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: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0425

Corresponding Measures:

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

Co.1.1. Measure Steward: Focus on Therapeutic Outcomes, Inc

De.3. Brief Description of Measure: This is a patient-reported outcome performance measure (PRO-PM) consisting of an item response theory-based patient-reported outcome measure (PROM) of risk-adjusted change in functional status (FS) for patients aged 14 years and older with low back impairments. The change in FS is assessed using the Low Back FS PROM. The measure is adjusted to patient characteristics known to be associated with FS outcomes (risk adjusted) and used as a performance measure at the patient, individual clinician, and clinic levels to assess quality. Scores are reported on a 0 to 100 continuous scale with higher scores indicating better FS. The Low Back 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: Patients with low back impairments with functional status deficits are very common in rehabilitation therapy. Functional deficits affect large numbers of people leading to substantial morbidity, high resources use, severity of illness and is a leading cause of poor quality of life for patients that negatively affects society In addition, functional status deficits may severely impact people of any age. Therefore, functional status change measurement during rehabilitation treatment is an important construct.

The Low Back FS PROM was designed to assess functional status and change in functional status in patients with low back impairments. Improved function is a primary goal of therapy for low back pain across the world. The primary purposes of the physical therapy profession according to the Guide to Physical Therapist Practice (American Physical Therapy Association. 2001) include enhancing physical functional abilities, restoring, maintaining, and promoting optimal physical function, wellness, fitness, and optimal quality of life as it relates to movement and health. The World Confederation for Physical Therapy has a similar purpose described in the Declarations of Principle and Position Statements (1999) that emphasizes the importance of the activities and participation component of the International Classification of Functioning, Disability and Health (ICF) (World Health Organization 2001). Therefore, functioning, as described by a patient's ability to perform and participate in different physical and social activities, is important when establishing treatment goals for patients attending physical therapy. The Guide offers clear recommendations for assessing functional status by physical therapists, but the recommendation is applicable to other types of providers treating patients with functional deficits.

The Low Back Patient Reported Outcome-Performance Measure (PRO-PM) begins with the patient reported status of function at the onset of care (intake). The specifics of the deficiency in function, reported by the

patient, provides data for the clinician to analyze and incorporate into the development of the plan of care by setting specific functional goals.

Repeated PROM assessments can assist the clinician in verifying the effectiveness of the plan of care implemented, or, conversely the need to adjust the plan of care to improve effectiveness.

The final measure quantifies the patient's perception of function at the end of care i.e., at discharge from rehabilitation services. Because the measure relies on patient self-report, the functional status outcomes measures are patient-centered and reflect the patient's perceived functional ability.

The measure of functional status change collected during rehabilitation is, by definition, an outcome measure of effectiveness or quality associated with the treatment provided.bMonitoring of aggregated clinician and clinic performance derived from the risk adjusted, aggregated outcome data (of all patients treated by a clinician or clinic) can be used to monitor quality and identify quality improvement to elevate the effectiveness of care for a specific provider. Thus, measurement of effectiveness of care for patients with low back impairments can help to promote quality, improve accountability, and ultimately reduce practice variation and enhance outcomes of care across therapy providers.

Low back pain is a heterogeneous condition, with little consensus on its diagnosis, classification or treatment. (Fourney, Andersson et al. 2011) Although a variety of clinical practice guidelines for the care of patients with low back pain exist, (Arnau, Vallano et al. 2006 ; Delitto, George et al. 2012) guidelines are not are not universally accepted or followed. Preliminary evidence suggests that implementation of evidenced based guidelines may result in improved patient functional status. (Rutten, Degen et al. 2010) However, the multitude of Clinical Guidelines developed in the US and abroad and promulgated by a variety of professional associations and societies may contribute to the variation in care. (Jackson, Hettinga et al. 2009; Chilibeck, Vatanparast et al. 2011; Guevara-Lopez, Covarrubias-Gomez et al. 2011; Rudwaleit and Marker-Hermann 2012) Yet, most guidelines lack specificity of detail for physical therapy practice. (Ladeira 2011)

It is widely known that high degrees of variability in the pathways of care for patients with low back impairment exist both within and between geographic regions, with variability in adherence to evidence based guidelines, use of invasive procedures, opioid use and advanced imaging. (Cherkin, Deyo et al. 1994 ; Deyo and Mirza 2006 ; Friedly, Chan et al. 2008 ; Webster, Cifuentes et al. 2009) (Ferguson, Holdsworth et al. 2010) Prior research also shows variation in risk-adjusted functional status outcome by clinic and therapist. (Resnik, Feng et al. 2006; Resnik, Liu et al. 2008)

S.4. Numerator Statement: The numerator is based on residual scores (actual change scores - predicted change after risk adjustment) of patients receiving care for Low Back impairments and who completed the Low Back PRO-PM.

The numerator, as it applies to the 3 levels, is defined as follows:

Patient Level: The residual functional status score for the individual patient with a low back impairment.

Individual Clinician Level: The average of residuals in functional status scores in patients who were treated by a clinician in a 12-month time period for a low back impairment.

Clinic Level: The average of residuals in functional status scores in patients who were treated by a clinic in a 12-month time period for a low back impairment.

S.6. Denominator Statement: The target population is all patients 14 years and older with a Low Back impairment who have initiated an episode of care and completed the Low Back FS PROM.

S.8. Denominator Exclusions: Patients who are not being treated for a Low Back 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: Jul 31, 2008 **Most Recent Endorsement Date:** Jul 07, 2015

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? N/A

Preliminary Analysis: Maintenance of Endorsement

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence

Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

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

- This is a patient-reported outcome performance measure (PRO-PM) consisting of an item response theory-based patient-reported outcome measure (PROM) of risk-adjusted change in functional status (FS) for patients aged 14 years and older with low back impairments. The change in FS is assessed using the Low Back FS PROM. The measure is adjusted to patient characteristics known to be associated with FS outcomes (risk adjusted) and used as a performance measure at the patient, individual clinician, and clinic levels to assess quality. Scores are reported on a 0 to 100 continuous scale with higher scores indicating better FS. The Low Back 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.
- To demonstrate evidence of a structure, process, intervention, or service that can influence the outcome of interest, the developer analyzed the relationship between their measure's score at discharge (the outcome) compared to the clinical process of administering the PROM assessment within the first two weeks of patient care.
- The developer matched patients using propensity scores to compare means/rates of included variables between patients both with and without an early interim assessment (the intervention). Only clinicians with at least 10 completed episodes in the 2016 were factored in. Data was aggregated at the clinician level to allow the developer to assess the relationship between early interim assessments and functional status outcomes at the PRO-PM provider score level.

- Patients who were treated by clinicians with high rates of early interim assessment (n=2,451) increased their functional status points by 2.5 at discharge over patients who were treated by clinicians with low rates of early interim assessment (n=2,273, p<0.001)

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates:

Question for the Committee:

- *The developer has provided testing that shows the effect of an intervention (interim assessment of the outcome using the same PROM instrument) on the outcome of interest. Does this meet the NQF criteria for evidence?*
- *This measure is derived from patient report. Does the target population value the measured outcome and find it meaningful?*

Guidance from the Evidence Algorithm

Measure assesses outcome (box 1) YES → relationship between outcome and at least one healthcare action (box 2) YES → PASS (Evidence Algorithm from pg. 15 of NQF Measure Evaluation Criteria)

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.

- The developer used three quality categories as well as deciles for clinicians and clinics to demonstrate performance gap, with a significant performance spread.
  - Quality categories for clinicians (figure 2b4.2iia):
    - Low performance: 18% of clinicians
    - Average performance: 68%
    - High performance: 14%
  - Quality categories for clinics (figure 2b4.2iiia):
    - Low performance: 29%
    - Average performance: 52%
    - High performance: 19%
- Difference in mean residual scores between the 1st and 10th decile for clinicians and clinics also showed a range of performance socres
  - Clinician performance gap by decile (table 2b4.2iib):
    - 1st decile: -7.1
    - 10th decile: 7.6
    - From 2016-2018 performance gap between 1st and 10th decile was -7.6 to 7.9 (2016) and -7.0 to 8.8 (2018)
  - Clinic performance gap by decile (table 2b4.2iiib):

- 1st decile: -6.2
- 10th decile: 6.3
- From 2016-2018 performance gap between 1st and 10th decile was -6.8 to 6.6 (2016) and -5.8 to 7.6 (2018)

Disparities

- The developer provides three sets of information each spanning including multiple years, regarding: age, gender, and insurance status (includes: indemnity, Medicaid, Medicare, HMO/PPO, and Workers Comp.)
    - 2002-2004: n=1285, age (Mean±SD): 46±16, gender (percent female): 59, largest payer by percentage: HMO/PPO (41%)
    - 2007-2008: n=17,439, age (Mean±SD): 51±17, gender (percent female): 60, largest payer by percentage: HMO/PPO (48%)
    - 2014-2016: n=414,125, age (Mean±SD): 57±16.8, gender (percent female): 60, largest payer by percentage: HMO/PPO (42%)
- Risk adjustment models over time (beta coefficient by submission) – the developer notes that though some beta coefficients vary in size, in general, the beta coefficients are stable over time
- Education level analysis did not show an important contribution to predicint FS at discharge (controlling for all other variables which were already included in the risk-adjustment model).

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?

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

Committee Pre-evaluation Comments:

Criteria 1: Importance to Measure and Report (including 1a and 1b)

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.

Insufficient evidence

It is unclear from the data provided whether autocorrelation within patients has been accounted for and whether a pre-post difference in functional status of 2.5 points is clinically meaningful.

Sufficient evidence

As a maintenance measure, I am not aware of any new information that changes the evidence base. As the measure is reported at the individual patient level (as well as provider and clinic)- I believe it demonstrates value to the target population.

Evidence is good, especially if assessment is given early in treatment and again at discharge.

The developer provided the following updates to the evidence of this patient-reported outcome performance measure (PRO-PM), which consist of an item response theory-based patient-reported outcome measure (PROM) of risk-adjusted change in functional status (FS) for patients aged 14 years and older with low back impairments. The change in FS is assessed using the Low Back FS PROM and the measure is adjusted to

patient characteristics known to be associated with FS outcomes (risk adjusted) and used as a performance measure at the patient, individual clinician, and clinic levels to assess quality. Scores are reported on a 0 to 100 continuous scale with higher scores indicating better FS. The Low Back 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. To demonstrate evidence of a structure, process, intervention, or service that can influence the outcome of interest, the developer analyzed the relationship between their measure's score at discharge (the outcome) compared to the clinical process of administering the first interim PROM assessment within the first two weeks of patient care. The developer matched patients using propensity scores to compare means/rates of included variables between patients both with and without an early interim assessment (the intervention). Only clinicians with at least 10 completed episodes in the 2016 were factored in. Data was aggregated at the clinician level to allow the developer to assess the relationship between early interim assessments and functional status outcomes at the PRO-PM provider score level. Patients who were treated by clinicians with high rates of early interim assessment (n=2,451) increased their functional status points by 2.5 at discharge over patients who were treated by clinicians with low rates of early interim assessment (n=2,273, $p<0.001$) Data provided by the developer also indicates the majority of patients found all or most of the items meaningful to them.

Demonstrated through patient survey

Does this meet the NQF criteria for evidence? As a first time reviewer, I will abstain from comment. I would like to learn more from my peers. This measure is derived from patient report. Does the target population value the measured outcome and find it meaningful? There is no data to evaluate at the individual patient level if the outcome measurement meaningful, is a personal assessment. From a macro perspective outcome results as a function of treatment is critical. As stated, " Thus, measurement of effectiveness of care for patients with low back impairments can help to promote quality, improve accountability, and ultimately reduce practice variation and enhance outcomes of care across therapy providers." With the variability in treatment and perceived benefits, evidence to evaluate care, is a tool to capture performance and provide learning across systems of care.

No concerns

Pass

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?

Uncertain; Disparities not investigated

It is not clear from the data provided how the standard error of measurement (appears to be >5 points in the highest and lowest quartiles) was taken into account when presented the categories of low, average and high performance for data provided in 1b.

Gap identified

Yes, there was performance data provided and yes it demonstrated opportunities for improvement. Data by specific sub populations was provided, however they only included age, gender and insurance status. I am curious if the original measure tested the impact of additional factors such as ethnicity and income.

Disparity information wasn't really there. Developer is looking at how to delve into SDOH issues going forward. However, Gap rating was High.

"There is opportunity for improvement in light of the reported data. The developer used three quality categories for clinicians and clinics relative to performance levels of low, average, or high. While there was a considerable spread in the performance scores, the majority of the scores were in the average performance category. Relative to disparities, the developer provides three sets of information each spanning multiple years, regarding: age, gender, and insurance status (includes: indemnity, Medicaid, Medicare, HMO/PPO, and

Workers Comp.) Requested is clarification of the context for how the findings presented below should be interpreted in relation to disparities.

2002-2004: n=1285, age (Mean±SD): 46±16, gender (percent female): 59, largest payer by percentage: HMO/PPO (41%)

2007-2008: n=17,439, age (Mean±SD): 51±17, gender (percent female): 60, largest payer by percentage: HMO/PPO (48%)

2014-2016: n=414,125, age (Mean±SD): 57±16.8, gender (percent female): 60, largest payer by percentage: HMO/PPO (42%)

Education level analysis did not show an important contribution to predict FS at discharge (controlling for all other variables which were already included in the risk-adjustment model)."

Adequate peformance gap demonstrated at clinic/clinician level; data on disparities was unclear

As a new reviewer, I would like to understand how my peers interpret the data. From a macro point of view, we need consistency, and this would warrant a national performance measure.

No concerns

Moderate

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.

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 the call. A summary of the measure and the Panel discussion is provided below.

Scientific Methods Panel Votes: Measure Passes

- Reliability: H-3, M-1, L-0, I-1
- Validity: H-4, M-1, L-0, I-0

Reliability

- Measure passed SMP with a High for reliability
- Cronbach's alpha and IRT person reliability
- Low Back FS PROM had very good internal consistency
- SMP assessment of results:
    - Signal to noise reported average reliability at the clinician level of 0.71
    - Clinicians reporting 10-19 cases: average reliability score was 0.71, with 42 percent of clinicians having reliability below 0.71
    - Clinicians reporting 20-29 cases: average reliability score was 0.77, with 28 percent of clinicians having reliability below 0.7
    - Other clinician categories and clinics: average reliability score was above 0.8, with 14 percent having reliability below 0.7
    - While the Adams paper that outlines reliability estimation for signal to noise puts for an ad hoc standard of 0.7, it also notes that at this level this a substantial level of misclassification, and as a committee, we have observed levels of misclassification for measures with 0.7 reliability that raised concerns about whether the measure had adequate reliability for use.
    - Average reliability score at the clinic level was 0.84.
- Note for measure developer: Patient level element reliability assessed on estimates of two versions of instrument. Some patient reliability testing appeared more directed toward testing validity and there is no reported direct comparison of consistency about patient level reliabity of instrument.
- Note for measure developer: Opportunities exist to improve upon the instructions of the survey to increase accuracy of responses due to "patients asking about how to complete items where activity is not routine".

Validity

- Measure passed SMP with a High for validity.
- Psychometric testing was conducted for instrument content and construct validity.
- External markers used to validate the measure score for measure score validity.
- IRT testing of consistency and unidimensionality of the Guttman scaling of the survey questions.
- Data element and score validity are shown in the validity testing evidence.
- Testing showed high face validity and multiple levels of validity evidence includent content and structural.

- The distribution of residuals between actual and predicted end FS are not directly presented, but can be inferred from Table 2b4.2i and Figure 2b4.2iia. The SD is about 13 and the IQ from Figure 2b4.2iia looks to be about 9. The range of variation in the percentage with meaningful patient reported improvement between the 2nd and 9th deciles of distribution of residuals is 58.9% to 80.8%. These look like meaningful differences but should be discussed by the substantive committee.
- The Methods Panel noted wanting to see results stratified by individuals who completed the form themselves versus those who had a proxy complete the form. Specifically, the Methods Panel wanted to see was the proxy a family member or employee of a rehab center.
- The Methods Panel was unsure if exclusions resulted in bias in the measurement: "For example, if patient did not have access to internet or were less computer savvy they may not complete the tool unless they received assistance, or patients with certain social risk factors may not complete the tool at two points in time or even complete their rehabilitation". Additionally, social risk factors were not assessed and determined if there was impact on completion rates.
- Methods Panel comments on Risk-adjustment:
    - More can and should be done on SES testing. The measure with the largest impact on predicted FS is acuity, measured by the time between onset of symptoms and initial evaluation. There is no discussion of the sources of or causes for the time between onset of symptons and initial evaluation. The basis for assessing education as a risk factor was fundamentally flawed. It treated each eduation level as an independent variable, when what should have been assessed was 1) the proportion of variance explained by all the education measures collectively, and 2) whether there was a consisten pattern of improvement in outcomes as educational levels of patients increased.
    - The measure developer said "we posit that the traits of having Medicaid or Medicare B under age 65 serve as proxy variables for socioeconomic factors." No rationale or literature is provided to support this cliam. While these measures are included in the model they also represent payer characteristics and do not likely capture the full effect of sodial determinants of health. Outside of using the payer characteristics, they did not adjust for social risk factors or for patients who were cognitively impaired or had language barriers and thus had someone else complete the PRO form**.** Concern this may introduce bias in results for those not able to complete the form, and this was not accounted for in the risk adjustment model

Questions for the Committee regarding reliability:

- Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?
- The Scientific Methods Panel is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?

Questions for the Committee regarding validity:

- Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?
- The Scientific Methods Panel is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?

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

Combined Methods Panel Scientific Acceptability Evaluation

Measure Number: 0425

Measure Title: Functional Status Change for Patients with Lumbar 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** ☒ **Other Panel Member #3:** proprietary clinical registry database

Level of Analysis:

☒ **Clinician: Group/Practice** ☒ **Clinician: Individual** ☐ **Facility** ☐ **Health Plan**
☐ **Population: Community, County or City** ☐ **Population: Regional and State**
☐ **Integrated Delivery System** ☒ **Other Panel Member #3:** patient level

Measure is:

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

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.**

   Panel Member #2: No substantial concerns. Some discussion of patients asking about how to complete items where activity is not routine, suggesting opportunities for improving instructions to increase accuracy of responses.

   Panel Member #3: The denominator statement does not define "who have initiated an episode of care"? How do they identify this population? The ICD codes for low back impairment are listed. Also "and who completed the low back FS PROM". Don't they mean at least twice? Does it have to be at initiation of the episode (admission?) and at discharge? How are those two specific patient reported outcome surveys identified?

   Panel Member #4: I have a few minor concerns about the measure specifications. Specifically, in the sampling section S.15 there are instructions for patients less than and 8 years old and for those over 8. I found this confusing since the measure description in DE3 indicated that the measure is to be used on those 14 and over.

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**

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**

6. **Assess the method(s) used for reliability testing Panel Member #1:**

   Submission document: Testing attachment, section 2a2.2

   Panel Member #1: The developer conducted extensive reliability testing for both data element and measure score with appropriate methods.

   Avearge measure score reliability is very good at both clinic and clinician level.

   Panel Member #2: Patient level element reliability was assessed based on consistency of estimates of two versions of instrument. Some of patient reliability testing appeared to be more directed toward testing validity and there is no reported direct comparison of consistency of responses by patients within same time frame, i.e., test-retest reliability. That said, I'm not concerned about patient level reliability of instrument.

   Panel Member #3: The methods are appropriate.

   Panel Member #4: Cronbach's alpha and IRT person reliability are appropriate.

   Panel Member #5: Provider to provider variance assessed using HLM

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

   Submission document: Testing attachment, section 2a2.3

   Panel Member #1: The developer showed that the Low Back FS PROM has very good internal consistency, the expanded took has even better internal consistency than the origian l25-item instrument.

   Panel Member #2: Reliability of testing at patient level consistent with adequate reliability to use instrument.

   Signal to noise testing reported average reliability at clinician level of .71, . For clinicians reporting 10-19 cases, average reliability was .71, with 42% of clincians reporting 10-19 cases having reliability below .. For clinicians with 20-29 cases, average reliability was .77, with28% cases having reliability below .7. For all other clinician categories and all clinic categories, average reliability was above .8 and proportion below 0.7 14% or less.

