
   b. Specifically, OM Lite provides the services of data collection, scoring for both the Neck FS PROM and the PRO-PM components, patient- and clinician-level reporting for the individual patient results for use in patient-clinician communication and engagement, aggregation of risk-adjusted benchmarked results on the clinician and clinic levels to assist in quality assurance/improvement initiatives.
   c. Pricing: $250 one-time set up fee, $20 per clinic/month, $15 per clinician/month.
3. FOTO Outcomes Manager (OM) services
   a. The OM level provides the same services described under OM Lite above. The OM level also provides additional services that promote the use of patient-reported outcomes in improving quality of care and costs, e.g., an effectiveness/efficiency ratio derived from aggregated risk adjusted functional status change relative to the number of visits used per episode of care are reported for each body part or impairment. The provider's utilization scores are compared to national utilization scores from all providers to identify performance areas that the provider is excelling at or needs to improve.
   b. Pricing $350 one-time set up fee, $50 per clinic/month, $25 per clinician/month

The feasibility (affordability) of the costs for OM and OM Lite is supported by the finding that, as of March 2019, 24,061 clinicians in 3837 clinics in the United States, were subscribed to the full service level (OM) and 206 clinics (with 694 clinicians) preferred the lower cost option of OM Lite. In total, 4043 clinics (consisting of 24,755 clinicians) across all 50 United States find the costs and operations to be feasible.

4. Usability and Use

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

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	**Current Use (for current use provide URL)**
	Public Reporting Physician Compare via MIPS as QCDR measure https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/physician-compare-initiative/ Payment Program MIPS as QCDR measure https://www.fotoinc.com/science-of-foto/qcdr-measure-specification Quality Improvement (Internal to the specific organization) Therapy Partners (TPI) https://therapypartners.com/foto-outcomes/

4a1.1 For each CURRENT use, checked above (update for maintenance of endorsement), provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included
- Level of measurement and setting

CURRENT USE: PAYMENT PROGRAM AND PUBLIC REPORTING

This measure is currently part of a payment program in that it is a QCDR measure for the 2019 CMS Merit-based Incentive Program (MIPS) performance year. CMS has stated that results from measures in the MIPS program will be publicly reported on Physician Compare.

CURRENT USE: PAYMENT PROGRAMS

Secondly, the PRO-PM measures in the FOTO system, including the Neck Functional Status (FS) PRO-PM (NQF measure 3461), are used in state-level payer initiatives. Below are two examples:

1. The Physical Therapy Provider Network (PTPN https://www.ptpn.com/) is a national network of over 700 private practice physical, occupational, and speech therapy providers. PTPN uses the FOTO Outcomes Management system, which includes the Neck FS PRO-PM. PTPN has an outcomes bonus programs with large health plan partners in California, Arizona, and Louisiana. For providers who provide effective and efficient care, the outcome bonus program rewards the providers with higher reimbursement per visit. Based on the provider's using FOTO risk adjusted outcome measures of functional status and number of visits, including the Neck FS PRO-PM, PTPN's data show that the providers who qualify for the bonuses get better than predicted functional outcomes in fewer than predicted visits. This results in a lower overall cost per case, even with the bonus reimbursement, with demonstrated quality and efficiency of care.
2. Therapy Partners (TPI) is a network of sixteen practices with thirty-five locations in Minnesota and western Wisconsin. TPI uses FOTO outcomes in value-based contracts with payers. The results from the FOTO PRO-PMs, including the Neck FS PRO-PM, are used in aggregate to determine a portion of the payment based on achieving certain standards of functional improvement (measured by the PRO-PM) and efficiency (measured by number of treatment visits). Because of the risk adjustment component of each PRO-PM, payers are able to differentiate levels of performance between practices and provider networks. The PRO-PM system allows practices to be compared by payers and identifies the higher quality practices.

Further information about TPI payment program:

1. https://therapypartners.com/services/aco-health-plans/. Accessed April 5, 2019
2. https://therapypartners.com/foto-outcomes/. Accessed April 5, 2019.
3. https://cdn2.hubspot.net/hubfs/442011/docs/P4P/TPI%20Statement%20for%20Ways%20and%20Means.pdf?t=1531375320446. Accessed April 5, 2019.