   While the Adams paper that outlines reliability estimation for signal to noise puts for an ad hoc standard of 0.7, it also notes that at this level this a substantial level of misclassification, and as a committee, we have observed levels of misclassification for measures with 0.7 reliability that raised concerns about whether the measure had adequate reliability for use. **I would like a discussion of whether this measure reaches the levels of reliability the committee expects for assessing the performance of clinicians with 10-19 and 20-29 cases.**

   Panel Member #3: Test results show high reliability. At the clinic level, the average reliability for clinics meeting the FOTO threshold of number of patients per clinic for quality reporting was 0.84. At the clinician level, average reliability for clinicians with 10 or more patients per calendar year was 0.71.

   Panel Member #5: High level of reliability – better after addition of new items.

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.

   Submission document: Testing attachment, section 2a2.2

   ☒ **Yes Panel Member #2:** …BUT PROPORTION OF VARIANCE NOT REPORTED.

   ☐ **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?

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 all testing results):

    ☒ **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 specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

    ☒ **Insufficient** (NOTE: Should rate INSUFFICIENT 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.**

    Panel Member #1: The reason for overall moderate rating is that measure reliability score is high only for clinic or clinician with high volume.

    Panel Member #2: I repeat the comment I made under question 7 above:

    Reliability of testing at patient level consistent with adequate reliability to use instrument.

    Signal to noise testing reported average reliability at clinician level of .71, . For clinicians reporting 10-19 cases, average reliability was .71, with 42% of clincians reporting 10-19 cases having reliability below .. For clinicians with 20-29 cases, average reliability was .77, with28% cases having reliability below .7. For all other clinician categories and all clinic categories, average reliability was above .8 and proportion below 0.7 14% or less.

    While the Adams paper that outlines reliability estimation for signal to noise puts for an ad hoc standard of 0.7, it also notes that at this level this a substantial level of misclassification, and as a committee, we have observed levels of misclassification for measures with 0.7 reliability that raised concerns about whether the measure had adequate reliability for use. **I would like a discussion of whether this measure reaches the levels of reliability the committee expects for assessing the performance of clinicians with 10-19 and 20-29 cases.**

    Panel Member #3: Comprehensive testing results show high reliability.

    Panel Member #4: No concerns.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    Submission document: Testing attachment, section 2b2.

    Panel Member #1: My concern is actually for lack of exclusion. In some testings, very old patients were included. For example, in Testing form Table 1.6 VII, it shows that patient as old as 116 was included.

    Panel Member #2: None

    Panel Member #3: They don't list as a specific exclusion but obviously patients without 2 assessments (those who didn't complete rehab or didn't complete both admission and completion assessment) are excluded. They did some anlaysis of these exclusion but I believe it was incomplete (see my response under 15 below).

    Panel Member #4: I have no concerns with exclusions.

Although I know this is not what the question is asking here… In section 2b1.3ii, I would have liked to have seen the items that were removed and their fit statistics/residual correlations as well as where they fit on the latent construct (person/item map).

Panel Member #5: None – submitter provided testing no bias with exclusions.

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

    Submission document: Testing attachment, section 2b4.

    Panel Member #2: The distribution of residuals between actual and predicted end FS are not directly presented, but can be inferred from Table 2b4.2i and Figure 2b4.2iia. The SD is about 13 and the IQ from Figure 2b4.2iia looks to be about 9. The range of variation in the percentage with meaningful patient reported improvement between the 2nd and 9th deciles of distribution of residuals is 58.9% to 80.8%. These look like meaningful differences but should be discussed by the substantive committee.

    Panel Member #3: The physician/clinic level scores are increasing quite a bit over time for those with data over 3 year period. It is not clear these providers are really producing dramatically improved outcomes in their patients or how the completion of the forms may have changed based on whether the patient or a proxy completed the assessment.

    Panel Member #4: No concerns. I was pleased to see an assessment of floor and ceiling effects in 2b1.2v and 2b1.3ii.

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.
    Panel Member #2: This is a paper based version of an assessment otherwise done by computer using an item response approach. Results look comparable.
    Panel Member #3: Would like to see results stratified by those who completed the form themselves versus those for whom a proxy was used, and specifically if that proxy respondent was a family member or employee of rehab center.
    Panel Member #4: No concerns.

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

    Submission document: Testing attachment, section 2b6.

    Panel Member #2: None. Based on analysis presented, data appears to be missing at random.

    Panel Member #3: Not complete testing to determine if patients excluded resulted in bias in measurement. For example, if patient did not have access to internet or were less computer savvy they may not complete the tool unless they received assistance, or patients with certain social risk factors may not complete the tool at two points in time or even complete their rehabilitation (need initiation and discharge survey completed to be included). While they did evaluate similarity of those completing vs not on some patient characteristics, they did not assess social risk factors and impact on completion rates. Also, patients with certain conditions could have someone else complete the form and I do not see testing as to impact of those who had an alternate complete the tool? Did this introduce bias, say if clinician completed and wanted to show good results? This should proably have been included in the risk adjustment model as another variable, or at least tested for significance.

    Panel Member #4: No concerns.

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

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

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

Panel Member #2: Risk adjustment adequately reported and tested.

That said, more can and should be done on SES testing. I would make two comments:

First, the measure with the largest impact on predicted FS is acuity, which is actually measured by the time between onset of symptoms and initial evaluation. There is no discussion of the sources of or causes for the time between onset of symptons and initial evaluation, but to the extent it is due to lack of access to providers, there may be an SES component to this . It's worth more reflection by the developer and committee.

Second, the basis for assessing education as a risk factor was fundamentally flawed. It treated each eduation level as an independent variable, when what should have been assessed was 1) the proportion of variance explained by all the education measures collectively, and 2)whether there was a consisten pattern of improvement in outcomes as educational levels of patients increased.

Panel Member #3: They said "we posit that the traits of having Medicaid or Medicare B under age 65 serve as proxy variables for socioeconomic factors. They do not provide rationale or cite literature to support this cliam. While these measures are included in the model they also represent payer characteristics and do not likely capture the full effect of sodial determinants of health. They did some testing of education level but with 9 response categories only bachelors degree was significant. Further testing is warranted using collapsed groups.

Outside of using the payer characteristics, they did not adjust for social risk factors or for patients who were cognitively impaired or had language barriers and thus had someone else complete the PRO form**.** Concern this may introduce bias in results for those not able to complete the form, and this was not accounted for in the risk adjustment model (not even fact that someone else completed for the patient).

<u>For cost/resource use measures ONLY</u>:

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**

    Submission document: Testing attachment, section 2b2.2

    Panel Member #1: Standard psychometric testings were conducted for instrument content and construct validitiy. For measure score validity, external markers were used to validate the measure score.

    Panel Member #2: IRT testing of consistency and unidimensionality of the Guttman scaling of the survey questions.

    Correlational consistency of scores with two other measures of improvement.

    Panel Member #3: Used multiple statistical approaches the measure various components of validity.

    Panel Member #4: No concerns.

    Panel Member #5: IRT modeling used – I am not experienced in these methods, but appear to be appropriate in this setting.

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

    Submission document: Testing attachment, section 2b2.3

    Panel Member #1: Validity tests provided supportive evidence for both data element and measure score validity.

    Panel Member #2: IRT score was suffient to establish scaling of instrument.

    Correlational analysis was consistent with measure assessing improvements in status.

    Panel Member #3: Results support strong validity of the measure.

    Panel Member #4: No concerns.

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.**

Panel Member #2: Instrument tested for scale characteristics.

Adequate correlation with other measures.

Face validity high.

No concerns.

Panel Member #3: Results produced multiple levels of validity evidence including content and structural validity.

Panel Member #4: I felt that some items I wanted to evaluate were missing from this report such as: IRT person/item/Keyform maps to evalauate content coverage; and scale functionality information and curves (especially since there are 2 different scales being used in 1 measure). Should the partial credit model have been used since there are 2 different response scales? Also, in 3c.1 it is indicated that the instrument is available in Engligh and Spanish in the US and in several other languages in Israel. I'm assuming that cross crultural equivalence testing was done and I would have liked to see that information so I could asses it. However, all things considered, I was very impressed by the analysis and the level of thought that went into this application. All of the IRT assumtions were addressed and reported.

ADDITIONAL RECOMMENDATIONS

27. **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.**

Panel Member #2: See comment above about assessing what level of change relative to expected is meaningful improvement.

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?*

- Assess different populations, e.g. mixture of acute and chronic LBP, young vs. elderly, surgical vs. non-surgical, medical comorbidities, spinal/orthopedic comorbidities
- The frequency of survey completion appears to be allowed to vary between patients. It is unclear whether the impact of the number of completed surveys on reliability of the patient, clinician and clinic data was accounted for, particularly for clinicians and clinics with greater numbers of patients.
- No concerns.
- Passes on reliability.
- This measure was reviewed and passed by the Scientific Methods Panel with a High for reliability.
- Reliability testing seems adequate
- Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)? Yes, consistency is an issue because of timing and subjective nature of outcome measure. The range of variation in the percentage with meaningful patient reported improvement between the 2nd and 9th decile of distribution of residuals is 58.9% to 80.8%. Patients were not sure how to answer the question when asked about performance for non-routine activities. Instead of activities the measure needs to be on the ability to functional.

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

- Yes as above
- The amount of variance explained by clinicians in this measure appears to be roughly 7% (Table 2a2.3iv), indicating than 93% may be explained by something other than quality of care.
- No concern
- Passes on reliability
- The Scientific Methods Panel noted patient level elements were assessed using two versions of the instrument, which gave the appearance that some testing was aligned more with testing validity. Additionally, there was no reported direct comparison of consistency about patient reliability of the instrument.
- The Scientific Methods Panel is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability? Yes we need to discuss and vote on reliability.

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

- Some as it relates to low reliability.
- I share concerns about who is completing the surveys (patient vs. family member or clinic staff) especially for those completing multiple surveys, vs. those completing only the intake survey.
- Results appear valid
- No concerns.
- No - they tested empiric validity against GROC and ODQ as external assessments
- Yes. Validity testing needs to adequately identify differences in quality. It is unclear if descriptive statistics for ability during activities is consistent enough for a valid measure.

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?*

- Same as above related to lack of adjustment for infrastrctural and patient characteristics
- No Concerns
- Developer analyzes improvement of patient outcomes and clinician performance over time. Developer reports testing results supported that missing data are mostly missing at random and that the risk-adjustment model was not impacted by missing data. No other threats to validy are noted.
- Unclear; although CAT is very useful for assessments with minimal burden it is unclear whether this approach would be applicable to other settings/populations
- The issue that we should discuss is the comparability of performance scores. If there is a recommendation that the measure should be based on activity vs. functional ability, then there has been bias identified. This may be a loop based on the output of other discussions. I support that function-specific questions added to the survey instead of the question relating to “usual hobbies, recreational or sporting activities”. For example bending/stooping, lifingt/carrying groceries, changing ones position, putting on socks and shoes, and standing for one hour.

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?*

- As above previously mentioned challenges with lack of adjustment methods
- Appears adequate.

- I wonder about potential impact of race/ethnicity and SES.
- No concerns
- Exclusions appear to be appropriate. Relative to risk adjustments: There appears to be potential opportunity to examine SES impacts more deeply. A few examples that provide opportunity to gain further insight into the role of SES on this issue include examining: 1) Reasons and/or causes for time intervals between the onset of symptoms and subsequent evaluation of the problem. 2) An alternative approach for assessing the association of education level with patterns of outcome improvement. That is, as noted by a SMP member-- assess the proportion of variance by all education levels collectively and determine whether there is a pattern of improvement as educational levels increase. 3) Issues related to service provider access. 4) Factors affecting treatment compliance (i.e., health literacy). The measure developer's response reflects their interpretation of Medicaid or Medicare B enrollment serving as a proxy for addressing SES, perhaps based on enrollment criteria. This approach would benefit from substantiation by including evidence from the literature to support this claim. It does not appear that the risk adjustment model considered certain subpopulation groups such as those with intellectual disabilities, cognitive impairments or those experiencing language barriers, or the introduction of potential biases introduced by individuals requiring assistance in completing the PRO form.
- Yes, seems appropriately adequate
- It is not clear who is filling in the forms. The social risk factor of having a person ask and input the answers was not clearly addressed.
- Risk adjustment: The only disparities examined were age, gender, education and insurance. One of the SMP members had trouble with the way the education variable was used. Including other patient-level disparities (eg race/ethnicity) would improve this.
- Mod-high rating; some discussion points raised by panel (education); can we get more information?

2c. Composite Performance Measure: *Composite Analysis (if applicable): Do analyses demonstrate the component measures fit the quality construct and add value? Do analyses demonstrate the aggregation and weighting rules fit the quality construct and rationale?*

- I view this as a "composite" of different populations
- No concerns, except to establish whether reliability results are related to the same individual completing the surveys over time and issues with autocorrelation.

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.

- Data generated during the provision of care; all data elements are found in electronic sources, unless provider chooses to print out the "short form" and provide manual administration and scoring of the survey.
- The survey is offered on a web based platform allowing for ease of access by patients either at home or at work. This accessibility has increased participation and reduced missing data.
- Entire survey (PROM and risk-adjustment sections) take five minutes to complete, on average.
- The developer offers clinicians different pricing options for software:
    - Free access to the components needed to calculate a reportable score can be found on FOTO's website.

- FOTO Outcomes Manager Lite services- provides data collection, scoring of PROM's components, patient- and clinician-level reporting for individual patient's results at $20 per clinic/month or $15 per clinician/month.
- FOTO Outcomes Manager services- provides the same services as *Outcomes Manager Lite services* with additional benefits such as, promotion of using patient-reported outcomes to improve quality of care and costs (efficiency). This level comes at a cost of $50 per clinic/month or $25 per clinician/month.

- The developer states that generating one or two new patient referrals can offset the cost of their software.

Questions for the Committee:

- Does the Committee agree with the staff assessment that there are no significant feasibility challenges associated with this 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?

- Relies on iPad data entry and data system, which I believe is/may be owned by FOTO; costs for some and data management processes may be a barrier
- The impact of cognitive impairment on survey completion may limit generalizability across facilities.
- Measure is feasible
- No concerns.
- Concerns were raised around proxy completions, but the developer notes that is a rarely utilized option so probably does not impact data significantly.
- There do not appear to be any significant challenges to implementing the performance measurement without undue burden. The survey is offered on a web-based platform, providing ease of access to patients and only takes 5 minutes to complete. Clinicians are offered reasonable pricing options for software.
- Requires proprietary software; low feasibility unless provider pays for software

The concern for the data collection strategy is the consistency of who is entering the data. It needs to be tracked if the patient is entering the data without oversight or if they are in the office with a clinician assisting them.

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**

OR

Planned use in an accountability program? ☒ **Yes** ☐ **No**

- AIM Specialty Health w/Anthem/BCBS

Accountability program details

- CMS payment program PQRS (2009-2016)
- Merit-based Incentive Program (2017-present)
- The Physical Therapy Provider Network – outcomes bonus program with large health plan partners in multiple states
- Therapy Partners – uses FOTO outcomes in value-baed contracts with payers

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

- Providers receive real time reports for individual patient results. This includes PROM scores, PRO-PM comparions of scores and end-of-episode results, and patients' responses to functional assessment questions.
- Measure developer notes that clinicians value the use of PROM and PRO-PM data to foster better understanding of the patient's perspectice, aid in goal-setting, and to assist in treatment and discharge planning.
- Measure developer notes clinicians have frequently stated that smoking and pregnancy be factors added into the risk-adjustment model.
- Multiple clinicians indicated that they would prefer to see function-specific questions added to the survey instead of the question relating to "usual hobbies, recreational or sporting activities". Some of the offered function-specific items are: bending/stooping, lifingt/carrying groceries, changing ones position, putting on socks and shoes, and standing for one hour.

Additional Feedback: N/A

Questions for the Committee:

- How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?
- How has the measure been vetted in real-world settings by those being measured or others?

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

- The measure developer provided performance data over a three-year time period (not broken down year over year) for clinicians and clinics:
    - Clinicians' mean residual scores using one-way ANOVA ($p<0.001$) with a monotonic analysis showed a statistically significant increase from -0.3 to +1.0.
    - Clinics' mean residual scores using one-way ANOVA ($p<0.001$) with a monotonic analysis showed a statistically significant increase from -0.3 to +1.2.
- The three-year time period data demonstrates that providers improve their ability to make decisions about patient care and enhance patient engagement as they gain skills over time using risk-adjusted PROM data.
- Performance results of the measure will be evaluated by CMS at the conclusion of the data collection/submission period for the 2019 MIPS performance year

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 received notice of a patient having a military application denied due to the words "Hepatitis, HIV, or AIDS" in their medical record. Because of this, the developer is looking into modifying or removing this specific item from the measure.

Potential harms

- None found

Additional Feedback: N/A

***Questions for the Committee**:*

- 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 and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

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?

Would not endorse this measure for public reporting at this time. Needs field testing for the impact of comparative non-risk adjusted results to determine discriminatory accuracy

Attribution to the clinician or clinic appears weak.

Use case acceptable

Used in Merit and Payment programs. Also used, most importantly, in patient goal setting and treatment planning.

"Developer indicates the following public reporting activities: Planned use in an accountability program?

AIM Specialty Health w/Anthem/BCBS Accountability program details

CMS payment program PQRS (2009-2016)

Merit-based Incentive Program (2017-present)

The Physical Therapy Provider Network – outcomes bonus program with large health plan partners in multiple states

Therapy Partners – uses FOTO outcomes in value-based contracts with payers. In terms of feedback: the developer provides the following:

1) Providers receive real time reports for individual patient results. This includes PROM scores, PRO-PM comparisons of scores and end-of-episode results, and patients' responses to functional assessment questions.

2) Measure developer notes that clinicians value the use of PROM and PRO-PM data to foster better understanding of the patient's perspective, aid in goal-setting, and to assist in treatment and discharge planning.

3) Measure developer notes clinicians have frequently stated that smoking and pregnancy be factors added into the risk-adjustment model.

4) Multiple clinicians indicated that they would prefer to see function-specific questions added to the survey instead of the question relating to "usual hobbies, recreational or sporting activities". Some of the offered function-specific items are: bending/stooping, lifting/carrying groceries, changing one's position, putting on socks and shoes, and standing for one hour."

Measure is in use in MIPS

I support that function-specific questions added to the survey instead of the question relating to "usual hobbies, recreational or sporting activities". For example bending/stooping, lifingt/carrying groceries, changing ones position, putting on socks and shoes, and standing for one hour.

No concerns

rating moderate; feedback incorporated

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.

Unclear in terms of the differences between intended and actual usage

Attribution to the clinician or clinic is of concern.

Measure is usable

It appears as though a patient had a military application denied due to information disclosed through reporting the PROM

Developer noted one item may be removed as the data may be used against potential patients (Hepatitis, HIV or AIDS question). Otherwise passes Usability

Providers (clinicians and clinics) demonstrated significant improvement in their performance over time. These findings support that providers may better learn and gain skills over time for using risk-adjustment PROM data in the context of everyday data-driven clinical decision making with the patient at the center. Providers may improve their ability to use the data to enhance their communication with the patient, promote patient engagement. Using risk-adjusted patient-reported outcome measures (PROMs) of function promotes a focus on patient-perceived function and encourages meaningful discussions about goals and expectations for the results of the care episode. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations appear to outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists). However, one unexpected

finding (positive or negative) during implementation included the developer's receipt of a notice from a patient having a military application denied due to the words "Hepatitis, HIV, or AIDS" in their medical record. As a result of this, the developer is looking into modifying or removing this specific item from the measure.

Benefits seem to outweigh risks, with exception of proprietary nature of product

The prefromance results can be used to further the goal of high-quality, efficient healthcare by standardizing measure and deploying thoughout the system based on function. Deviations in results will provide an opportunity for learning and best practice sharing.

No concerns; moderate-high

Criterion 5: Related and Competing Measures

Related or competing measures

This measure is related to, but not competing with, the following measures:

- 0422: Functional status change for patients with Knee impairments
- 0423: Functional status change for patients with Hip impairments
- 0424: Functional status change for patients with Foot and Ankle impairments
- 0426: Functional status change for patients with Shoulder impairments
- 0427: Functional status change for patients with elbow, wrist and hand impairments
- 0428: Functional status change for patients with General orthopaedic impairments
- 3461: Functional status change for patients with Neck impairments

Harmonization

- This measure is not fully harmonized with the related measures. Measure 0425 and its related measures complement one another to measure impairment across multiple areas of the human body.

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?

Measure developers to not appear to have identified competing measures or similar systems

There are related measures, which address FS changes due to impairments in other areas of the body.

While there are many related measures, the specifics are significantly different. The measures are not fully harmonized, although the related measures may be used if multiple body areas require therapeutic interventions.

N/A - No measures were identified, No concerns, Unclear

Harmonized with related measures

Yes. Knee, hip, ankle, shoulder, neck etc need to follow the same measurement philosopy as function. There may be an ablity to reduce data collection if the measures are based on standards like movement.

Yes -- is there a way to link? many patients have multiple conditions

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 2/7/2020

- No NQF members have submitted a support/non-support choice

No NQF members have commented

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_FOTO_Low_Back_0425-637088195491643687.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.

Yes

1a. Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0425

Measure Title: Functional Status Change for Patients with Low Back Impairments

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

Date of Submission: 11/8/2019

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

Outcome

☐ Outcome:

☒ Patient-reported outcome (PRO): The Low Back Functional Status patient-reported outcome measure

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 low back 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 Low Back Functional Status (FS) Patient Reported Outcome Measure (PROM).

Step #3: A patient-specific report is produced that describes the data entered, the Low Back 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 Low Back 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.)

Method:

During April-June 2019, we surveyed 57,210patients who presented for a new episode of care for a low back impairment and who completed the Low Back FS PROM. Immediately after each patient responded to the Low Back FS PROM, administered via computer adaptive testing (CAT), each patient was presented with the following question:

"In the last few questions, we asked you about your ability to do several physical activities. To what extent did you find those questions meaningful or important?"

__ All or most questions were meaningful or important
__ Some questions were meaningful or important
__ Few or no questions were meaningful or important

Since the Low Back 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 depending on their functional status level. For example, an individual with high (good) functional status might only find the most high functioning items to be valuable and meaningful because they have no difficulty with the rest of the items. However, the items were administered via CAT with each patient administered, on average, 5-7 of the items from the full item bank, and those items should be mostly tailored to their level of ability. Therefore, we hypothesized that the majority of participants would find most of the Low Back FS PROM items administered to be valuable and meaningful to them. In addition to levels of functional status, we tested a series of baseline patient characteristics for their perception of level of meaningfulness of the Low Back FS PROM items. Identifying specific patient groups that find more or fewer items to be meaningful would provide opportunity of improvements to the item bank as it evolves over time.

Results:

The sample of patients surveyed is described in Table 1, including patient characteristics tested for item meaningfulness.

Table 1: Patient sample surveyed for item meaningfulness (N=57,210)	
Admission functional status [mean(sd)]	48.8(14.0)
Gender (% Female)	59.6
Age groups (%)	
14 to <18	2.2
18 to <45	25.2
45 to <65	35.0
65 to 89	37.5
Acuity (%)	
0-7 days	4.0
8-14 days	6.2
15-21 days	8.1
22-90 days	22.6
91 days to 6 months	13.5
Over 6 months	45.7
Payer (%)	
Indemnity	4.2
Medicaid	5.9
Medicare A	1.5
Patient	0.7
Workers Comp	4.6
HMO, PPO	44.6
No Fault, Auto	1.4
Medicare B under 65	3.4
Medicare B 65 or above	22.6
Other (including Litigation, School, NoCharge, MedC, Commercial)	11.0
Exercise History (%)	
At least 3 times/wk	39.0
1 to 2 times/wk	24.6

Table 1: Patient sample surveyed for item meaningfulness (N=57,210)	
Seldom or Never	36.4
Previous treatment (%)	49.8
Language (%)	
English	98.6
Spanish	1.4

Abbreviations: HMO=health maintenance organization; PPO=preferred provider organization.

Differences in rate of response categories by the above participant characteristic were tested using Pearson Chi-squared, and are presented in Table 2. Overall, 79.8% of patients thought that all or most items were meaningful.

Table 2: Level of meaningfulness of the low-back items (N=57,210)			
Intake	**All or most items meaningful**	**Some items meaningful**	**Few or no items meaningful**
**Intake functional status (%)			
1st quartile (0-39.7)	89.0	8.4	2.6
2nd quartile (39.7-48.4)	84.9	11.8	3.3
3rd quartile (48.4-57.3)	80.2	15.8	4.1
4th quartile (57.4-100)	65.0	21.8	13.3
**Gender (%)			
Male	78.0	15.5	6.5
Female	80.9	13.7	5.5
**Age groups (%)			
14 to <18	72.2	18.8	9.0
18 to <45	79.2	14.7	6.1
45 to <65	81.4	13.1	5.5
65 to 89	79.1	15.2	5.7
**Acuity (%)			
0-7 days	84.1	11.7	4.2
8-14 days	81.9	13.1	4.9
15-21 days	79.7	14.7	5.6
22-90 days	78.8	15.3	5.9

Range	All or most items meaningful	Some items meaningful	Few or no items meaningful
91 days to 6 months	79.4	14.7	5.9
Over 6 months	79.7	14.2	6.1
**Payer			
Indemnity	77.7	15.6	6.8
Medicaid	84.6	10.5	4.9
Medicare A	77.5	16.2	6.3
Patient	81.1	15.0	3.9
Workers Comp	83.8	12.0	4.2
HMO, PPO	78.6	14.9	6.5
No Fault, Auto	82.5	13.7	3.9
Medicare B under 65	85.5	10.0	4.5
Medicare B 65 or above	79.2	15.3	5.6
Other (including Litigation, School, NoCharge, MedC, Commercial)	80.5	14.4	5.0
**Exercise History			
At least 3 times/wk	78.7	15.2	6.1
1 to 2 times/wk	79.5	14.9	5.7
Seldom or Never	81.2	13.2	5.6
**Previous treatment (%)			
Yes	81.0	13.8	5.1
No	78.5	15.0	6.5
*Language (%)			
English	79.8	14.4	5.8
Spanish	78.1	13.4	8.5

Abbreviations: HMO=health maintenance organization; PPO=preferred provider organization.

**Chi2 significant at P<0.001; *Chi2 significant at P<0.01

Interpretation:

As mentioned above, the majority of patients found all or most of the items meaningful to them, confirming our hypothesis. When examining level of perceived functional ability at Intake (i.e., functional status score), patients with lower levels of function tended to more strongly perceive the items as meaningful, and these results ranged from 89% of patients in the lowest functioning quartile reporting most or all items as meaningful to 65% in the highest functioning quartile. With respect to age, the youngest patient group (age 14-<18) had the second lowest rate of patients finding most or all items meaningful (72%), providing an

opportunity to improve the meaningfulness of the item bank for this age group. All other patient groups assessed had 78% or more patients finding most or all items meaningful, providing strong evidence supporting the overall meaningfulness of the Low Back FS PROM item bank.

****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 Low Back PRO-PM (NQF measure 0425) and the clinical process of administering interim PROM assessments during the episode of care

Background:

We assessed the relationship between NQF measure 0425 scores at discharge(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. This analysis was part of larger study recently published in Quality of Life Research.[1] 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 two patient groups that were administered either one interim assessment, or two or more 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 one early interim assessment, or two or more early interim assessments, we matched 1 patient without an interim assessment. Matching was done on all variables used in the Low Back 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 low back 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.[2,3] 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. Additionally, only clinicians who achieved a completion rate of at least 50% were included. "Completion rate" was defined as the percentage of a clinicians' patients with an intake FS PROM for whom a discharge PROM was also recorded. **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 PRO-PM provider score level. For each clinician, a rate (in percent) of early interim PROM administration was 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 0425.

Results:

Patients with one early interim PROMs (n=9,092), or two or more early interim PROMs (n=6,894), were each matched with one patient that had no interim assessment, selected from all available patients that had no interim assessment (n=83,101). 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 samples with no interim assessment

Levels	Early 1 interim	Matched: no interim	Early 2+ interims	Matched: no interim
	n=9,092	**n=9,092**	**n=6,894**	**n=6,894**
Patient characteristics used for risk-adjustment				
Functional status at admission	48.9	49.0	47.3	47.7
Age	56.3	56.3	56.9	57.0
Female	58.2%	57.2%	61.0%	61.8%
Days from onset to admission (acuity)				
0-7 days	5.7%	6.3%	5.0%	4.4%
8-14 days	8.2%	8.6%	7.3%	6.7%
15-21 days	9.6%	10.4%	8.1%	7.6%
22-90 days	24.1%	23.5%	23.4%	23.6%

Levels	Early 1 interim	Matched: no interim	Early 2+ interims	Matched: no interim
91 days to 6 months	11.9%	12.0%	11.5%	11.9%
Over 6 months	40.5%	39.2%	44.7%	45.8%
Payer				
Indemnity insurance	2.9%	2.7%	3.0%	2.8%
Medicaid	3.3%	2.7%	2.9%	3.2%
Medicare A	1.0%	0.9%	1.3%	0.9%
Medicare B Under Age 65	3.1%	3.0%	3.2%	2.8%
No fault, Auto insurance	1.1%	1.3%	1.5%	1.5%
Workers compensation	8.3%	9.0%	7.2%	6.9%
Other (Litigation, Medicare C, School, No charge, Early Intervention, Commercial Insurance)	6.5%	5.8%	6.5%	6.7%
Health Maintenance Organization, Preferred Provider	73.8%	74.6%	74.4%	75.2%
Surgical history				
No related surgery	83.5%	83.6%	80.4%	79.4%
1 related surgery	10.6%	10.8%	12.8%	12.8%
2 related surgeries	3.3%	3.2%	4.0%	4.4%
3 or more related surgeries	2.6%	2.4%	2.8%	3.4%
Exercise history				
At least 3x/week	37.8%	37.7%	37.3%	39.4%
1-2x/week	24.5%	24.5%	24.7%	25.5%
Seldom or Never	37.7%	37.8%	38.0%	35.1%
Medication use at intake	53.3%	53.6%	57.0%	57.0%
Received Previous treatment	46.9%	46.0%	50.2%	51.5%
Post-surgical: Lumbar Fusion	0.8%	1.0%	1.3%	1.9%
Post-surgical: Laminectomy / Foraminectomy / Discectomy	1.0%	0.8%	1.4%	1.5%
Specific comorbidities:				
Angina	1.4%	1.2%	1.4%	1.6%
Anxiety	14.4%	13.4%	15.8%	16.7%

Levels	Early 1 interim	Matched: no interim	Early 2+ interims	Matched: no interim
Arthritis	45.6%	44.5%	48.1%	48.5%
Asthma	10.4%	10.1%	11.4%	11.8%
Chronic obstructive pulmonary disease	3.7%	3.6%	3.6%	4.0%
Depression	16.7%	15.6%	18.5%	18.5%
Diabetes type I or II	15.2%	15.2%	15.3%	14.8%
Headache	22.7%	21.6%	24.5%	23.7%
Incontinence	6.1%	5.6%	7.3%	6.8%
Kidney, bladder, prostate, or urination problems	10.8%	10.7%	11.6%	10.7%
Neurological Disease	1.4%	1.4%	1.7%	1.7%
Obesity (BMI ≥30 kg/m2)	41.2%	41.3%	43.0%	43.1%
Osteoporosis	9.1%	8.5%	10.4%	10.3%
Previous Accident	11.8%	11.9%	12.5%	11.5%
Sleep Dysfunction	19.8%	19.3%	21.4%	20.1%
Stroke	3.5%	3.6%	3.6%	3.4%
Additional confounders				
Number of visits	10.6	10.5	15.9	16.0
Duration of episode in days	31.4	31.4	55.1	54.7

Patients treated by clinicians with high rates of early interim assessment (n=2,451) had on average 2.5 additional functional status points at discharge, compared to those treated by clinicians with low rates of early interim assessment (n=2,273 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

Levels	Clinicians	Mean rate of early interim assessments	Mean FS at discharge (95% confidence interval)
Low rates of early interim assessment	2,451	2.7%	61.3 (60.9-61.7)
High rates of early interim assessment	2,273	72.7%	63.8 (63.4-64.2)

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 0425.

References:

1. Werneke MW, Deutscher D, Fritz J, et al. Associations between interim patient-reported outcome measures and functional status at discharge from rehabilitation for non-specific lumbar impairments. *Qual Life Res.* 2019.
2. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Stat Med.* 2014;33(6):1057-1069.
3. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70(1):41-55.

1a.3. SYSTEMATIC REVIEW(SR) 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.

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

Systematic Review	Evidence
Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	**NA**
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If	**NA**

Systematic Review	Evidence
not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**NA**
Provide all other grades and definitions from the evidence grading system	**NA**
Grade assigned to the **recommendation** with definition of the grade	**NA**
Provide all other grades and definitions from the recommendation grading system	**NA**
Body of evidence: Quantity – how many studies? Quality – what type of studies?	**NA**
Estimates of benefit and consistency across studies	**NA**
What harms were identified?	**NA**
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	**NA**

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.

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.

Patients with low back impairments with functional status deficits are very common in rehabilitation therapy. Functional deficits affect large numbers of people leading to substantial morbidity, high resources use, severity of illness and is a leading cause of poor quality of life for patients that negatively affects society In addition, functional status deficits may severely impact people of any age. Therefore, functional status change measurement during rehabilitation treatment is an important construct.

The Low Back FS PROM was designed to assess functional status and change in functional status in patients with low back impairments. Improved function is a primary goal of therapy for low back pain across the world. The primary purposes of the physical therapy profession according to the Guide to Physical Therapist Practice (American Physical Therapy Association. 2001) include enhancing physical functional abilities, restoring, maintaining, and promoting optimal physical function, wellness, fitness, and optimal quality of life as it relates to movement and health. The World Confederation for Physical Therapy has a similar purpose described in the Declarations of Principle and Position Statements (1999) that emphasizes the importance of the activities and participation component of the International Classification of Functioning, Disability and Health (ICF) (World Health Organization 2001). Therefore, functioning, as described by a patient's ability to perform and participate in different physical and social activities, is important when establishing treatment goals for patients attending physical therapy. The Guide offers clear recommendations for assessing functional status by physical therapists, but the recommendation is applicable to other types of providers treating patients with functional deficits.

The Low Back Patient Reported Outcome-Performance Measure (PRO-PM) begins with the patient reported status of function at the onset of care (intake). The specifics of the deficiency in function, reported by the patient, provides data for the clinician to analyze and incorporate into the development of the plan of care by setting specific functional goals.

Repeated PROM assessments can assist the clinician in verifying the effectiveness of the plan of care implemented, or, conversely the need to adjust the plan of care to improve effectiveness.

The final measure quantifies the patient's perception of function at the end of care i.e., at discharge from rehabilitation services. Because the measure relies on patient self-report, the functional status outcomes measures are patient-centered and reflect the patient's perceived functional ability.

The measure of functional status change collected during rehabilitation is, by definition, an outcome measure of effectiveness or quality associated with the treatment provided.bMonitoring of aggregated clinician and clinic performance derived from the risk adjusted, aggregated outcome data (of all patients treated by a clinician or clinic) can be used to monitor quality and identify quality improvement to elevate the effectiveness of care for a specific provider. Thus, measurement of effectiveness of care for patients with low back impairments can help to promote quality, improve accountability, and ultimately reduce practice variation and enhance outcomes of care across therapy providers.

Low back pain is a heterogeneous condition, with little consensus on its diagnosis, classification or treatment. (Fourney, Andersson et al. 2011) Although a variety of clinical practice guidelines for the care of patients with low back pain exist, (Arnau, Vallano et al. 2006 ; Delitto, George et al. 2012) guidelines are not are not universally accepted or followed. Preliminary evidence suggests that implementation of evidenced based guidelines may result in improved patient functional status. (Rutten, Degen et al. 2010) However, the multitude of Clinical Guidelines developed in the US and abroad and promulgated by a variety of professional associations and societies may contribute to the variation in care. (Jackson, Hettinga et al. 2009; Chilibeck, Vatanparast et al. 2011; Guevara-Lopez, Covarrubias-Gomez et al. 2011; Rudwaleit and Marker-Hermann 2012) Yet, most guidelines lack specificity of detail for physical therapy practice. (Ladeira 2011)

It is widely known that high degrees of variability in the pathways of care for patients with low back impairment exist both within and between geographic regions, with variability in adherence to evidence based guidelines, use of invasive procedures, opioid use and advanced imaging. (Cherkin, Deyo et al. 1994 ; Deyo and Mirza 2006 ; Friedly, Chan et al. 2008 ; Webster, Cifuentes et al. 2009) (Ferguson, Holdsworth et al. 2010) Prior research also shows variation in risk-adjusted functional status outcome by clinic and therapist. (Resnik, Feng et al. 2006; Resnik, Liu et al. 2008)

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.*

Performance scores are detailed in section 2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE of the Testing Form with relevant excerpts provided here:

TABLE 2b4.2iia: Performance at the Clinician Level Over Time

Clinician performance (n=2,552)

Year Mean Residuals ± SD (95%CI)

Minimum-Maximum

2016 -0.3±4.3 (-0.42 to -0.09)

-13.6 to 26.4

2017 0.2±4.2 (0.01 to 0.35)

-16.1 to 23.0

2018 1.0±4.5 (0.86 to 1.20)

-13.3 to 22.7

TABLE 2b4.2iiia: Performance at the Clinic Level Over Time

Clinic performance (n=1,182)

Year Mean Residuals±SD (95%CI)

Minimum-Maximum

2016 -0.3±3.56 (-0.48 to -0.08)

-11.5 to 21.1

2017 0.2±3.40 (-0.01 to 0.40)

-9.2 to 20.6

2018 1.2±3.62 (0.99 to 1.39)

-10.2 to 17.7

SCORES BY DECILE

TABLE 2b4.2iib: Performance Gap at the Clinician Level Over Time

Performance gap over time (years) at the clinician level

Decile ranking by average clinic residuals 2016

(5,772 clinicians) 2017

(6,800 clinicians) 2018

(7,899 clinicians) Total

(12,025 clinicians)

1	-7.6	-7.5	-7.0	-7.1
2	-4.9	-4.7	-4.3	-4.4
3	-3.5	-3.3	-2.8	-3.1
4	-2.3	-2.1	-1.6	-2.0
5	-1.3	-1.0	-0.5	-1.0
6	-0.2	-0.1	0.6	-0.1
7	0.8	1.1	1.7	1.0
8	2.1	2.4	3.1	2.2
9	3.8	4.1	4.8	3.9
10	7.9	8.1	8.8	7.6
Total	-0.5	-0.3	0.3	-0.3

Values are mean residuals by deciles of average clinician residuals.

Residuals represent the difference between actual and predicted outcomes at discharge.

A residual of 0 represents no difference between actual and predicted outcomes.

Higher residuals represent better outcomes.

TABLE 2b4.2iiib: Performance Gap at the Clinic Level Over Time

Performance gap over time (years) at the clinic level

Decile ranking by average clinic residuals	2016	2017	2018	Total
	(1,757 clinics)	(2,029 clinics)	(2,440 clinics)	(3,098 clinics)
1	-6.8	-6.3	-5.8	-6.2
2	-4.3	-3.9	-3.5	-3.8
3	-3.1	-2.6	-2.3	-2.7
4	-2.1	-1.8	-1.3	-1.8
5	-1.2	-0.9	-0.4	-1.1
6	-0.4	-0.1	0.4	-0.2
7	0.5	0.8	1.3	0.7
8	1.5	1.8	2.5	1.7
9	2.9	3.2	4.1	3.0
10	6.6	6.6	7.6	6.3
Total	-0.6	-0.3	0.3	-0.4

Values are mean residuals by deciles of average clinic residuals.

Residuals represent the difference between actual and predicted outcomes at discharge.

A residual of 0 represents no difference between actual and predicted outcomes.

Higher residuals represent better outcomes.

NUMBER OF PATIENTS: (from TABLE 2b4.2i: Performance at the patient level)

All patients with complete outcomes data during 2016-2018

Year N (%) Mean Residuals±SD
(95%CI)
Minimum-Maximum
2016 183,113 (28) -0.4±12.9
(-0.43 to -0.31)
-77.2 to 64.3
2017 217,651 (33) -0.2±13.1
(-0.21 to -0.10)
-77.7 to 69.7
2018 251,911 (39) 0.4±13.3
(0 .35 to 0.46)
-73.5 to 61.1
Total 652,675 0.0±13.2
(-0.03 to 0.03)
-77.7 to 69.7

CHARACTERISTICS OF THE MEASURED ENTITIES: To view this lengthy table, please see Testing Form section 1.6 TABLE 1.6.VI: Patients with FS measures at Initial Evaluation & Discharge Aged 14 to 89: patient exclusion testing by age (n = 625,675 patients)

PERFORMANCE GAP was demonstrated using 1) 3 quality categories, and 2) deciles. The methods for the 3 QUALITY CATEGORIES and DECILES approaches were first detailed in Testing Form in the Validity section 2b1.2vii-viii and further examined with respect to demonstrating

Performance Gap in the Testing Form in section 2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE. A summary of findings is as follows:

PERFORMANCE GAP BY 3 QUALITY CATEGORIES:

Results were that for each of the 3 quality categories for low, average, and high performance, 18%, 68% and 14% of CLINICIANS and 29%, 52%, and 19% CLINICS, respectively. Please see FIGUREs 2b4.2iia and 2b4.2iiia for visual illustrations of these differences which suggest ample room for providers to make meaningful improvements in their quality as measured by 0425.