Current use: Quality improvement (internal to the specific organization)

As described above, Therapy Partners (TPI) is a network of sixteen practices with thirty-five locations in Minnesota and western Wisconsin. TPI has used the FOTO system of PRO-PMs, including the Neck FS PRO-PM, for several years for a number of quality assurance and improvement efforts. Some examples of this include:

- Training, policies, and operational processes to support data collection integrity related to the PRO-PMS such as standards for administration of patient-reported outcome measures (PROMs) and holding clinicians and staff accountable to high PROM completion rates. A designated "FOTO Champion" at each practice location is responsible for carrying out the trainings and insuring policies and processes are followed.
- Each FOTO Champion additionally provides training for clinicians on clinical interpretation and application in patient care.
- Quality Assurance/Improvement-opportunities are regularly measured for each practice based on established thresholds for PRO-PM performance and efficiency of care (i.e., risk-adjusted results for number of visits)
- PRO-PM and efficiency results are shared with physicians and other referral sources as evidence of quality and to assist interdisciplinary communication regarding patient care.
- PRO-PM and efficiency results are shared with individual clinicians as part of the clinician's annual review as a basis for discussion of the clinician's performance.

4a1.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
It has not yet been publicly reported. However, as described above in section 4a1.1 this measure is currently part of a payment program in that it is a QCDR measure for the 2019 CMS MIPS performance year. CMS has stated that results from measures in the MIPS program will be publicly reported on Physician Compare. Furthermore, the measure has been submitted as a potential MIPS Clinical Quality Measure (CQM) for the 2020 MIPS performance year via the CMS Measures Under Consideration (MUC) process. The measure received a rating of "pass, pending NQF endorsement" in the MUC process.

4a1.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

As described above, currently in use as a QCDR measure for the 2019 MIPS performance year and anticipated as a MIPS CQM for the 2020 performance year. Physician Compare is the public reporting mode for MIPS.

4a2.1.1. **Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.**

How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.

DURING DEVELOPMENT

The development of the original item pool of 54 functional questions ("candidate items") for the Neck FS PROM was led by a clinician researcher with experience treating patients with neck impairments and based on review of ex¬isting measures in the literature and input from additional physical therapists with clinical experience treating patients with neck impairments. As a next step, the 54 original candidate items were formally assessed for clinical relevance by an expert panel of 8 physical therapists experienced in treat¬ing patients with neck impairments (3 women, 5 men; mean +/- SD years of clinical experience, 16 +/- 9 years) was assembled. Therapists rated the clinical relevance of the 54 items as (1) highly rel¬evant; (2) partially relevant, beneficial to ask; (3) neutral, not certain; and (4) not relevant at all. Items were considered for inclusion if they were rated as "highly rel¬evant" by at least half the therapists. After the expert panel consultation, 19 items were removed, resulting in the remain¬ing 35 items moving toward subsequent processing and analyses.

PERFORMANCE RESULTS, DATA, AND ASSISTANCE WITH INTERPRETATION PROVIDED TO THOSE BEING MEASURED (CLINICIANS AND CLINICS) DURING IMPLEMENTATION AND ON AN ONGOING BASIS

On the patient level

- Real time reports for individual patient results including PROM scores, PRO-PM (risk-adjusted) comparisons of scores and end-of-episode results (i.e., “predicted” results) and patient responses to individual functional questions
- Facilitates clinician communication with patient and clinician understanding of patient’s perception of function/functional change, clinical decision-making, treatment and discharge planning.
- Includes comparative data about # Visits to promote efficiency of care.
- Includes both a clinician-facing and patient-facing version (examples shown in link below)

On the clinician and clinic levels

- Risk adjusted, benchmarked comparative reporting (PRO-PM)
- easy accessibility via web-based portal with multiple filtering options (example of portal shown in link below)
- at a glance comparisons of statistically at-, below and above benchmark averages
- at a frequency of every 3 months, including both 3-month and rolling 12-month periods

Assistance with interpretation and ongoing education is provided via

- patient reports designed to make them easy to interpret
- new user orientations and ongoing opportunities for training sessions
- instructions and guides on both the report portal and web-based survey administration site
- easy access to specialized provider relations representatives via training sessions (both live and recorded), email, phone, web-conferencing and chat options

For examples of provider-level (clinic and clinician) reporting (FOTO Report Portal) and patient level reporting, please view https://www.fotoinc.com/science-of-foto/nqf-measure-specifications-1

Other Users

- Payers are potential other users. Education information that specifically targets payers is included on the FOTO website. The information includes how payers may be interested in interpreting and utilizing FOTO data to support quality and efficiency initiatives. https://www.fotoinc.com/payer accessed April 5, 2019.