PERFORMANCE GAP BY DECILES:

In this case, the overall performance gap was represented by differences in mean residuals between the 1st and 10th decile. CLINICIAN PERFORMANCE GAP BY DECILES was demonstrated in table 2b4.2iib. Overall, average residual scores by clinic ranks based on deciles of their average residual scores ranged from -7.1 to +7.6 for 1st and 10th decile ranks, respectively. Over the three-year period assessed, performance gap between 1st and 10th decile ranks were from -7.6 to +7.9 in 2016 to -7.0 to +8.8 in 2018. CLINIC PERFORMANCE GAP BY DECILES was demonstrated in table 2b4.2iiib below. Overall, average residual scores by clinic ranks based on deciles of their average residual scores ranged from -6.2 to +6.3 for 1st and 10th decile ranks, respectively. Over the three-year period assessed, performance gap between 1st and 10th decile ranks were from -6.8 to +6.6 in 2016 to -5.8 to +7.6 in 2018.

IMPROVEMENT OVER TIME is detailed in the Use and Usability tab section 4b1.

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.

N/A

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.*

Below we comment first on age, gender, and insurance status; these factors were examined in the context of risk adjustment analyses. Second, education was examined.

Here we present descriptive disparities data over time. We note trends for increasing age and payer type of Medicare.

(from Testing Form Tables 1.6.i, 1.6.ii, and 1.6.vii):

	2002-2004		
n=1285	2007-2008		
n=17,439	2014-2016		
n=414,125,			
Age (Mean±SD)	46±16	51±17	57±16.8
Gender (% female)	59	60	60
Payer source (%)			
Indemnity Insurance	6	5	3
Medicaid	4	4	5
Medicare	11	20	32
HMO/PPO	41	48	42
Workers Comp	34	10	6

We further note the evolving role of these variables with respect to their impact within 2 risk adjustment models over time as shown in the Table below:

Beta coefficient by submission (2014 submission for 0425 and current):

	Original submission beta (2011-2013 data)	Maintenance submission beta (2014-2016 data)
Age:	-0.1	-0.1
Sex: Female	-1.2	-0.3
Payer (HMO, Preferred Provider as reference)		
Indemnity insurance	0.1	-2.6
Medicaid	-4.6	-4.7
Medicare A	-0.5	-1.4
Medicare B Under Age 65		-3.0
Medicare B	-1.2	
No fault, Auto insurance		-6.3 -4.2
Workers compensation	-5.1	-5.7
Other (Litigation, Medicare C, School, No charge, Early Intervention, Commercial Insurance)	-1.0	-1.1

While the size of some beta coefficients may be different because of changes in/additions to the overall model (as described in detail in the Testing Form), it is interesting to note that in general, the sizes of the beta

coefficient remain fairly stable over time. One difference is that we now examine Medicare B under the age of 65 separately from Medicare B for ages 65 and older, positing that the former represents patients who receive social security disability.

Our examination of educational level variables is detailed in the Testing Form in section 2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES; please see that section for details. In short, the results did not support an important contribution of the education variable to the prediction of FS at discharge, after having controlled for all other variables already included in the RA model. However, additional testing would be required, possibly testing collapsed groups of educational level, before making a final conclusion on its appropriateness as a social risk factor that needs to be adjusted for.

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

N/A

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):*

Musculoskeletal, Musculoskeletal : Joint Surgery, Musculoskeletal : Low Back Pain, Musculoskeletal : Osteoarthritis, Musculoskeletal : Osteoporosis, Musculoskeletal : Rheumatoid Arthritis, Surgery, Surgery : Perioperative and Anesthesia

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

Health and Functional Status : Change, Health and Functional Status : Physical Activity, Primary Prevention

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

Elderly, Populations at Risk, Populations at Risk : Dual eligible beneficiaries, Populations at Risk : Individuals with multiple chronic conditions, Populations at Risk : Veterans

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.)*

www.fotoinc.com/science-of-foto/nqf0425

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:** Low_Back_Data_Dictionary_-_RA_Coefficients_NQF2019July-637001594271537751.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:** Low_Back_Item_Bank_NQF2019July-637001594378571060.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.

Yes

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.

The description and the numerator statements in the original application now appear to conflict. The Title was and is "Functional Status change for patients with lumbar spine impairments." The description was expressed as a percent of patients who were measured. Finally, the Numerator was described again as the number of eligible patients who were measured at the beginning and the end of care. The title and the risk adjustment process describes an outcome measure, but the numerator statement describes a process measure for measuring functional status change. In this application we are changing the numerator statement and the brief description to describe the measurement of the risk adjusted benchmarked effectiveness measure derived from aggregated functional status data submitted by patients with lumbar spine impairments who were treated by participating providers.1. Changes to the wording of the specifications to clarify the intent of the measure.

2. ICD-10 codes added to S.7. Denominator Details to further clarify measure intent.

3. The item bank of the Low Back FS PROM was expanded from 25 to 28 items as part of the measure maintenance over time. Details are provided in the Testing Form sections 2a2.2iii and 2a2.3iii.

4. The risk adjustment model was updated to include additional factors and variables including exercise history, previous treatment, medication use, 30 specific comorbidities, and post-surgical categories. This is described below in section S.10. Stratification Information.

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 Low Back impairments and who completed the Low Back PRO-PM.

The numerator, as it applies to the 3 levels, is defined as follows:

Patient Level: The residual functional status score for the individual patient with a low back impairment.

Individual Clinician Level: The average of residuals in functional status scores in patients who were treated by a clinician in a 12-month time period for a low back impairment.

Clinic Level: The average of residuals in functional status scores in patients who were treated by a clinic in a 12-month time period for a low back impairment.

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 score for the individual patients with low back impairments is derived by applying the statistical risk adjustment model described in S.10 and applying steps 1-5 as described in S.14.

Individual Clinician Level: The average of residuals in functional status scores in patients who were treated by a clinician in a 12-month time period for low back impairment. Average scores are calculated for all clinicians, but performance is evaluated only for those clinicians that had a minimum of 10 patients in the previous 12 months to maximize stability of the benchmarking estimates. The score is derived by applying steps 1-6 as described in S.14.

Clinic Level: The average of residuals in functional status scores in patients who were treated within a clinic in a 12-month time period for lumbar impairments. Average scores are calculated for all clinics, but performance is evaluated only for large clinics (5 or more clinicians) that had a minimum of 40 patients, and small clinics (1-4 clinicians) that had a minimum of 10 patients per clinician, in the previous 12 months to maximize stability of the benchmarking estimates. The score is derived by applying steps 1-6 as described in S.14.

Items and response options are provided in the attachment in section S.2c. above.

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

The target population is all patients 14 years and older with a Low Back impairment who have initiated an episode of care and completed the Low Back FS PROM.

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).

The ICD-10 codes relevant for this measure are:

G54.1, G54.4, G57.0, M43.06, M43.07, M43.08, M43.16, M43.17, M43.18, M43.26, M43.27, M43.28, M43.5X6 , M43.5X7, M43.5X8, M43.8X6, M43.8X7, M43.8X8, M45.6, M45.7, M45.8 M46.1, M46.46, M46.47, M46.48, M47.16, M47.26, M47.27, M47.28, M47.816, M47.817, M47.896, M47.897, M47.898, M48.06, M48.07, M51.06, M51.16, M51.17, M51.26, M51.27, M51.36, M51.37, M51.46, M51.47, M51.86, M51.87, M51.9, M53.2X6, M53.2X7, M53.2X8, M53.88, M54.16, M54.17, M54.18, M54.3, M54.4, M54.5, M99.73, S32.0, S32.1, S32.2, S33.0, S33.1, S33.2, S33.3, S33.5, S33.10, S33.11, S33.12, S33.13, S39.002, S39.012

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

Patients who are not being treated for a Low Back 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.)*

This measure is risk-adjusted, not risk-stratified. The methods used to develop the FOTO risk-adjustment Low Back model were the same as the methods described in detail in a recent publication by Deutscher et at, 2018 [Deutscher, D., Werneke, M. W., Hayes, D., Mioduski, J. E., Cook, K. F., Fritz, J. M., et al. (2018). Impact of Risk Adjustment on Provider Ranking for Patients with Low Back Pain Receiving Physical Therapy. J Orthop Sports Phys Ther, 48(8), 637-648] Briefly, we used data from adult patients with Low Back pain treated in outpatient rehabilitation clinics during 2014-2016, 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 (low back fusion, laminectomy or other); and 31 comorbidities.

For further details, please see Measure Testing Form section 2b3. Risk Adjustment/Stratification for Outcome or Resource Use Measures. The model variables and coefficients are contained in the document attached above in section S.2b. Data Dictionary, Code Table, or Value Sets.

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.)*

DEFINITIONS:

Patient's Functional Status Score. A Functional Status (FS) Score is produced when the patient completes the FOTO Low Back FS PROM administered via computer adaptive testing or short form.

Patient's FS Change Score. An FS Change Score is calculated by subtracting the Patient's FS Score at the Initial Evaluation (i.e., the start of the care episode) from the Patient's FS Score at Discharge (i.e., the end of the care episode).

Predicted FS Change Score. FS Change Scores for patients are risk adjusted with a model developed using multiple linear regression methods that account for the following independent variables: Patient's FS Score at Initial Evaluation, patient age, symptom acuity, surgical history, gender, specific co-morbidities, payer type, use of medication for the low back impairment at Initial Evaluation, previous treatment for the low back impairment, exercise history, and post-surgical category if applicable. The Patient's FS Change Score is the dependent variable. The statistical regression method provides a set of coefficients that accounts ("adjusts") for the association of each variable with the FS outcome as it applies to each patient, resulting in a risk-adjusted Predicted FS Change Score.

Residual Score: The Residual Score is calculated as the difference between the actual change and risk-adjusted predicted change scores and should be interpreted as the unit of FS change different than predicted given the risk-adjustment variables of the patient being treated. As such, the risk-adjusted Residual change score represents risk-adjusted change corrected for patient characteristics. Risk-adjusted Residual change scores of

zero (0) or greater (>0) should be interpreted as functional status change scores that were predicted or better than predicted given the risk-adjustment variables of the patient. Risk-adjusted residual change scores less than zero (<0) should be interpreted as functional status change scores that were less than predicted given the risk-adjustment variables of the patient.

Aggregated Residual Scores: The average of Residual scores of FS (actual change - predicted change after risk adjustment) from a provider (clinician or clinic). The aggregated scores are used to make comparisons between clinicians or clinics.

STEPS TO CALCULATE THE PRO-PM SCORE, APPLYING THE ABOVE DEFINITIONS:

Patient level measures use steps 1-5.

Clinician and clinic level measures use steps 1-6.

1) The patient is identified as age 14 or older and presenting for an episode of care for a low back impairment and completing the FOTO Low Back FS PROM which generates the Patient's FS Score at Initial Evaluation.

2) The patient completes the FOTO Low Back FS PROM at or near Discharge, which generates the Patient's FS Score at Discharge.

3) The Patient's FS Change Score (raw, non-risk-adjusted) is generated.

4) A Predicted FS Change Score is generated for the patient using the risk-adjustment model.

5) A Residual Score is generated for the patient.

6) The average Residual Scores per clinician and/or clinic are calculated, and scores for all clinicians/clinics in the database are ranked. The quality score is the percentile of the clinician and/or clinic ranking. The quality scores and its 95% CI can be compared to the benchmark (a score of zero) to determine if the performance is below, at, or above the predicted average. FOTO recommends that clinicians have a minimum of 10 patients/year and clinics have a minimum of 10 patients/therapist per year for small clinics or 40 patients per year for larger clinics (5 or more clinicians) in order to obtain stable estimates of provider performance.

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 Initial Evaluation 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 2a2.3iv: Reliability (R) at the Provider Level).

In summary, the average reliability of clinics meeting the FOTO unique threshold of number of patients per clinic for quality reporting was 0.84. At the clinician level, average reliability for clinicians with 10 or more patients per year was 0.71.

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 years 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 rare within our data (0.03%). 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.

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.

Calculation of response rates to be reported with performance measure results:

All eligible patients are expected to be surveyed if they fit the target population description of age 14 or older and present for an episode of care for a low back impairment. At this time, no minimum response rate is required because FOTO does not have data to determine participation rate, defined as the percentage of patients completing the survey at admission from all eligible patients. FOTO does not yet have visibility into all patients within the clinic, although levels of integration between FOTO and multiple electronic medical record systems are advancing in a manner that this may become feasible in the future.The instructions vary somewhat on the timing and format of the survey process. The timing involves initial or Intake surveys and status (or follow-up) surveys.

The paper/pencil version available on the sponsor's website at http://www.fotoinc.com/science-of-foto/NQF0425.html applies to Intake and Status surveys and includes the following instructions:

FOTO Lumbar Functional Status 10-Item Paper Short Form

(Date of last update: 2/08. Date of planned update: none)

We are interested in knowing whether you are having any difficulty at all with the activities listed below because of your low back problem for which you are currently seeking attention. Please provide an answer for each activity.

The local survey is when the patient is logged into the web browser and the remote survey is when the patient is responding to a link that has been emailed to the patient.

Intake Local

Welcome {{PatientFirstName}}

The comprehensive evaluation that you will have to start your therapy treatment at {{ClinicName}} includes a computerized functional assessment that will help your clinician better understand your condition and how it impacts your quality of life. This information will help your clinician develop treatment goals with you and is an important part of your treatment.

When you are ready to get started, click the ´Begin´ button. Please respond to each question with the response that best describes you or your level of function at this time.

The information you share is confidential, a part of your medical record, and is subject to all protected health care information regulations.

Intake Remote

Welcome

The comprehensive evaluation that you will have to start your therapy treatment at {{ClinicName}} includes a computerized functional assessment that will help your clinician better understand your condition and how it impacts your quality of life. This information will help your clinician develop treatment goals with you and is an important part of your treatment.

You have the option of completing the survey online prior to your first appointment, rather than in the clinic before your first treatment.

When you are ready to get started, click the ´Begin´ button. Please respond to each question with the response that best describes you or your level of function at this time. If you do not complete the entire survey, you may resume it by clicking the link in this email again.

The information you share is confidential, a part of your medical record, and is subject to all protected health care information regulations

Status Local

Welcome {{PatientFirstName}}

At the beginning of your treatment at {{ClinicName}} you completed a computerized functional assessment related to your impairment. Please complete the questionnaire again to reassess how the treatment for your impairment has helped to improve your function and pain. You will also have the opportunity to respond regarding your satisfaction with several aspects of your treatment.

Please complete the survey as it relates to your impairment at this present time. You can use the information that you learned in therapy to help you answer the questions. This will help your clinician assess how your treatment has or has not helped you.

When you are ready to get started, click the ´Begin´ button. Please respond to each question with the response that best describes you or your level of impairment at this time.

Status Remote

Welcome

At the beginning of your treatment at {{ClinicName}} you completed a computerized functional assessment related to your impairment. Please complete the questionnaire again to reassess how the treatment for your impairment has helped to improve your function and pain. You will also have the opportunity to respond regarding your satisfaction with several aspects of your treatment.

Please complete the survey as it relates to your impairment at this present time. You can use the information that you learned in therapy to help you answer the questions. This will help your clinician assess how your treatment has or has not helped you.

When you are ready to get started, click the ´Begin´ button. Please respond to each question with the response that best describes you or your level of impairment at this time. if you do not complete the entire survey, you may resume it by clicking the link again.

IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

All eligible patients are expected to be surveyed if they have functional deficits for the applicable body part. At this time, no minimum response rate is required because FOTO does not have data to determine participation rate, defined as the percentage of patients completing the survey at admission from all eligible patients. FOTO does not have visibility into the true denominator of patients within the clinic, because FS data collection is elective, not required and because for most providers FOTO is not yet linked to an electronic medical record system that includes all patients treated.

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 Low Back FS PROM and risk adjustment questions (as described in the measure Testing Form). A patient completes the FS PROM and respond to risk adjustment questions at the start of an episode of care. The patient again responds to the FS PROM, at a minimum, at or near the time of discharge from the episode of care.

The Low Back FS PROM may be administered via computer adaptive testing (CAT) or 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. The components needed to complete NQF 0425 are publicly available on the FOTO website at no charge.

Proxy and Recorder modes of administration are described above in section S.15. Sampling.

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.)*

N/A

2. Validity – See attached Measure Testing Submission Form

NQF_testing_attachment_Sep2017_Low_Back_0425_Aug_1_2019-637014541692429107.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.

Yes

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.

Yes

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.

Yes - Updated information is included

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

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Functional Status Change for Patients with Low Back Impairments
Date of Submission: 8/1/2019

Type of Measure:

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

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*).

The FOTO database has existed since 1994 and consists of approximately 21 million patient assessments. FOTO provides web-based data collection and reporting for roughly 23,000 clinicians in 5200 outpatient clinics across all 50 states.

Swinkels, I. C., van den Ende, C. H., de Bakker, D., Van der Wees, P. J., Hart, D. L., Deutscher, D., et al. (2007). Clinical databases in physical therapy. *Physiother Theory Pract, 23*(3), 153-167, doi:10.1080/09593980701209097.
Swinkels, I. C., van den Ende, C. H., van den Bosch, W., Dekker, J., & Wimmers, R. H. (2005). Physiotherapy management of low back pain: does practice match the Dutch guidelines? *AUST J PHYSIOTHER, 51*(1), 35-41.

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)	Hart et al 2006:[23] (2002-2004)	1285	NR+	56	18

Analysis	Data source (years)	Entities tested			
		Patients	Clinicians	Clinics	States
2a2ii. **Data elements (patient) level:** Reliability of point estimates and change scores.	Hart et al 2010[29] & Wang et al 2010[61] (2007-2008)	17,439	NR+	377	30
2a2iii. **Clinician performance score level**: at different sample thresholds per clinician per calendar year*	FOTO internal analysis (2016-2018)	585,357	12,025	3,409	50+DC
2a2iv. **Clinic performance score level:** at different sample thresholds per clinic per calendar year**	FOTO internal analysis (2016-2018)	618,472	19,704	3,098	50+DC
2b1. VALIDITY TESTING					
2b1i. **Data elements (patient) level:** Content validity (coverage), i.e., analysis examined if test items covered the content area of functional status); Structural validity (uni-dimensionality, local independence and item fit); Differential Item Functioning	Hart et al 2006:[23] (2002-2004)	1,285	NR+	56	18
2b1ii. **Data elements (patient) level:** Construct validity; Sensitivity to change; Clinically important improvement	Hart et al 2010[29] Wang et al 2010[61] (2007-2008)	17,439	NR+	377	30
2b1ii. **Data elements (patient) level:** Construct validity & discriminating ability testing	Hart et al 2012:[24] (2007-2008)	8,198	382	111	24
2b1iv. **Clinician performance score level:** Construct Validity of performance score level; Validity of performance classification*	FOTO internal analysis (2016-2018)	585,357	12,025	3,409	50+DC
2b1v. **Clinic performance score level:** Construct Validity of performance score level; Validity of performance classification**	FOTO internal analysis (2016-2018)	618,472	19,704	3,098	50+DC
2b2. EXCLUSIONS ANALYSIS					
2b2. Age exclusion	FOTO internal analysis (2016-2018)	652,675	23,430	4,156	50 states and DC
2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES					
2b3.Risk adjustment model development	Deutscher et al 2018:[14] (2014-2016)	414,125	12,569	3,048	50 states and DC
2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE					

Analysis	Data source (years)	Entities tested			
		Patients	Clinicians	Clinics	States
2b4i. Data elements (patient) level:	Deutscher et al 2018:[14] (2014-2016)	652,675	23,430	4,156	50 states and DC
2b4ii. Clinician performance score level: *	FOTO internal analysis (2016-2018)				
2b4iii. Clinic performance score level: **	FOTO internal analysis (2016-2018)				
2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS					
2b6i. Comparing patients with or without complete outcomes; assessing impact of adjusting for risk of patient censoring using inverse-probability-weighting on the risk-adjustment model and provider ranking	FOTO internal analysis (2016-2018)	977,155	25,893	4,263	50+DC
2b6ii. Correlations between clinician residuals and their completion rates for clinicians participating in the performance analyses*	FOTO internal analysis (2016-2018)	585,357	12,025	3,409	50+DC
2b6iii. Correlations between clinic residuals and completion rates for clinics participating in the performance analyses**	FOTO internal analysis (2016-2018)	618,472	19,704	3,098	50+DC
2b6iv. Average residuals at the clinician level by completion rate categories with or without the use of Inverse Probability Weighting*	FOTO internal analysis (2016-2018)	585,357	12,025	3,409	50+DC
2b6v. Average residuals at the clinic level by completion rate categories with or without the use of Inverse Probability Weighting**	FOTO internal analysis (2016-2018)	618,472	19,704	3,098	50+DC

+ NR=not reported

*Clinicians with 10+ patients per calendar year with FS measures at initial evaluation & 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 initial evaluation & 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*)

See Tables below:

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 IRT person reliability)	Hart et al 2006:[23] (2002-2004)	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.	Hart et al 2010[29] & Wang et al 2010[61] (2007-2008)	Patients with FS scores at initial evaluation or initial evaluation & discharge	TABLE 1.6.II
2a2iii. **Data elements (patient) level:** Construct validity & discriminating ability testing	Hart et al 2012:[24] (2007-2008)	Patients with FS scores at initial evaluation or initial evaluation & discharge	TABLE 1.6.III
2a2iii. **Clinician performance score level:** at different sample thresholds per clinician per calendar year*	FOTO internal analysis (2016-2018)	Patients treated by clinicians with 10+ patients per calendar year with FS scores at initial evaluation & discharge	TABLE 1.6.IV
2a2iv. **Clinic performance score level:** at different sample thresholds per clinic per calendar year**	FOTO internal analysis (2016-2018)	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 initial evaluation & discharge	TABLE 1.6. V
2b1. VALIDITY TESTING			
2b1i. **Data elements (patient) level:** Content validity (coverage), i.e., analysis examined if test items covered the content area of functional status); Structural validity (uni-dimensionality, local independence and item fit); Differential Item Functioning	Hart et al 2006:[23] (2002-2004)	Patients responding to the full item bank considered for the measure development	TABLE 1.6.I

Analysis	Data source (years)	Sample selection criteria	TABLE
2b1ii. **Data elements (patient) level:** Construct validity; Sensitivity to change; Clinically important improvement	Hart et al 2010[29] Wang et al 2010[61] (2007-2008)	Patients with FS scores at initial evaluation & discharge who also complete the patient global rating of change at discharge	TABLE 1.6.II
2b1ii. **Data elements (patient) level:** Construct validity & discriminating ability testing	Hart et al 2012:[24] (2007-2008)	Patients with FS measures at initial evaluation & discharge	TABLE 1.6.III
2b1iv. **Clinician performance score level:** Construct Validity of performance score level; Validity of performance classification*	FOTO internal analysis (2016-2018)	Patients treated by clinicians with 10+ patients per calendar year with FS measures at initial evaluation & discharge	TABLE 1.6.IV
2b1v. **Clinic performance score level:** Construct Validity of performance score level; Validity of performance classification**	FOTO internal analysis (2016-2018)	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 initial evaluation & discharge	TABLE 1.6. V
2b2. EXCLUSIONS ANALYSIS			
2b2. Age exclusion	FOTO internal analysis (2016-2018)	Patients with FS measures at initial evaluation & discharge from age 14 to 98	TABLE 1.6.VI
2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES			
2b3.Risk adjustment model development	Deutscher et al 2018:[14] (2014-2016)	Patients with FS measures at initial evaluation & discharge	TABLE 1.6.VII
2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE			
2b4i. Performance patient level	FOTO internal analysis (2016-2018)	Patients with FS measures at initial evaluation & discharge	TABLE 1.6.VI
2b4ii. Performance individual clinician level	FOTO internal analysis (2016-2018)	Patients treated by clinicians with 10+ patients per calendar year with FS measures at initial evaluation & discharge	TABLE 1.6.IV
2b4iii. Performance clinic/group practice level	FOTO internal analysis (2016-2018)	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 initial evaluation & discharge	TABLE 1.6. V

Analysis	Data source (years)	Sample selection criteria	TABLE number
2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS			
2b6i. Comparing patients with or without complete outcomes; assessing impact of adjusting for risk of patient censoring using inverse-probability-weighting on the risk-adjustment model and provider ranking	Deutscher et al 2018:[14] (2014-2016)	Patients with FS measures at initial evaluation	TABLE 1.6.VIII
2b6ii. Correlations between clinician residuals and their completion rates for clinicians participating in the performance analyses	FOTO internal analysis (2016-2018)	Patients treated by clinicians with 10+ patients per calendar year with FS measures at initial evaluation & discharge	TABLE 1.6.IV
2b6iii. Correlations between clinic residuals and completion rates for clinics participating in the performance analyses	FOTO internal analysis (2016-2018)	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 initial evaluation & discharge	TABLE 1.6. V
2b6iv. Average residuals at the clinician level by completion rate categories with or without the use of inverse-probability-weighting	FOTO internal analysis (2016-2018)	Patients treated by clinicians with 10+ patients per calendar year with FS measures at initial evaluation & discharge	TABLE 1.6.IV
2b6v. Average residuals at the clinic level by completion rate categories with or without the use of inverse-probability-weighting	FOTO internal analysis (2016-2018)	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 initial evaluation & discharge	TABLE 1.6. V
Abbreviations: FS = functional status			

TABLE 1.6.I: Patient Characteristics at Initial evaluation: original development sample (n = 1285 patients)

Characteristic	Value
Age (Mean±SD; Min-Max)	46±16; 14-100
Age 14 to <45 (%)	52
Age 45 to 65 (%)	37
Age >65 (%)	11
Gender (% female)	59

Characteristic	Value
Symptoms acuity of (%)	
Acute (0 to 21 days)	22
Subacute (22-90 days)	22
Chronic (>90 days)	57
Surgical history (%)	
None	72
One or more	24
Missing	0
Payer source (%)	
Indemnity Insurance	6
Litigation	1
Medicaid	4
Medicare	11
Patient	1
HMO	19
PPO	22
Workers Comp	34
Other	1

TABLE 1.6.II: Patient Characteristics at Initial evaluation: data elements testing (n=17,439)

Characteristic	Value
Diagnoses* (%)	
Spinal pathology including lumbago, intervertebral disc, sciatica, spondylosis, spinal stenosis (ICD-9 codes 720–724)	29
Soft tissue disorders of muscle, synovium, tendon, bursa, or enthesopathies† (ICD-9 codes 725–729)	18
Sprains and strains including sacroiliac region, lumbar spine, sacrum, (ICD-9 codes 846–848 including unspecified sprain or strain)	4
Post-surgical conditions including discectomy and fusion (CPT codes 22224 and 22612)	5
Not otherwise classified	2
Missing	43
Age (mean±SD, min, max in yr)	51±17, 18, 100
Age 18 to <45 (%)	36
Age 45 to 65 (%)	40
Age >65 (%)	24
Gender (% female)	60
Acuity of symptoms (%)	

Characteristic	Value
Acute (0–21 days)	23
Subacute (22–90 days)	25
Chronic (>90 days)	52
Surgical history (%)	
None	82
1	12
2	3
3	2
4 or more	1
Exercise history (%)	
At least 3 /week	37
1–2 /week	26
Seldom or never	37
Payer source (%)	
PPO	38
HMO	10
Workers' compensation	10
Medicare part B	16
Indemnity	5
Medicaid	4
Medicare part A	4
Other	12
Missing	1
No. functional comorbidities (%)	
None	15
1	25
2	21
3 or more	39
Global rating of change (%)	
Improved (≥+3 to +7)	12
Not improved (-7 to <+3)	3
Missing 85	85
*Diagnoses are groups of ICD-9-CM codes or surgical CPT codes.	

Characteristic	Value
†Enthesopathies are disorders of peripheral ligamentous or muscular attachments. SD indicates standard deviation; min, minimum; max, maximum; HMO, health maintenance organization; PPO, preferred provider organization; ICD-9, International Classification of Diseases 9th revision; CPT, current procedural terminology.	

TABLE 1.6.III: Patient Characteristics: construct validity & discriminating ability testing (n = 8198)

Patient Characteristic	Initial evaluation Data Only (n = 4819)	Initial evaluation and Discharge Data (n = 3379)
Initial evaluation FS (Mean±SD; Min-Max) Initial evaluation ODQ (Mean±SD; Min-Max)	52±13; 3-95 66±17; 0-100	51±13; 3-94 65±17; 14-100
Age (Mean±SD; Min-Max)	49±16; 18-93	52±16; 18-100
Age (%)		
18 to 45	41	35
>45 to 65	40	42
>65	16	22
Missing (%)	3	1
Gender (%)		
Male	44	44
Female	56	56
Missing	<1	0
Symptom acuity (%)		
Acute	25	23
Sub-acute	24	26
Chronic	51	51
Missing	<1	<1
Surgical history (%)		
None	81	79
1 or more	19	21
Missing	<1	0
Number of comorbid conditions (%)		
None or 1	27	23
2 or 3	30	30
4 or 5	20	23
6 or more	23	24
Missing	<1	0
Fear-Avoidance of physical activities (%)		
Not elevated	68	69
Elevated	28	27
Missing	4	4
Payer source (%)		
Automobile	1	2
Fee-for-service	8	16
Medicaid	3	2
Medicare Part A	3	4

Patient Characteristic	Initial evaluation Data Only (n = 4819)	Initial evaluation and Discharge Data (n = 3379)
Medicare Part B	11	13
Patient private pay	2	1
HMO	2	2
PPO	49	42
Workers' compensation	12	15
Other	8	4
Missing	<1	1

TABLE 1.6. IV: Patient Characteristics at Initial Evaluation: clinician level testing (n = 585,357 patients)

Characteristic	Values
Age (mean±SD); min/max	56.4(17.7); 14-89
Sex (female)	59.9
Acuity of Symptoms	
0-7 days	3.9
8-14 days	6.4
15-21 days	7.9
22-90 days	23.4
91 days to 6 months	12.5
Over 6 months	45.8
Surgical History	
None	81.4
1	11.9
2	3.8
3 or more	2.9
Number of Comorbid Conditions	
None	3.9
1	7.7
2	11.9
3 or more	76.4
Exercise History	
At least 3 times/wk	38.0

Characteristic	Values
1 to 2 times/wk	24.7
Seldom or Never	37.3
Payer Source	
Indemnity Insurance	2.8
Medicaid	5.5
Medicare A	1.5
Medicare B under 65	3.6
Medicare B 65 or above	26.5
Patient	0.5
Workers' compensation	5.7
HMO /PPO	42.8
No Fault, Auto insurance	1.5
Other	9.6
Medication use at initial evaluation	52.3
Previous treatment	49.6
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 Initial evaluation: clinic level testing (n = 618,472 patients)

Characteristic	Values
Age (mean±SD); min/max	56.4(17.7); 14-89
Sex (female)	59.9
Acuity of Symptoms	
0-7 days	3.9
8-14 days	6.4
15-21 days	8.0
22-90 days	23.4
91 days to 6 months	12.5
Over 6 months	45.8
Surgical History	

Characteristic	Values
None	81.4
1	11.9
2	3.8
3 or more	2.9
Number of Comorbid Conditions	
None	3.9
1	7.8
2	11.9
3 or more	76.4
Exercise History	
At least 3 times/wk	38.1
1 to 2 times/wk	24.7
Seldom or Never	37.2
Payer Source	
Indemnity Insurance	2.7
Medicaid	5.5
Medicare A	1.5
Medicare B under 65	3.6
Medicare B 65 or above	26.6
Patient	0.5
Workers' compensation	5.6
HMO /PPO	42.8
No Fault, Auto insurance	1.5
Other	9.6
Medication use at initial evaluation	52.2
Previous treatment	49.6

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: Patients with FS measures at Initial Evaluation & Discharge Aged 14 to 89: patient exclusion testing by age (n = 625,675 patients)

Patient characteristics	(N 625,675)
Functional Status score at initial evaluation: Mean ± SD (Min to Max)	48.8±13.1 (0-98)
FS score at discharge: Mean ± SD (Min to Max)	63.0±16.7 (0-98)
Age (years): Mean ± SD (Min to Max)	56.4±17.7 (14-89)
Sex: Female	59.9
Acuity:	
0-7 days	3.9
8-14 days	6.4
15-21 days	8.0
22-90 days	23.5
91 days to 6 months	12.5
Over 6 months	45.6
Payer:	
Indemnity insurance	3.0
Medicaid	5.4
Medicare A	1.6
Medicare B Under Age 65	3.6
Medicare B Age 65 or above	26.7
Patient	0.5
Workers compensation	5.7
Other (Litigation, Medicare C, School, No charge, Early Intervention, Commercial Insurance)	9.7
No fault, Auto insurance	1.5
HMO, Preferred Provider	42.3
Surgical history:	
No related surgery	81.4
1 related surgery	11.9
2 related surgeries	3.8
3 or more related surgeries	2.9
Exercise history:	

Patient characteristics	(N 625,675)
At least 3x/week	38.2
1-2x/week	24.6
Seldom or Never	37.1
Medication use at initial evaluation	52.1
Previous treatment	49.6
Low Back surgery procedure	
Fusion	1.5
Laminectomy/Foramenectomy/Discectomy	1.5
Other surgical codes	0.1
Number of comorbidities: Mean ± SD (Median, IQR ‡)	5.0±3.2(5,4)
Specific comorbidities:	
Allergy	26.2
Angina	1.5
Anxiety or Panic Disorders	16.9
Arthritis	47.6
Asthma	11.5
Back pain (neck pain, low back pain, degenerative disc disease) †	84.7
Cancer	8.7
Chronic Obstructive Pulmonary Disease (COPD)	4.1
Congestive Heart Failure	5.3
Depression	18.6
Diabetes Type I or II	14.6
Gastrointestinal	18.6
Headaches	23.0
Hearing	6.4
Hepatitis / HIV-AIDS	1.1
High Blood Pressure	39.2
Heart Attack (Myocardial Infarction)	3.1
Incontinence	6.9
Kidney, Bladder, Prostate or Urination Problems	11.4
Neurological Disease	1.7
Obesity (BMI>=30)	40.8

Patient characteristics	(N 625,675)
Osteoporosis	10.5
Other disorders	3.4
Peripheral Vascular Disease (or claudication)	1.7
Previous accidents (Motor vehicle, work, or other accident)	12.7
Previous Surgery	38.1
Prosthesis / Implants	7.6
Sleep dysfunction	20.3
Stroke or Transient Ischemic Attack	3.5
Visual Impairment	10.8
Pacemaker	1.7
Seizures	1.4

Abbreviations: BMI, body mass index; HMO, health maintenance organization.

*Patient characteristics at initial evaluation to physical therapy for the sample used to develop the risk-adjusted model (Total), the sample used for the ranking analyses (Selected) and the sample excluded from the ranking analyses (Not selected).

Values are percent unless otherwise indicated.

†Back pain was not allowed to enter the risk-adjusted model.

‡IQR, inter quartile range. Median and IQR are reported for number of comorbidities due to the skewed distribution

TABLE 1.6.VII: Patients with FS Measures at Intake & Discharge: patient level performance testing and risk-adjustment modeling (n = 414,125 patients)

Patient characteristics	(N 414,125)
Functional Status score at admission: Mean ± SD (Min to Max)	48.8±12.6 (0-98)
FS score at discharge: Mean ± SD (Min to Max)	62.7±16.6 (0-98)
Age (years): Mean ± SD (Min to Max)	57.0±16.8 (18-116)
Sex: Female	59.9
Acuity:	
0-7 days	4.0
8-14 days	6.4

Patient characteristics	(N 414,125)
15-21 days	7.8
22-90 days	23.6
91 days to 6 months	12.6
Over 6 months	45.5
Payer:	
Indemnity insurance	3.1
Medicaid	4.9
Medicare A	1.4
Medicare B Under Age 65	4.0
Medicare B Age 65 or above	28.2
Patient	0.5
Workers compensation	6.1
Other (Litigation, Medicare C, School, No charge, Early Intervention, Commercial Insurance)	8.1
No fault, Auto insurance	1.6
HMO, Preferred Provider	42.1
Surgical history:	
No related surgery	80.9
1 related surgery	12.4
2 related surgeries	3.8
3 or more related surgeries	2.8
Exercise history:	
At least 3x/week	39.2
1-2x/week	24.1
Seldom or Never	36.7
Medication use at intake	55.0
Previous treatment	49.6
Low Back surgery procedure	
Fusion	1.5
Laminectomy/Foramenectomy/Discectomy	1.6
Other surgical codes	0.1
Number of comorbidities: Mean ± SD (Median, IQR ‡)	4.9±3.3(4,5)

Patient characteristics	(N 414,125)
Specific comorbidities:	
Allergy	27.1
Angina	1.6
Anxiety or Panic Disorders	15.6
Arthritis	48.3
Asthma	11.0
Back pain (neck pain, low back pain, degenerative disc disease) †	80.1
Cancer	8.5
Chronic Obstructive Pulmonary Disease (COPD)	4.2
Congestive Heart Failure	5.5
Depression	17.9
Diabetes Type I or II	13.9
Gastrointestinal	18.9
Headaches	22.2
Hearing	6.9
Hepatitis / HIV-AIDS	1.0
High Blood Pressure	38.1
Heart Attack (Myocardial Infarction)	3.2
Incontinence	6.7
Kidney, Bladder, Prostate or Urination Problems	11.4
Neurological Disease	1.9
Obesity (BMI>=30)	39.9
Osteoporosis	10.5
Other disorders	5.1
Peripheral Vascular Disease (or claudication)	1.8
Previous accidents (Motor vehicle, work, or other accident)	13.3
Previous Surgery	37.6
Prosthesis / Implants	7.3
Sleep dysfunction	19.9
Stroke or Transient Ischemic Attack	3.4
Visual Impairment	11.3
Pacemaker	0.8
Seizures	0.7

Patient characteristics	(N 414,125)
Abbreviations: BMI, body mass index; HMO, health maintenance organization. *Patient characteristics at admission to physical therapy for the sample used to develop the risk-adjusted model (Total), the sample used for the ranking analyses (Selected) and the sample excluded from the ranking analyses (Not selected). Values are percent unless otherwise indicated. †Back pain was not allowed to enter the risk-adjusted model. ‡IQR, inter quartile range. Median and IQR are reported for number of comorbidities due to the skewed distribution	

TABLE 1.6.VIII: Patients with FS Measures at Initial Evaluation: missing outcomes data testing (n = 977,155 patients)

Patient characteristics	(N 977,155)
Functional Status score at initial evaluation: Mean ± SD (Min to Max)	48.7±13.3 (0-98)
FS score at discharge (n=652,675): Mean ± SD (Min to Max)	63.0±16.7 (0-98)
Age (years): Mean ± SD (Min to Max)	55.0±17.7 (14-89)
Sex: Female	60.0
Acuity:	
0-7 days	4.1
8-14 days	6.4
15-21 days	7.9
22-90 days	23.0
91 days to 6 months	12.4
Over 6 months	46.1
Payer source:	
Indemnity insurance	3.4
Medicaid	6.5
Medicare A	1.5
Medicare B Under Age 65	3.9
Medicare B Age 65 or above	23.5
Patient	0.6
Workers compensation	5.2

Patient characteristics	(N 977,155)
Other (Litigation, Medicare C, School, No charge, Early Intervention, Commercial Insurance)	10.2
No fault, Auto insurance	1.4
HMO, Preferred Provider	43.8
Surgical history:	
No related surgery	82.2
1 related surgery	11.3
2 related surgeries	3.7
3 or more related surgeries	2.8
Exercise history:	
At least 3x/week	37.8
1-2x/week	24.7
Seldom or Never	37.5
Medication use at initial evaluation	52.8
Previous treatment	48.9
Low Back surgery procedure	
Fusion	1.3
Laminectomy/Foramenectomy/Discectomy	1.3
Other surgical codes	0.1
Number of comorbidities: Mean ± SD (Median, IQR ‡)	5.0±3.3(4,4)
Specific comorbidities:	
Allergy	25.8
Angina	1.5
Anxiety or Panic Disorders	18.3
Arthritis	46.1
Asthma	11.8
Back pain (neck pain, low back pain, degenerative disc disease) †	84.3
Cancer	8.1
Chronic Obstructive Pulmonary Disease (COPD)	4.2
Congestive Heart Failure	5.1
Depression	19.8
Diabetes Type I or II	14.2

Patient characteristics	(N 977,155)
Gastrointestinal	18.3
Headaches	24.3
Hearing	6.0
Hepatitis / HIV-AIDS	1.1
High Blood Pressure	37.7
Heart Attack (Myocardial Infarction)	3.1
Incontinence	6.6
Kidney, Bladder, Prostate or Urination Problems	11.1
Neurological Disease	1.7
Obesity (BMI>=30)	41.0
Osteoporosis	10.0
Other disorders	3.6
Peripheral Vascular Disease (or claudication)	1.7
Previous accidents (Motor vehicle, work, or other accident)	12.8
Previous Surgery	36.9
Prosthesis / Implants	7.1
Sleep dysfunction	20.7
Stroke or Transient Ischemic Attack	3.5
Visual Impairment	10.1
Pacemaker	1.6
Seizures	1.5

Abbreviations: BMI, body mass index; HMO, health maintenance organization.