TYPES OF MEASURED ENTITIES

In the a recent 12-month period ending February 28, 2019, risk-adjusted functional status outcomes (Neck PRO-PM) data was captured in the FOTO system for 128,868 completed episodes for patients with neck impairments. The patient episodes were incurred by 13,299 clinicians in 3840 clinics. All patients with neck impairments were eligible for inclusion.

4a2.1.2. Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

This is described above in section 4a2.1.1.

Additionally, providers receive email alerts when reports are ready for them to access on the report portal. The report portal has education built in such as footnote explanations. Contact information for more assistance is provided in multiple locations. Direct feedback is encouraged through providers’ contact with specialized FOTO provider relations representatives as described in 4a2.1.1.

When feedback suggests need for higher-level education related to the science of PRO measurement, the FOTO Director of Research and/or scientists are consulted to help with education and receive/consider feedback. Needs for science-related education may also be addressed by directing the individual to the Science of FOTO website at: http://www.fotoinc.com/science-of-foto

4a2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.

Describe how feedback was obtained.

As described above, the FOTO provider relations representatives have ongoing and frequent (daily) contact with clinicians who see patients with conditions such as neck impairments. The provider relations representatives frequently share clinician feedback with the FOTO Director of Research and scientist team. Examples of common themes from this feedback include:

1. Clinicians value the use of PROM and PRO-PM data to promote clinician understanding of the patient's perspective, enhance goal-setting and other communication between clinician and patient, utility in clinical decision-making and treatment/discharge planning with the patient
2. Clinicians have expressed a consistent desire for ongoing risk-adjustment model development with consideration of more variables/constructs such as post-surgical types and weighting of individual comorbidities.

Additionally, 4 physical therapists (from 4 different states) who care for patients with neck impairments in outpatient settings responded to a recent inquiry by the FOTO Research department. We asked the physical therapists for their feedback pertaining to how the Neck FS PROM/PRO-PM was implemented back in January 2016 and how the measure has performed so far. The following is a summary of the responses from each of the 4 clinicians:

1. " I think it was implemented well. sorry, not a lot of detail on this one….I think the item bank is a very good and realistic set of items to consider for the majority of patients." (SP - Wisconsin)
2. "I have no concerns or criticisms on my end."(DG - Tennessee)
3. "The implementation of this measure pre-dates my involvement in FOTO." Regarding how the measure has been performing, "I am certainly not an expert – nor do I have a good appreciation for the differences between body part reports. I haven't received many inquiries about the questions in the CAT. I view that as a positive. I will typically receive questions when the CAT selects functional deficits that are not directly related to what the patient or therapist believes are typical functional activities associated with that region, and for the neck, I don't recall receiving any feedback. My primary request would be to integrate some of the features from the other [FOTO] measures into this one. These would include [functional] staging, as well as the anticipated level of function for a list of activities upon completion of that stage. Without the predictive levels of function, the utility of the report as a tool is diminished. Is that something that you are working on integrating as part of this project?" (Note: We responded that functional staging for the Neck PRO-PM is ready and is awaiting programming development time.) (BK – Illinois)
4. "It's going smoothly….The Neck CAT [PROM] definitely performs better than some. However," the clinician stated that patients sometimes complain about the question for placing a 25 lb. box on a shelf overhead. We followed up on his comment to ask for more detail about the concern. We learned that the clinician feels it is problematic that the 25 lb. box question is often followed by the question about reaching to work overhead for more than 2 minutes. He said the fact that there are 2 "overhead" questions in a row is redundant and causes some patients to feel frustrated. (SK – Pennsylvania)

4a2.2.2. Summarize the feedback obtained from those being measured.

In summary of the details provided above in 4a2.2.1., providers (i.e., those being measured) seem pleased with the performance of NQF measure 3461. They feel it is relevant to the patients they care for, and they have suggestions for ongoing improvement.

4a2.2.3. Summarize the feedback obtained from other users

In summary of the details provided above in 4a1.1., provider networks are working in partnership with payers with feedback being general positive, particularly with respect to lower costs with quality and efficiency of care.

4a2.3. Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

While the Neck FS PRO-PM is too new yet for revisions, based on FOTO's history with 7 other NQF endorsed PRO-PMs, clinician feedback remains an important ongoing driver in both development and revision phases of PRO-PMs. As one example, a consistent desire for ongoing risk-adjustment model development with consideration of more variables/constructs, post-surgical types, and inclusion of individual comorbidities, together with literature reviews caused us to collect and analyze data related to the new risk adjustment model changes for all PRO-PMs, including the Neck PRO-PM.