*Patient characteristics at initial evaluation to physical therapy for the sample used to develop the risk-adjusted model (Total), the sample used for the ranking analyses (Selected) and the sample excluded from the ranking analyses (Not selected).

Values are percent unless otherwise indicated.

†Back pain was not allowed to enter the risk-adjusted model.

‡IQR, inter quartile range. Median and IQR are reported for number of comorbidities due to the skewed distribution

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 Low Back 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.

Payer:

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.

Education level:

A standard data point to ask all respondents their level of education was added for a limited period of time to the FOTO system during the year 2018. Because this was a standard question asked of all patients, we acquired a large sample size for this variable enabling us to test its impact on the RA model. See section 2b3 for more details.

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: RELIABILITY OF CRITICAL DATA ELEMENTS (patient level)

2a2.2i: Internal Consistency

Reliability-based estimates of internal consistency were calculated using data collected at initial evaluation from the measure development sample of patients answering all items.[23] Reliability was examined through item response theory (IRT) person reliability analysis, which is equivalent to the calculation of Cronbach's alpha.[40]

2a2.2ii: 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 estimate of functional status (FS) ability. The SEM reports measurement error in the same units as the original measurement and was calculated as the initial evaluation standard deviation times the square root of 1 minus the internal consistency reliability. Because each FS score represented a point estimate for each patient's FS, the 95% confidence interval (CI) band associated with the point estimate FS score was constructed to provide an estimate of precision of the measure (i.e., FS score ± 1.96 X CSEM, where CSEM is the conditional standard error of measurement). We estimated 10 CSEMs, 1 for each of the 10 scale ranges (0 –10, 11–20, . . . , 91–100) by averaging the SEMs in each FS scale range. Based on IRT measurement models, SEM varies by level of FS. Therefore, different score ranges have different magnitudes of SEMs. Extreme scores are expected to have larger SEs because less information is obtained from patients at the extremes (i.e., patients with very low or very high functioning).

In addition to the interpretation of a point estimate, clinicians are faced with the need to interpret change in scores during treatment. Statistically reliable change, as described by Schmitt and Di Fabio,[50] reflects the statistical significance of individual change. To assess statistically reliable change, we computed the minimal detectable change (MDC)[4] as: MDC_{95}=1.96 X √2 X CSEM, where 1.96 represents the z value associated with a 95% CI. As above, we estimated 10 CSEMs, 1 for each of the 10 scale ranges (0 –10, 11–20, . . . , 91–100). For each CSEM, we multiplied the result by the square root of 2 to accommodate the 2 measurements involved in measuring change: initial evaluation and discharge.

As computed, MDC_{95} represents the smallest threshold for identifying statistically reliable change greater than random measurement error.[4]

2a2.2iii: Reliability Improvement Following Item Bank Expansion as Part of the Measure Maintenance Over Time.

In response to clinician (user) feedback, 3 new items were developed internally, with the direct involvement of a small panel of 3 physical therapists with experience treating back pain. These 3 new items were then added to the FOTO measurement system for patient administration and data collection during 2013. The new items asked about the level of difficulty when using a broom, getting down to and up from the floor, and changing positions quickly like sitting to standing.

To assess whether the expanded (28-item) Low Back FS bank provided improvements to score reliability (compared to the original 25-item bank), we began by obtaining score-level-specific item information values for each of the 25 original items and each of the 3 new items. We then calculated score-level-specific test information, standard error, and reliability, based on (a) the original 25-item bank, and (b) the expanded 28-item bank. We then compared observed 25-item vs. 28-item test information, standard error, and reliability values in 2 ways. First, we calculated average test information, standard error, and reliability across the score range of 20 to 80 (i.e., across approximately +/-3 standard deviations). Second, we graphed original vs. expanded Low Back FS item bank test information, standard error, and reliability by overall Low Back FS ability (0-100). Thus, this combination of numeric and visual evidence allowed us to quantify overall measurement precision gains and identify specific Low Back FS score levels with improved precision when using the expanded (28-item) bank vs. the original (25-item) bank.

METHODS: RELIABILITY OF PERFORMANCE MEASURE SCORE (e.g., *signal-to-noise analysis*)

2a2.2iv: 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.[1] **For the purpose of these analyses we defined the term 'provider' as either the clinic or the clinician, depending on the analysis conducted.** 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):

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

The variance between all provider groups (signal) was estimated using **risk-adjusted residuals** calculated using a **mixed-effects hierarchical linear model (HLM)** with patients nested within the provider (i.e. either the clinic or the clinician). 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.

The variance within each provider (noise/error) 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.[1]

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). The average reliability for all providers was also tested using a more conservative threshold by increasing the number of patients per provider, i.e., a minimum of 20, 30 and 40 patients per provider.

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: RELIABILITY OF CRITICAL DATA ELEMENTS

2a2.3i: Internal Consistency

Person reliability estimate was 0.92.[23]

2a2.3ii: Reliability of Point Estimates and Change Scores

Because the SEMs at discharge were similar to the SEs at initial evaluation, with an average of 0.15 score unit differences (minimum=0.00, maximum =0.33), for brevity we report only SEs associated with initial evaluation FS score estimations. TABLE 2a2ii shows the CSEMs, one for each of the 10 initial evaluation scale ranges (0 – 10, 11–20, . . . , 91–100), as well as their associated 95%CI. The CSEM was smallest (i.e., 2.2) over the scale range of 41 to 60 and increased at both ends of the scale range, with a range of 2.2 to 3.7 for the score range of 20 to 80 which included 97% of patients.

On average, the mean of 95% CI upper limits of MDC_{95} values for all patients was 13.9, but the mean MDC_{95} value for 97% of patients with FS initial evaluation scores between 20 and 80 was 7.8 (TABLE 2a2ii).

TABLE 2a2.3ii: Reliability of Point Estimates for Baseline FS Scores and Change Scores (n=17,439)

FS Range at Initial evaluation	No. of Patients (%)	CSEM	95% CI of CSEM	MDC_{95}
0–10	0.5	11.4	22.2	31.5
11–20	0.6	5.3	10.4	14.7
21–30	3.3	3.5	6.9	9.8
31–40	15.9	2.6	5.1	7.2
41–50	27.2	2.2	4.2	6
51–60	30.8	2.2	4.3	6.1
61–70	15.6	2.7	5.3	7.5
71–80	4.4	3.7	7.3	10.3
81–90	0.8	5.4	10.6	15
91–100	0.9	11.4	22.2	31.5
Abbreviations: FS=functional status, CSEM=conditional standard error of measurement at initial evaluation, 95% CSEM=1.96 X CSEM, MDC95=minimal detectable change (95% confidence interval) given initial evaluation FS.				

2a2.3iii: Reliability Improvement Following Item Bank Expansion as Part of the Measure Maintenance Over Time.

Measurement precision gains from the original (25-item) to the expanded (28-item) bank are presented in TABLE 2a2.3iii, reporting average test information, standard error, and reliability across score levels 20 to 80 (i.e., approximately +/- 3 standard deviations). Increased test information inevitably leads to decreased standard error and increased reliability. The additional test information provided by the 3 new Low Back FS items increased average reliability across score levels 20 to 80 from 0.88 to 0.90; 0.90 is a reliability standard for making individual score-level comparisons.[5, 8, 43]

TABLE 2a2.3iii: Measurement Precision Gains with Expanded 28-item Low Back FS Item Bank

Item Bank	Test Information *	Standard Error *	Reliability *
25-item	10.06	2.64	0.88
28-item	11.52	2.48	0.90

* Average across score levels 20 to 80 (approximately +/- 3 standard deviations)

The following figures visually present measurement precision gains for the expanded (28-item) Low Back FS bank. Test information and corresponding standard error (FIGURE 2a2.3iiia), and reliability (FIGURE 2a2.3iiib) are reported across the full ability score range (i.e., 0-100), comparing original vs. expanded item bank performance.

In general, the increase in total item information provided by the 3 new Low Back FS items contributed to observable improvements in measurement precision across essentially the full 0-100 ability score range, as reflected in reduced score-level standard error and increased score-level reliability.

FIGURE 2a2.3iiia: Test Information & Standard Error by Level of Functional Status Score

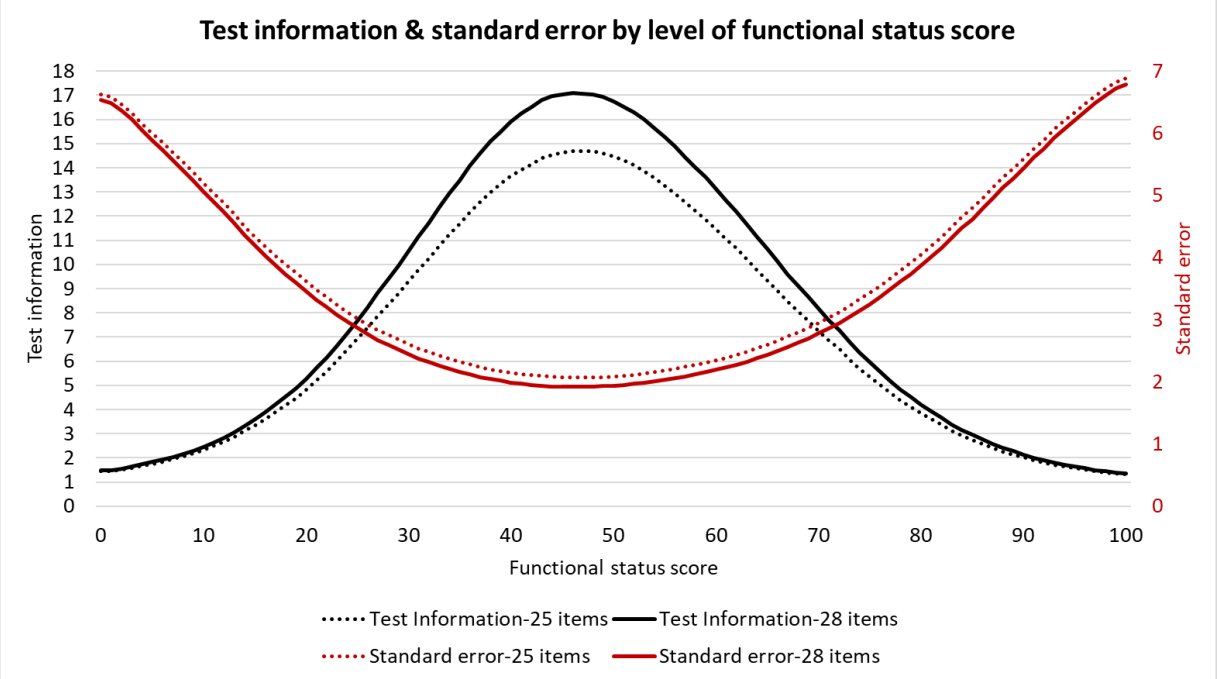
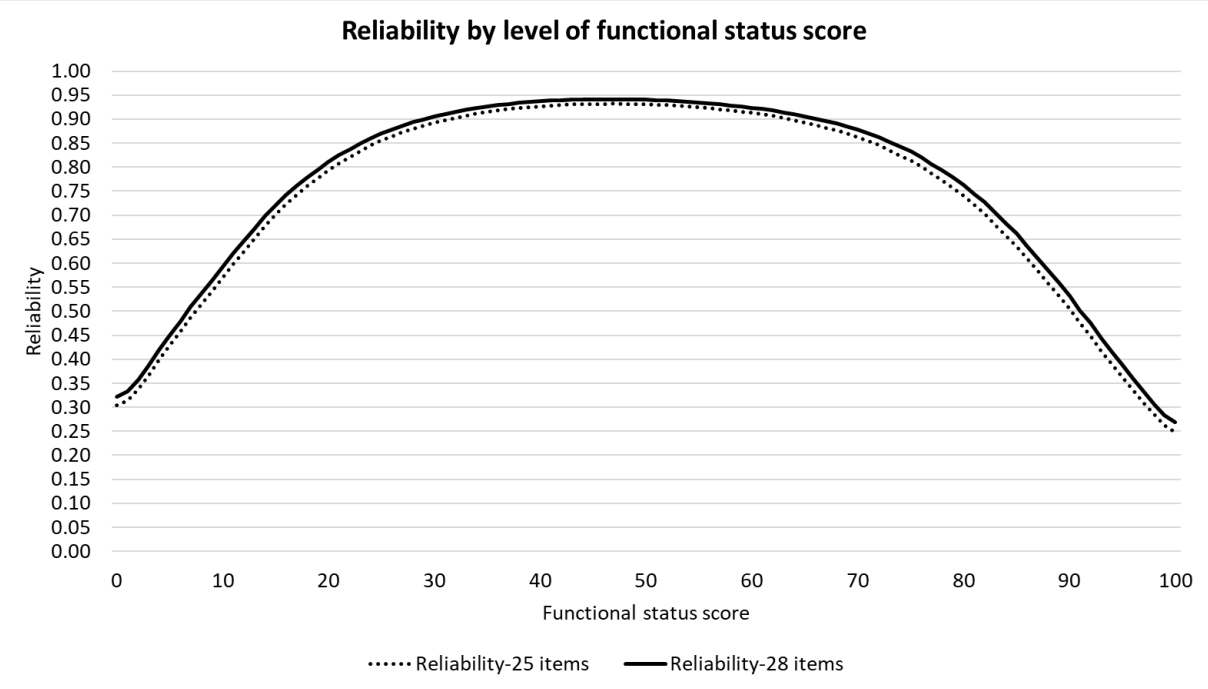


FIGURE 2a2.3iiib: Reliability by Level of Functional Status Score



RESULTS: RELIABILITY OF PERFORMANCE MEASURE SCORE (e.g., signal-to-noise analysis)

2a2.3iv: 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 initial evaluation 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.

At the clinic level, the average reliability for clinics meeting the FOTO threshold of number of patients per clinic for quality reporting was **0.84**. At the clinician level, average reliability for clinicians with 10 or more patients per calendar year was **0.71**.

TABLE 2a2.3iv: 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

Reliability (R) at the provider level: 2016-2017								
Clinic	*FOTO	5.8	3098	0.84	0.21	1.00	2674	86
	20+	5.8	2942	0.86	0.41	1.00	2636	90
	30+	5.5	2732	0.87	0.48	1.00	2523	92
	40+	5.5	2520	0.88	0.55	1.00	2397	95
Clinician	10+	6.8	12025	0.71	0.19	0.98	7029	58
	20+	6.8	7787	0.77	0.37	0.98	5618	72
	30+	6.9	4849	0.81	0.50	0.98	4191	86
	40+	7.4	2867	0.84	0.57	0.98	2799	98
*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?*)

2a2.4i: INTERPRETATION: RELIABILITY OF DATA ELEMENTS (patient level)

The results suggest that scores on the Low Back FS PROM have strong internal consistency (0.92) with an SEM of 2.2 to 3.7 out of 100 points for the score range of 20 to 80 including 97% of patients (approximately +/- 3 SDs).
The combination of numeric and visual evidence quantifying overall measurement precision gains identified the wide range of score levels with modest but notably improved precision when using the expanded 28-item Low Back FS bank, compared to the original 25-item bank. This supports the effort to improve content coverage and measurement precision as part of the ongoing measure maintenance over time.

2a2.4ii: INTERPRETATION: RELIABILITY OF PERFORMANCE MEASURE SCORE

Based on these findings and using the minimum threshold of a reliability of $\geq$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)]. Findings also suggest that the threshold of 10 patients for the clinician level PRO-PM is sufficient to reliably differentiate between levels of clinicians. 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: VALIDITY OF CRITICAL DATA ELEMENTS (patient level)

2b1.2i: Content Validity (Do test items appear to be measuring the construct of interest?)

The Low Back FS PROM was developed using an item bank established by concurrently calibrating items from the Back Pain Functional Scale (BPFS) [54, 55] and additional physical functioning (PF) items.[21, 30, 62] A small, internal clinician panel of 3 physical therapists with experience treating patients with back pain was involved in the selection of these items for testing. The BPFS is a patient self-reported outcomes instrument for patients with Low Back impairments; it was designed as a measure to estimate functional status. The PF items, although not designed specifically for patients with Low Back impairments, have also been used to assess the functional status of such patients.[35, 47] The latent trait of interest is the patient's perception of his or her ability to perform the functional tasks described in the BPFS/PF items.

As part of the Low Back FS measure's maintenance over time, as described above (reliability testing section 2a2.2iii), 3 new items were added as a response to clinician feedback with the goal of improving the measure's content validity. The clinician users provided ideas for content based on activities their patients with low back pain told them were important. These new items asked about the level of difficulty when using a broom, getting down to and up from the floor, and changing positions quickly like sitting to standing.

2b1.2ii: Structural Validity (uni-dimensionality, local independence and item fit)

Data were analyzed to determine how well uni-dimensionality and local independence IRT assumptions were met. Presence of a dominant factor was assessed with exploratory factor analyses (EFA) of latent trait variables followed by confirmatory factor analyses (CFA).[23]

The Andrich [2] rating scale IRT model (RSM) was selected to assess item fit because it is a latent structure model for polytomous responses to a set of test items, which is the format of the

BPFS and PF items. Assessing item structure and data fit investigates the success of the selected model in predicting or explaining the data. Infit (i.e., weighted mean squared fit index) and

outfit (i.e., unweighted mean squared fit index) mean square statistics for the sample were examined as an assessment of whether the data fit the RSM.[65] Item infit provides information about responses given to items near patient ability. Item outfit is an outlier-sensitive statistic that assesses items that are far from patient ability levels. A recommended criterion for reasonable fit for clinical rating-scale data is infit and outfit values of 1.4 or smaller.[64] Items whose infit and outfit values were greater than 1.4 were dropped.

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.[40] We conducted EFAs of latent trait variables, followed by CFAs on all item responses. Items were considered for removal if factor loadings were below 0.40.[42] Local independence requires that, after taking into account patient ability (in this case, functional status related to a Low Back impairment), 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.[45] 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.[33] 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.

To assess if the expanded item bank maintained essential uni-dimensionality, we used CAT response data from the expanded item bank to perform a principal component analysis (PCA) on Rasch residuals using Winsteps.[39] An unexplained variance in the first contrast below 2 Eigenvalue would support that there is no other dimension in the data that contradicts the Rasch dimension. For Eigenvalues above 2, a disattenuated correlation (what the Pearson correlation would be if we could measure without error) above 0.7, would support that the two or more 'dimensions' are measuring the same construct and, therefore, can be interpreted as unidimensional. A disattenuated correlation below 0.3 would suggest multidimensionality.[38]

2b1.2iii: 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 by the amount of the trait being measured, the difficulty of the items should not vary by group membership. That is, when level of Low Back-related function is constant, there should be no differences in how subgroups of patients answer particular items.[20, 32]

Items were assessed for DIF by selecting groups of patients by gender (male, female), surgical history (yes/no), acuity of symptoms (number of calendar days between date of onset of symptoms and date of initial evaluation, i.e., acute=21 days or less, subacute=22 to <90 days,

Chronic=90 days or more), and age group (young=14 to <45, middle=45 to 65, and older>65 years). We assessed the presence of (1) uniform DIF (i.e., the interference related to demographic groups between ability or trait level and item responses is the same across the entire range measured by the test) and (2) nonuniform DIF (i.e., the interference varies at different levels of the trait being measured).[23] 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.

2b1.2iv: Construct Validity

We used known group construct validity methods to assess the ability of the Low Back FS PROM CAT generated scores to discriminate among groups of patients expected to have different levels of Low Back function. The independent variables assessed included initial evaluation FS, age, symptom acuity, surgical history, condition complexity, and prior exercise history. We used one-way ANCOVAs with FS change as the dependent variable, initial evaluation FS as the covariate, with one ANCOVA for each risk-adjustment variable as the independent variable. Post hoc Sheffe analyses were run for significant main factors of the independent variable.[29]

2b1.2v: Sensitivity to Change, Responsiveness, and Content Range Coverage

Sensitivity to change and responsiveness were assessed using two distribution-based approaches. First, effect size statistics were estimated as follows: (discharge FS minus initial evaluation FS)/(initial evaluation FS standard deviation (SD)). Second, the proportion of patients with change scores greater than the conditional MDC at the upper 95% confidence interval (TABLE 2a2ii) was reported. Additionally, content range coverage was assessed by measuring floor and ceiling effects of the scale at initial evaluation. We operationally defined a floor effect as a measure from 0 to 5 and a ceiling effect as a measure from 95 to 100 FS scores.

2b1.2vi: 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 defined as minimum threshold of improvement that may likely be considered important to the patient. To incorporate the patient's perspective on the clinical importance of an FS score change, we used a patient-reported global rating of change (GROC) scale as the external anchor.[34] 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.[26-28] Because of the large body of evidence that MCII levels are dependent on baseline FS,[18, 25-29, 37, 44, 48, 53, 56-61] 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, were estimated.

METHODS: CONSTRUCT VALIDITY OF PERFORMANCE MEASURE SCORE

2b1.2vii-viii: Performance Score Level by MCII achievement:

Providers' (i.e., clinics' 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 using the external marker of global rating of change (GROC) as described above.

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

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 providers' (individual clinics or clinicians) performance scores, we plotted each provider'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 provider 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.

2b1.2ix: Empirical Validity: Correlation of Performance Scores and Two External Markers

We assessed the correlation between clinic and clinician performance scores and two external measures that assess a similar construct: the patient-reported global rating of change (GROC) [34] assessed at discharge, and the Modified Oswestry Low Back Pain Disability Questionnaire (ODQ)[17] as change from admission to discharge.[16]

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.[10, 61] Here, the GROC was used as an external measure to assess construct validity of the provider score level of Low Back PRO-PM.

The ODQ is a widely accepted legacy measure of patient-reported functional status including 10 items related to Low Back impairment. The ODQ evolved from the original Oswestry Low Back Pain Disability Questionnaire,[16] which was designed to quantify disability in patients with Low Back syndromes. Briefly, the ODQ consists of 10 items with 6 response choices (0-5) that are summed and multiplied by 2 to produce a nonlinear [63] disability scale from 0 to 100, where higher scores represent more disability and lower function. The ODQ has been reported to demonstrate lower discriminant ability and efficiency compared to the Low Back FS PROM,[24] insufficient uni-dimensionality, and a large floor effect.[6] However, its clinical utility has been supported,[7, 9] thus rendering it appropriate as an external measure for construct validity assessment of the Low Back PRO-PM scores for measuring the performance of providers (clinics and clinicians).

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 the Low Back PRO-PM with GROC and ODQ 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.[31] Due to the scale direction, a positive correlation with the GROC and a negative correlation with the ODQ were expected. We hypothesized that the Low Back PRO-PM 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 Low Back PRO-PM and at least one of the external measures. Since the validity correlations were tested for the provider levels (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*].

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

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

RESULTS: VALIDITY OF CRITICAL DATA ELEMENTS
2b1.3i: Content Validity (Do test items appear to be measuring the construct of interest?)

Twelve BPFS items characterizing functional activities commonly affected in people with Low Back impairments were co-calibrated with 16 PF items commonly used to assess similar patients. The BPFS items are scored on a 1 to 6 response option scale representing ''Unable to perform activity'' to ''No difficulty,'' while the PF items are scored on a 1 to 3 scale representing ''Yes, limited a lot,'' ''Yes, limited a little,'' and ''No, not limited.'' Of the 16 PF items, nine items were derived from the Medical Outcomes Study Short Form 36 physical functioning scale (PF-10).[62] The PF-10 item describing bending, kneeling, and stooping was not used because the BPFS contained an item describing bending or stooping. Six additional, internally developed PF items were added because they improved the effective measurement range of patients with Low Back impairments and thus were able to assess lower functioning status compared to the PF-10. A PF item describing lifting overhead to a cabinet completed the PF 16-item pool because lifting is considered important for patients with Low Back impairments and because it also improved the effective measurement range of patients with Low Back impairments, compared to the PF-10. For all BPFS and PF items, as noted above, higher responses represented better perceived functioning. After removing 3 items with low factor loadings and/or poor fit, the 25-item pool represented a unidimensional pool with strong local independence.[23]

As described above, as part of the measure's maintenance and based on specific clinician feedback and requests for additional item content, three new items were subsequently added, with content asking about the level of difficulty when using a broom, getting down to and up from the floor, and changing positions quickly like sitting to standing.

2b1.3ii: Structural Validity (uni-dimensionality, local independence and item fit)

After removing three items, confirmatory factor analysis results for the remaining 25 items were CFI = 0.87, TLI = 0.98, and RMSEA = 0.09 for the one-factor solution, demonstrating a unidimensional item pool with strong local independence.

One BPFS item (i.e., driving 1 hour) had infit and outfit statistics >1.4. Because driving is a clinically important task for patients with Low Back impairments, and the sitting item had been deleted, and factor analytic results did not support deleting the driving item, we decided to keep the item in the pool.[23]

The unexplained variance in the first contrast was 1.83, with a disattenuated correlation above 0.97, supporting the uni-dimensionality of the expended item bank of the Low Back FS PROM.

2b1.3iii: Differential Item Functioning

Only the BPFS items describing working and driving displayed nonuniform DIF by gender and age, respectively ($P<0.002$). No items displayed uniform DIF. DIF adjusted and unadjusted Low Back FS ability estimates were highly correlated (i.e., r values all >0.9992). We believe these results represented clinically negligible DIF for the variables assessed.[23]

2b1.3iv: Construct Validity

Results supported known group construct validity of the FS measures estimated. Briefly, patients who were younger, had more acute symptoms, fewer surgeries, fewer comorbidities, and exercised more frequently before receiving rehabilitation had better discharge FS.[29]

2b1.3v: Sensitivity to Change, Responsiveness, and Content Range Coverage

Results support that the Low Back FS PROM was sensitive to change. The initial evaluation FS measures averaged 51 (SD=12), discharge FS scores were 65 (SD=16), and FS change scores were 14 (SD=16), which produces an effect size ([discharge minus initial evaluation]/[standard deviation of initial evaluation]) of 14/12=1.17, which is considered large. There were 66% of patients attaining FS change scores equal to or greater than conditional MDC at the 95% confidence interval.[29]

Of 17,439 FS estimates at initial evaluation, 21 patients had a score of 0, and 84 (0.5%) had scores between 0 and 5, which we judged as a negligible floor effect. Of those same initial evaluation data, no patient had a score of 100, and 7 had scores between 95 and 100 for negligible ceiling effect.[29]

2b1.3vi: Clinically Important Improvement

There were 2612 patients with both GROC and FS change data. Of these patients, 449 (17.2%) reported no improvement (i.e., GROC scores <3), and 2163 (82.8%) reported improvement (i.e., GROC scores ≥3). ROC analyses (TABLE 2b1.3vi) supported 5 or more FS change units represented clinically meaningful improvement. Thus, 4673 (71%) patients with discharge data reported FS change equal to or greater than MCII. When patients were grouped by baseline FS measures and 4 ROC analyses were run (1 per quartile of FS initial evaluation measures), 4626 (70%) patients reported FS change scores equal to or greater than MCII. Results suggested that the Low Back FS PROM was responsive, and MCII was dependent on initial evaluation FS with patients perceiving improvement with fewer FS units as initial evaluation FS scores increase.[29]

TABLE 2b1.3vi: Anchor-based Estimate of Minimal Clinically Important Improvement*

Baseline FS score	N	% improved (GROC≥3)	MCII / ROC cut point	MCII 95%CI	AUC	AUC 95%CI	% ≥ MCII
Overall score range	2612	82.8	5		0.781	0.758,0.803	71
1st quartile (FS 0-43)	589	78.3	9		0.814	0.772,0.856	72
2nd quartile (FS>43-51)	728	80.5	5		0.815	0.777,0.852	74
3rd quartile (FS>51-58)	580	84.3	3		0.815	0.767,0.862	75
4th quartile (FS>58-100)	715	87.7	5		0.802	0.759,0.845	59

Abbreviations: FS, Functional Status; 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

RESULTS: CONSTRUCT VALIDITY OF PERFORMANCE MEASURE SCORE

2b1.3vii-viii: Performance Score Level by MCII achievement

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).

2b1.3vii: Clinician Performance Score Level

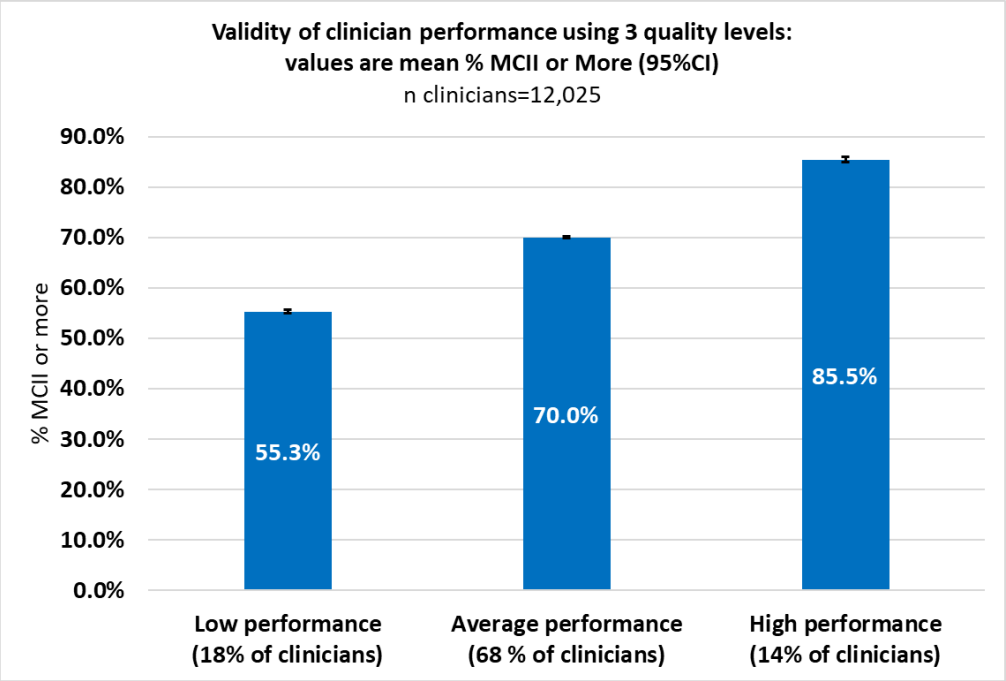
Method 1: Validity of clinician performance based on uncertainty assessments (3 levels):
The three performance levels had statistically significant differences between groups as determined by one-way ANOVA ($F(2,12022) = 4342.7$, $p < 0.001$) with a monotonic increase in rates of MCII achievement (TABLE 2b1.3vii-a).

TABLE 2b1.3vii-a: Validity of Performance at the Clinician Level Using 3 Quality Levels

Performance level	N Clinicians (%)	% MCII or more (%)
Low	2,181 (18.1)	55.3
Average	8,218 (68.3)	70.0
High	1,626 (13.5)	85.5

A Tukey post-hoc test revealed that all groups were significantly different from one another ($p<0.001$) (FIGURE 2b1.3vii -a).

FIGURE 2b1.3vii -a: Validity of Performance at the Clinician Level Using 3 Quality Levels



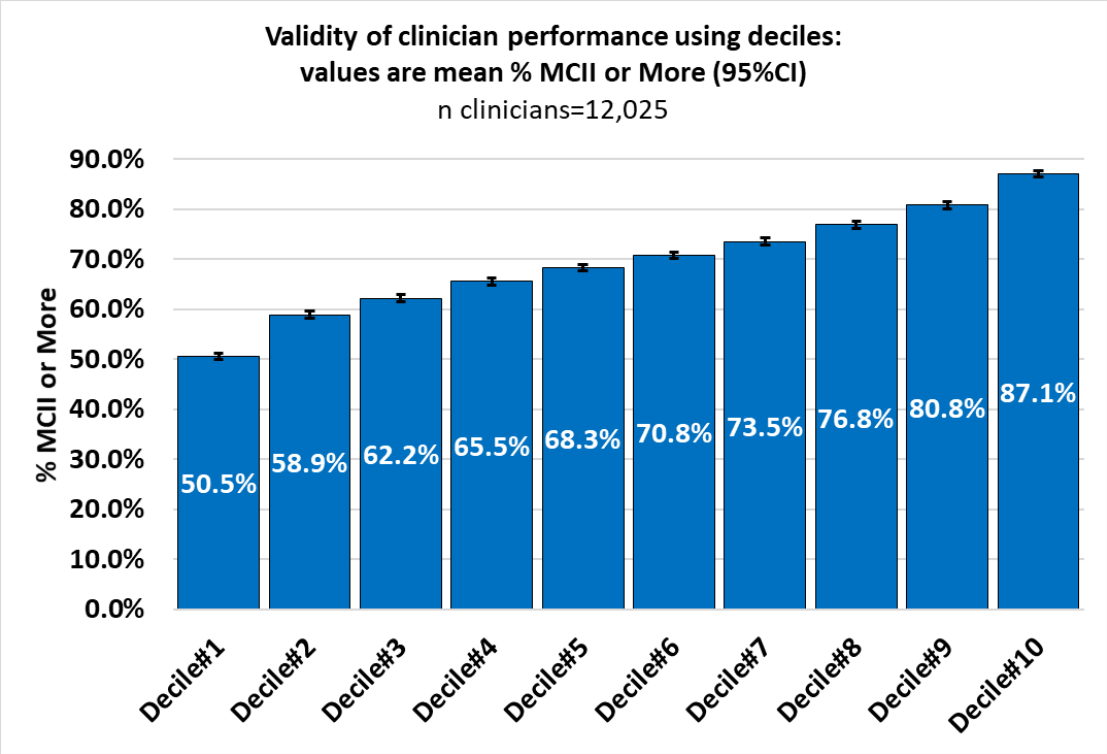
Method 2: Validity of clinician performance based on percentile ranking (10 levels):
The ten performance levels had statistically significant differences between groups as determined by one-way ANOVA ($F(9,12015) = 2175.1$, $p < 0.001$), with a monotonic increase in rates of MCII achievement (TABLE 2b1.3vii-b)

TABLE 2b1.3vii-b: Validity of Performance at the Clinician Level Using Decile Ranking

Performance level	N Clinicians	% MCII or more
Decile 1	1,203	50.5
Decile 2	1,202	58.9
Decile 3	1,203	62.2
Decile 4	1,202	65.5
Decile 5	1,203	68.3
Decile 6	1,202	70.8
Decile 7	1,203	73.5
Decile 8	1,202	76.8
Decile 9	1,203	80.8
Decile 10	1,202	87.1

A Tukey post-hoc test revealed that all groups were significantly different from one another ($p < 0.001$) (FIGURE 2b1.3vii-b).

FIGURE 2b1.3vii-b: Validity of Performance at the Clinician Level Using Decile Ranking



2b1.3viii: Clinic Performance Score Level

Similar to the above results pertaining to clinicians, a higher proportion of patient episodes managed by the higher performing clinics also experienced change equal to or greater than the MCII as compared to lower performing clinics. This pattern was observed using both methods of clinic performance ranking; uncertainty assessments (3 levels) and percentile ranking (10 levels).

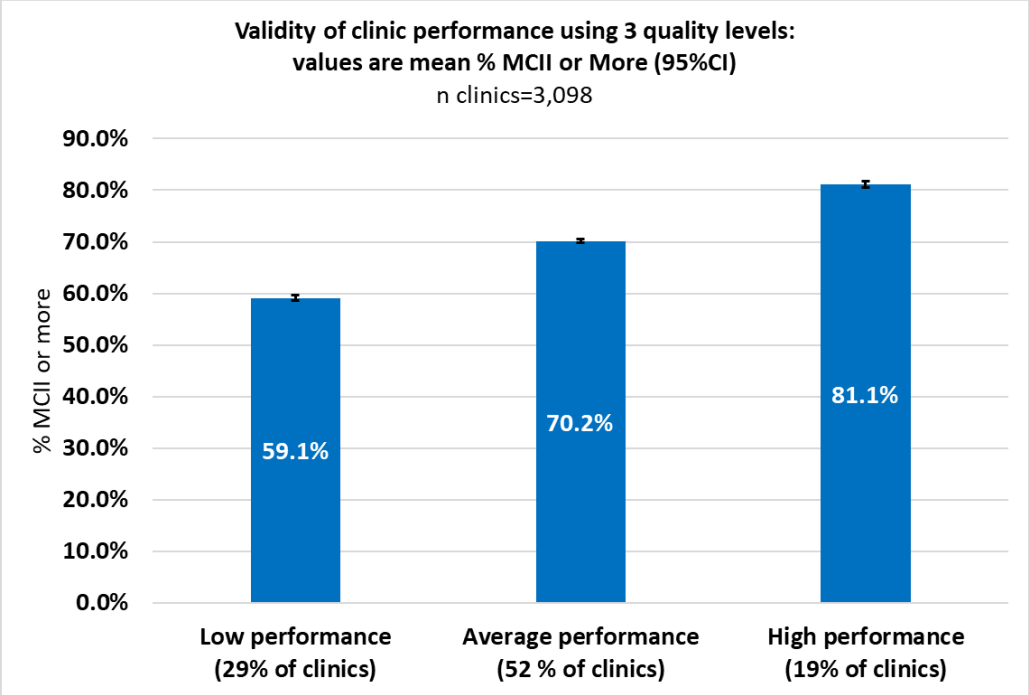
Method 1: Validity of clinic performance based on uncertainty assessments (3 levels):
The three performance levels had statistically significant differences between groups as determined by one-way ANOVA ($F(2, 3095) = 1613.3$, $p < 0.001$) with a monotonic increase in rates of MCII achievement (TABLE 2b1.3viii-a).

TABLE 2b1.3viii-a: Performance at the Clinic Level Using 3 Quality Levels

Performance level	N Clinics (%)	MCII or more(%)
Low	894 (28.9)	59.1
Average	1605 (51.8)	70.2
High	599 (19.3)	81.1

A Tukey post-hoc test revealed that all groups were significantly different from one another ($p<0.001$) (FIGURE 2b1.3viii-a).

FIGURE 2b1.3viii-a: Validity of Performance at the Clinic Level Using 3 Quality Levels



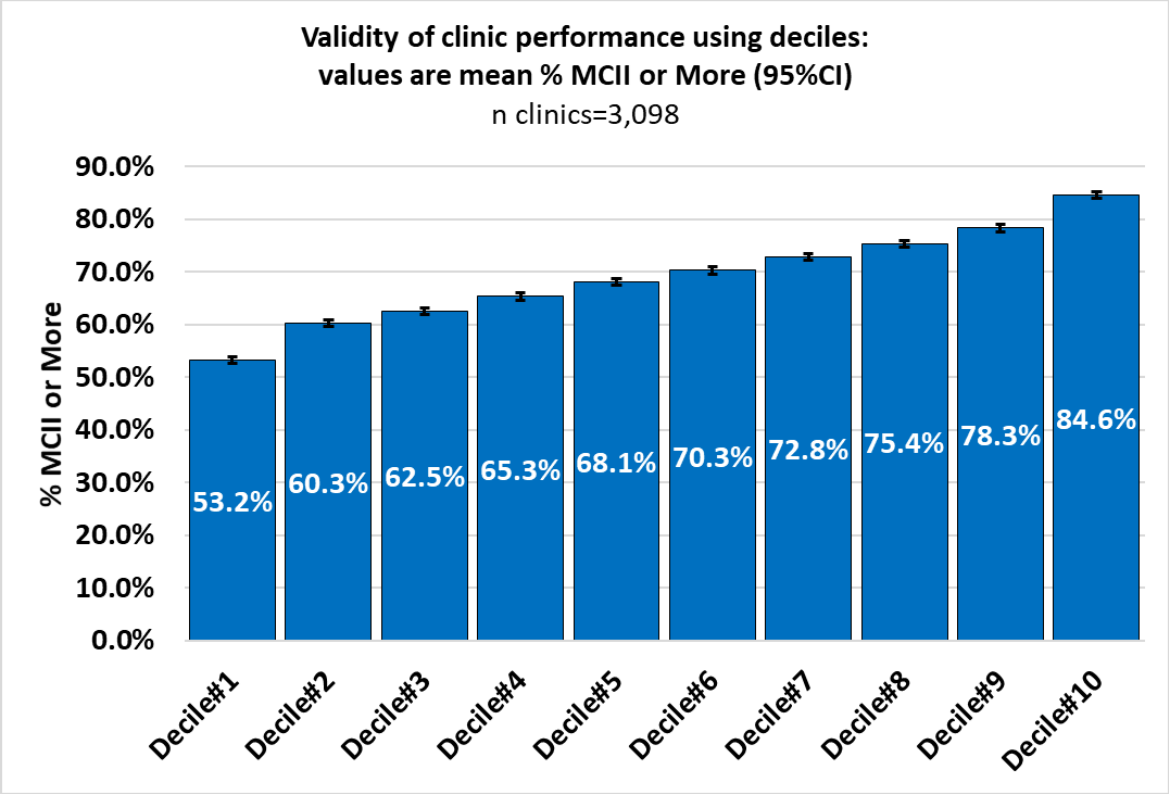
Method 2: Validity of clinic performance based on percentile ranking (10 levels):
The ten performance levels had statistically significant differences between groups as determined by one-way ANOVA ($F(9,3088) = 746.2$, $p < 0.001$), with a monotonic increase in rates of MCII achievement (TABLE 2b1.3viii-b).