The feedback from clinician respondent # 4 in section 4a2.2.1. is an example of a scenario in which the scientific/mathematical nature of the CAT may be functioning correctly, yet the clinical/patient experience may suggest a need for changes. Clinician # 4's feedback is the first time we've received a concern about the 2 overhead questions in the Neck PROM that we are aware of; we will monitor and consider further action should we receive similar feedback from more clinicians and/or patients.

Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b1. Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)

If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

The performance results of the Neck FS PRO-PM (NQF measures 3461) will be evaluated by CMS after data collection/submission is completed for the 2019 MIPS performance year. As of March 15, the FOTO QCDR had 645 clinicians across 69 clinics participating in the 2019 MIPS performance year. It is unknown how many of them will use the Neck QCDR measure, nor how many patients will be included, but given the high prevalence of patients presenting for rehabilitation care for neck impairments, we anticipate strong sample sizes of patient episodes of care.

Further, following a longer data collection period within the FOTO system, we will be able to examine improved quality over time. For example, previous analyses of other NQF endorsed FOTO PRO-PM's have suggested trends toward clinician improvement in performance over time with using PROM and PRO-PM data in clinical care.

4b2. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4b2.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.

We did not detect any unexpected findings during implementation of this measure.

4b2.2. Please explain any unexpected benefits from implementation of this measure.

We did not detect any unexpected benefits from implementation of this measure.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

No

5.1a. List of related or competing measures (selected from NQF-endorsed measures)

5.1b. If related or competing measures are not NQF endorsed please indicate measure title and steward.

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

5a.1. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):

Are the measure specifications harmonized to the extent possible?

5a.2. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.

NA

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);

OR

Multiple measures are justified.

5b.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

NA

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

No appendix **Attachment:**

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Focus on Therapeutic Outcomes

Co.2 Point of Contact: Deanna, Hayes, deannahayes@fotoinc.com, 800-482-3686-230

Co.3 Measure Developer if different from Measure Steward: Focus on Therapeutic Outcomes

Co.4 Point of Contact: Deanna, Hayes, deannahayes@fotoinc.com, 800-482-3686-230

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Dennis Hart, PT, PhD (Focus on Therapeutic Outcomes) and Ying-Chih Wang, OTR, PhD (University of Wisconsin-Milwaukee, Milwaukee, WI) were the original developers of this measure. Dr. Hart died in 2012. The expert panel for continued development, analysis, maintenance and re-submission, include: Daniel Deutscher, PT, PhD (Maccabi Healthcare Services, Tel Aviv, Israel); Deanna Hayes, PT, DPT, MS (FOTO); Mark Werneke, PT, MS, Dip MDT (FOTO); Karon Cook, PhD (Northwestern University, Chicago, IL); Michael Kallen, PhD, MPh (Northwestern University, Chicago, IL); and Jerome Mioduski, MS (FOTO).

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2016

Ad.3 Month and Year of most recent revision: 01, 2016

Ad.4 What is your frequency for review/update of this measure? 3-6 years

Ad.5 When is the next scheduled review/update for this measure? 01, 2020

Ad.6 Copyright statement: copyright ©2019. Focus on Therapeutic Outcomes, Inc. All rights reserved.

Ad.7 Disclaimers: NA

Ad.8 Additional Information/Comments: When I tried to edit the Authorized Users above, and I tried on more than one day, I got the following message:

This site can't be reached i15.qualityforum.org's server IP address could not be found.

Construct Validity of performance score level

Additional information following the Scientific Methods Panel review

NQF Measure 3461: Functional Status Change for Patients with Neck Impairments

Background: The scientific methods panel requested additional information supporting the construct validity of NQF measure 3461. Specifically, the panel's request was to assess a correlation at the performance level between the submitted measure and an external measure that assesses a similar construct.

Data from two external patient-reported outcome measures were available to us: the patient-reported global rating of change (GROC) [6] assessed at discharge, and the neck disability index (NDI) [8] as change from admission to discharge.

The GROC used for this analysis is a 15-point scale administered at follow up or discharge. It includes one question on the degree of change (-7 to +7), with zero representing no change. It is often used as an external anchor to estimate a minimal clinically important improvement threshold for the scale of interest.[2, 9] Here, the GROC was used as an external measure to assess construct validity of the provider score level of NQF measure 3461.

The NDI is a widely accepted legacy measure of patient-reported functional status including 10 items related to neck function. Scores range from 0 to 100 with higher scores representing lower functioning levels and higher disability. The NDI has been reported to demonstrate insufficient unidimensionality,[4] moderate responsiveness [5] and a large floor effect.[4, 7] However, its clinical utility has been supported,[1] thus appropriate as an external measure for construct validity assessment of the NQF 3461 provider scores.

Method: We tested validity at the score level by generating Pearson correlation coefficients of mean risk-adjusted residual scores (actual change minus predicted change using the risk-adjusted model) of provider scores using NQF 3461 with GROC and NDI mean scores. Correlations were tested at the clinic and clinician level. Correlations of 0.3 to 0.5, 0.5 to 0.7, and 0.7 to 0.9 were interpreted as supporting low, moderate and high levels of construct validity, respectively.[3] Due to the scale direction, a positive correlation with the GROC, and a negative correlation with the NDI, were expected. We hypothesized that the NQF 3461 measure would be strongly correlated with both external measures examined.

A testing sample was selected separately for each external measure and included patients that had responded to both the NQF 3461 and at least one of the external measures. Since the validity correlations were tested for the provider level (clinics and clinicians), only data from providers who met the threshold used for all other provider-level testing were included [i.e., *clinicians with 10+ patients per calendar year for the clinician level, and clinics with 10+ patients per clinician per calendar year for small clinics (up to 4 clinicians) or 40+ patients per calendar year for large clinics (5 or more clinicians) for the clinic level*].

Results: For NQF measure 3461 correlations with the GROC, a sample of 967 clinics and 3,206 clinicians were included. For NQF measure 3461 correlations with the NDI, a sample of 81 clinics and 267 clinicians were included. Absolute correlations for the two measures and provider levels ranged from 0.64 to 0.73 (see table below) and were highly significant (P<0.001).

Correlation with provider level Neck FS-CAT PRO-PM risk-adjusted outcomes (residuals)

	mean patient-reported GROC* at discharge		**mean NDI** change (discharge-admission)**	
Provider level	**Clinic**	**Clinician**	**Clinic**	**Clinician**
N Patients	78,224	72,315	5,918	5,463
N Clinics	**967**	1,451	**81**	133
N Clinicians	4,960	**3,206**	452	**267**
N States	45	46	25	33
Pearson correlation coefficient	**0.68**	**0.64**	**-0.73**	**-0.69**

*GROC; global rating of change

**NDI; neck disability index (negative change represents a positive outcome)

Interpretation: The correlations reported above were interpreted as moderate to high,[3] confirming our hypothesis that that NQF measure 3461 would be strongly correlated with both external measures, supporting its construct validity.

References:

1. Childs JD, Cleland JA, Elliott JM, et al. Neck pain: Clinical practice guidelines linked to the International Classification of Functioning, Disability, and Health from the Orthopedic Section of the American Physical Therapy Association. *J Orthop Sports Phys Ther*. 2008;38:A1-A34.
2. Deutscher D, Cook KF, Kallen MA, et al. Clinical Interpretation of the Neck Functional Status Computer Adaptive Test. *J Orthop Sports Phys Ther*. 2019;Accepted:
3. Hinkle DE. *Applied statistics for the behavioral sciences*. 5th ed. Boston: Houghton Mifflin; 2003.
4. Hung M, Cheng C, Hon SD, et al. Challenging the norm: further psychometric investigation of the neck disability index. *Spine J*. 2015;15:2440-2445.

5. Hung M, Saltzman CL, Voss MW, et al. Responsiveness of the Patient-Reported Outcomes Measurement Information System (PROMIS), Neck Disability Index (NDI) and Oswestry Disability Index (ODI) instruments in patients with spinal disorders. *Spine J*. 2019;19:34-40.

6. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials*. 1989;10:407-415.

7. Moses MJ, Tishelman JC, Stekas N, et al. Comparison of Patient Reported Outcome Measurement Information System With Neck Disability Index and Visual Analog Scale in Patients With Neck Pain. *Spine (Phila Pa 1976)*. 2019;44:E162-E167.

8. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. *J Manipulative Physiol Ther*. 1991;14:409-415.

9. Wang YC, Hart DL, Werneke M, Stratford PW, Mioduski JE. Clinical interpretation of outcome measures generated from a lumbar computerized adaptive test. *Phys Ther*. 2010;90:1323-1335.