TABLE 2b1.3viii-b: Validity of Performance at the Clinic Level Using Decile Ranking

Performance level	N Clinics (%)	% MCII or more
Decile 1	310	53.2%
Decile 2	310	60.3%
Decile 3	310	62.5%
Decile 4	310	65.3%
Decile 5	309	68.1%
Decile 6	310	70.3%
Decile 7	310	72.8%
Decile 8	310	75.4%
Decile 9	310	78.3%
Decile 10	309	84.6%

A Tukey post-hoc test revealed that all groups were significantly different from one another ($p < 0.001$) (FIGURE 2b1.3viii-b).

FIGURE 2b1.3viii-b: Validity of Performance at the Clinic Level Using Decile Ranking



2b1.3ix: Empirical Validity: Correlation of Performance Scores and Two External Markers

For Low Back PRO-PM correlations with the GROC, a sample of 202 clinics and 924 clinicians were included. Low Back PRO-PM correlations with the ODQ, a sample of 208 clinics and 669 clinicians were included. Absolute correlations for the two measures and provider levels ranged from 0.62 to 0.78 (see TABLE 2b1.3ix below) andGr were highly significant (P<0.001).

TABLE 2b1.3ix: Correlation of Performance Scores and Two External Markers

Correlation with provider level Low Back PRO-PM risk-adjusted outcomes (residuals)

Markers	mean patient-reported GROC* at discharge		mean ODQ** change (discharge-admission)	
Provider level	Clinic level	Clinician level	Clinic level	Clinician level
N Patients	53,020	50,191	23,652	21,966
N Clinics	202	239	208	243
N Clinicians	1,483	924	1,099	669
N States	35	35	39	40
Pearson correlation coefficient	0.78	0.70	-0.69	-0.62

*GROC; global rating of change

**ODQ; Modified Oswestry Low-Back-Pain Disability Questionnaire (negative change represents a positive outcome)

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?*)

2b1.4i: INTERPRETATION: VALIDITY OF DATA ELEMENTS (patient) level

Results produced multiple levels of validity evidence including content and structural validity. Retained items demonstrated essential uni-dimensionality, local independence and item fit. The expanded item bank was supported for its uni-dimentionality. 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.

2b1.4ii: INTERPRETATION: VALIDITY OF PERFORMANCE MEASURE 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 increased monotonically between consecutive performance levels, supporting clinically logical expectations.

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

Furthermore, the correlations with two external markers (GROC and ODQ) reported above were interpreted as moderate to high,[31] confirming our hypothesis that that the Low Back PRO-PM would be strongly correlated with both external measures, supporting its construct and concurrent validity.

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 Low Back FS PROM was designed and tested for 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 Low Back 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).[14] 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 Low Back FS PROM 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 0.986 (P<0.001), with an ICC(2,1) of 0.985 (P<0.001). FIGURE 2b2.2 plots the association between these two sets of residuals.

FIGURE 2b2.2: Age Exclusion Sensitivity Analysis

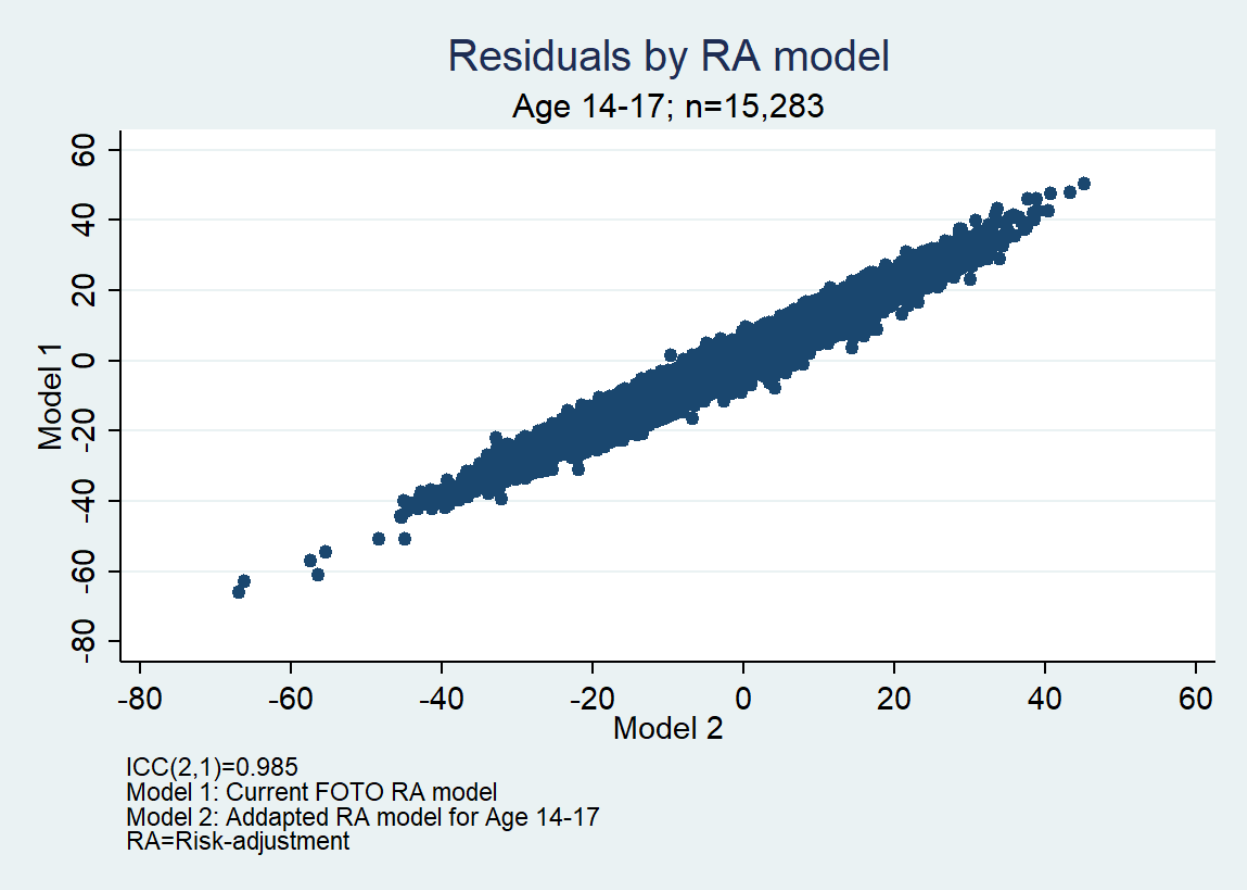


TABLE 2b2.2 compares the coefficients from model 1 & 2. As described above, only significant coefficients were allowed to enter model 2.

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	652,675			15,283		
Adjusted R-squared	37.8%			20.7%		
Independent variables	Beta	95% CI		Beta	95% CI	
		Lower	Upper		Lower	Upper
Constant	43.7	43.5	44.0	53.1	49.5	56.7
Initial evaluation FS	0.6	0.6	0.6	0.5	0.4	0.5
Age (continuous)	-0.1	-0.1	-0.1	-0.2	-0.4	0.0
Gender: Female	-0.6	-0.7	-0.6	-2.8	-3.3	-2.3
Acuity						
0-7 days	12.1	11.9	12.3	10.0	8.8	11.1
8-14 days	8.7	8.6	8.9	7.0	6.1	8.0

Dependent Variable: FS at discharge	Model 1: All ages (14-89)			Model 2: Age 14-17		
15-21 days	6.6	6.5	6.8	5.9	5.0	6.7
22-90 days	4.1	4.0	4.2	4.2	3.6	4.8
91 days to 6 months	1.7	1.6	1.9	2.2	1.5	2.9
Over 6 months (Ref)						
Payer						
Indemnity	-3.2	-3.4	-3.0	-3.6	-4.7	-2.6
Medicaid	-4.5	-4.6	-4.4	-1.4	-2.0	-0.7
Medicare A	-2.1	-2.3	-1.8			
Medicare B, age 65 or above	-3.0	-3.2	-2.8			
Medicare B, under age 65	-4.3	-4.6	-4.1			
No fault, Auto	-1.0	-1.1	-0.9			
Other (Litigation, Medicare C, School, No charge, Early Intervention, Commercial Insurance)	-5.9	-6.1	-5.8	-1.1	-1.8	-0.5
Workers' compensation	-3.2	-3.4	-3.0	-6.6	-11.1	-2.2
HMO, preferred provider (Ref)						
Surgical history						
1 related surgery	-1.9	-2.0	-1.8	-3.9	-5.3	-2.4
2 related surgeries	-3.0	-3.1	-2.8	-3.1	-6.7	0.4
3 or more related surgeries	-3.9	-4.1	-3.7	-8.4	-12.6	-4.1
No related surgery (Ref)						
Exercise history						
At least 3x/week	1.6	1.5	1.7	1.2	0.6	1.8
1-2x/week	0.8	0.8	0.9	-0.1	-0.8	0.6
Seldom or Never (Ref)						
Medication use at initial evaluation	-1.3	-1.4	-1.2			
Previous treatment	-1.7	-1.8	-1.6	-1.0	-1.5	-0.5
Low Back surgery procedure (No surgical codes)						
Fusion	1.4	1.1	1.6			
Laminectomy/Foramenectomy/Discectomy	2.0	1.7	2.3			
Comorbidities						
Angina	-0.4	-0.6	-0.1			
Anxiety	-1.0	-1.1	-0.9	-1.4	-2.1	-0.7
Arthritis	-1.2	-1.2	-1.1	-3.1	-4.7	-1.4
Asthma	-0.2	-0.3	-0.1	-0.6	-1.2	0.0
Chronic Obstructive Pulmonary Disease (COPD)	-1.0	-1.2	-0.9			
Depression	-1.1	-1.2	-1.0			
Diabetes Type I or II	-0.7	-0.8	-0.6			
Headaches	-1.3	-1.4	-1.2	-1.6	-2.1	-1.1
Incontinence	-0.9	-1.0	-0.7	-7.2	-12.0	-2.4
Kidney, Bladder, Prostate or Urination	-0.4	-0.6	-0.3			
Neurological Disease	-1.6	-1.8	-1.3	-6.4	-10.7	-2.2

Dependent Variable: FS at discharge	Model 1: All ages (14-89)			Model 2: Age 14-17		
Obesity (BMI>=30)	-0.8	-0.9	-0.7	-0.9	-1.6	-0.2
Osteoporosis	-0.6	-0.7	-0.5	-8.4	-15.8	-1.1
Previous accidents	-0.6	-0.7	-0.5			
Sleep dysfunction	-1.1	-1.2	-1.1	-2.5	-3.5	-1.5
Stroke	-0.6	-0.8	-0.4			
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 significant coefficients from the two models used to calculate the two sets of residuals 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, specific Low Back surgery procedure, specific comorbidities). Overall, we interpret these results as supporting the validity of the current FOTO risk-adjustment model for the Low Back 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** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

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.

Since the endorsement of NQF 0425 in the year 2015 we have updated the risk-adjustment (RA) model to include additional significant predictors of FS outcomes that were also found to be clinically relevant. These included: exercise history, having received previous treatment for the same condition, use of medication for the condition being treated, specific post-surgical types, and the inclusion of specific comorbidities as opposed to using a comorbidity index as recommended previously.[46] Additionally, the previously adjusted factor of fear avoidance beliefs was removed from the RA model due to new findings of low predictive power and a clinical consideration of this factor being modifiable during treatment.

The methods used to develop the FOTO risk-adjustment Low Back model were described in detail in a recent publication by Deutscher et at, 2018.[14] Briefly, we used data from patients with Low Back impairments treated in outpatient physical therapy clinics during 2014-2016 that had complete outcomes data at initial evaluation 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 initial evaluation (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 initial evaluation for the treatment of LBP (yes/no), previous treatment for LBP (yes/no), post-surgical type (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. To adjust for the large sample size, variables were entered if significance of their T value was less than 0.005 (entry level) and removed if significance was greater than 0.01 (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.[3] 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.[41] 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 above in section 2b3.1.1 and in further detail in a recent publication by Deutscher et at, 2018.[14]

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.[12, 19, 22] We selected non-modifiable patient factors to avoid misclassification of provider performance and control for their relationships with outcomes of interest.

Social factors

Payer:

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.

Education level:
A standard data point to ask all respondents their level of education was added for a limited period of time to the FOTO system during the year 2018. Because this was a standard question asked of all patients, we acquired a large sample size for this variable enabling us to test its impact on the RA model. The education item included 9 response categories for 8 educational levels and one response being 'prefer not to answer' for those who would not feel comfortable responding (see TABLE 2b3.4b). To test the contribution of the educational level variable to the existing RA model, we conducted an ordinary least square linear regression model for the sample of patients that had complete outcomes data at both admission and discharge and had also completed the educational variable during its testing period.

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 37.3%.

TABLE 2b3.4a: Risk-adjusted model: **Associations between patient characteristics at initial evaluation and FS at discharge.**

Significant Predictors of FS at Discharge (Reference group for categorical variables)	β†	t‡
Intercept	42.4 (42.1, 42.7)	280.9
FS score at initial evaluation§	0.6 (0.6, 0.6)	320.6
Age	-0.1 (-0.1, -0.1)	-69.3
Sex: Female	-0.3 (-0.4, -0.3)	-8.0
Acuity (Over 6 months)		
0-7 days	12.5 (12.3, 12.7)	116.4

Significant Predictors of FS at Discharge (Reference group for categorical variables)	β†	t‡
8-14 days	9.2 (9.0, 9.3)	105.8
15-21 days	7.0 (6.8, 7.1)	88.0
22-90 days	4.2 (4.1, 4.3)	78.7
91 days to 6 months	1.8 (1.7, 1.9)	27.7
Payer (HMO, Preferred Provider)		
Indemnity insurance	-2.6 (-2.9, -2.4)	-22.5
Medicaid	-4.7 (-4.9, -4.5)	-47.7
Medicare A	-1.4 (-1.7, -1.1)	-8.4
Medicare B Under Age 65	-3.0 (-3.2, -2.8)	-28.2
No fault, Auto insurance	-4.2 (-4.5, -3.8)	-25.3
Workers compensation	-5.7 (-5.9, -5.5)	-64.0
Other (Litigation, Medicare C, School, No charge, Early Intervention, Commercial Insurance)	-1.1 (-1.3, -1.0)	-15.0
Surgical history (No related surgery)		
1 related surgery	-1.8 (-1.9, -1.7)	-27.4
2 related surgeries	-2.9 (-3.1, -2.6)	-26.3
3 or more related surgeries	-3.7 (-4.0, -3.5)	-29.9
Exercise history (Seldom or Never)		
At least 3x/week	1.3 (1.2, 1.4)	27.0
1-2x/week	0.6 (0.5, 0.7)	12.1
Medication use at initial evaluation	-1.3 (-1.4, -1.2)	-29.9
Previous treatment	-1.5 (-1.6, -1.5)	-36.4
Low Back surgery procedure (No surgical codes)		
Fusion	1.5 (1.2, 1.9)	9.2
Laminectomy/Foramenectomy/Discectomy	2.3 (1.9, 2.6)	13.5
Specific comorbidities:		
Angina	-0.6 (-1.0, -0.3)	-3.9
Anxiety	-0.9 (-1.1, -0.8)	-14.7
Arthritis	-1.1 (-1.2, -1.0)	-23.2
Asthma	-0.3 (-0.4, -0.1)	-3.9
Chronic Obstructive Pulmonary Disease (COPD)	-1.0 (-1.2, -0.8)	-9.5
Depression	-1.1 (-1.2, -1.0)	-18.1
Diabetes Type I or II	-0.6 (-0.7, -0.5)	-10.0
Headaches	-1.2 (-1.3, -1.1)	-23.3
Incontinence	-0.8 (-1.0, -0.6)	-9.1
Kidney, Bladder, Prostate or Urination	-0.4 (-0.5, -0.2)	-5.6
Neurological Disease	-1.3 (-1.6, -1.0)	-8.7
Obesity (BMI>=30)	-0.6 (-0.7, -0.5)	-14.5
Osteoporosis	-0.5 (-0.6, -0.4)	-7.3
Previous accidents	-0.5 (-0.6, -0.4)	-8.7
Sleep dysfunction	-1.2 (-1.3, -1.1)	-22.3
Stroke	-0.5 (-0.7, -0.3)	-4.6

Significant Predictors of FS at Discharge (Reference group for categorical variables)	β†	t‡
Abbreviations: BMI, body mass index; FS, functional status; HMO, health maintenance organization. Number of patients, n =414,125. Adjusted R-squared=37.3% †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. Values in parentheses are 95% confidence interval. ‡t values indicate the importance of each independent variable for predicting discharge FS (dependent variable). All t values were significant at the 0.001 level. §Higher FS scores represent higher level of functioning.		

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.

The sample of patients that had responded to the temporarily introduced education variable included 41,889 patients with a mean age of 57(SD=17.8) ranging from 14 to 89 years of age, and 60% were female. Their mean (SD) FS scores at admission and discharge were 48.9(13.6) and 63.4(16.9), respectively. These characteristics were practically identical to those of the full sample of patients that had complete outcomes data (see TABLE 1.6.VI: Patients with FS measures at initial evaluation & discharge aged 14 to 89, N = 625,675), supporting the external validity of this partial sample. The response rates for each education response category are presented in Table 2b3.4b below. Since all categories were above 1%, we determined that they could each be tested as a separately for the education level construct. The lowest level of education (less than high-school degree) was set as the reference standard for the regression model.

Results showed that all educational levels were not significant except for Bachelor's degree with only being borderline significant [unstandardized beta coefficient=0.7 (95%CI=0.09 to 1.3), P=0.025]. The squared semi-partial correlation for bachelor's degree, representing the amount of explained variance decrease if the variable would be removed from the model, was 0.01%. The category of 'prefer not to answer' was also borderline significant [unstandardized beta coefficient=0.8 (95%CI=0.00 to 1.51), P=0.049]. These preliminary results do not support an important contribution of the education variable to the prediction of FS at discharge, after having controlled for all other variables already included in the RA model. However, additional testing would be required, possibly

testing collapsed groups of educational level, before making a final conclusion on its appropriateness as a social risk factor that needs to be adjusted for.

TABLE 2b3.4b: Education Levels Tested as a Social Risk factor (N=41,889)

Level of education	Frequency
Less than high-school degree	6%
High-school degree or equivalent	22%
Trade/technical/vocational training	6%
Some college but no degree	16%
Associate degree	8%
Bachelor's degree	20%
Master's degree	11%
Other advanced degree beyond a Master's	4%
Prefer not to answer	6%

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 [3] and the predictive ratio [22]. 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.[3] 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 [52] was applied using the 'regvalidate' STATA program.[15] 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.[36] 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.[22, 36]

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 37.3% of the variance in discharge FS, with FS at initial evaluation, acuity, payer type and age being the most important predictors.
R-squared shrinkage was less than 0.2% for both methods used to assess shrinkage.
The average predicted discharge FS of the test sample (n/2= 207,063), estimated using the development sample, was practically identical to the average actual discharge FS obtained by the test sample (62.743 and 62.737, 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.

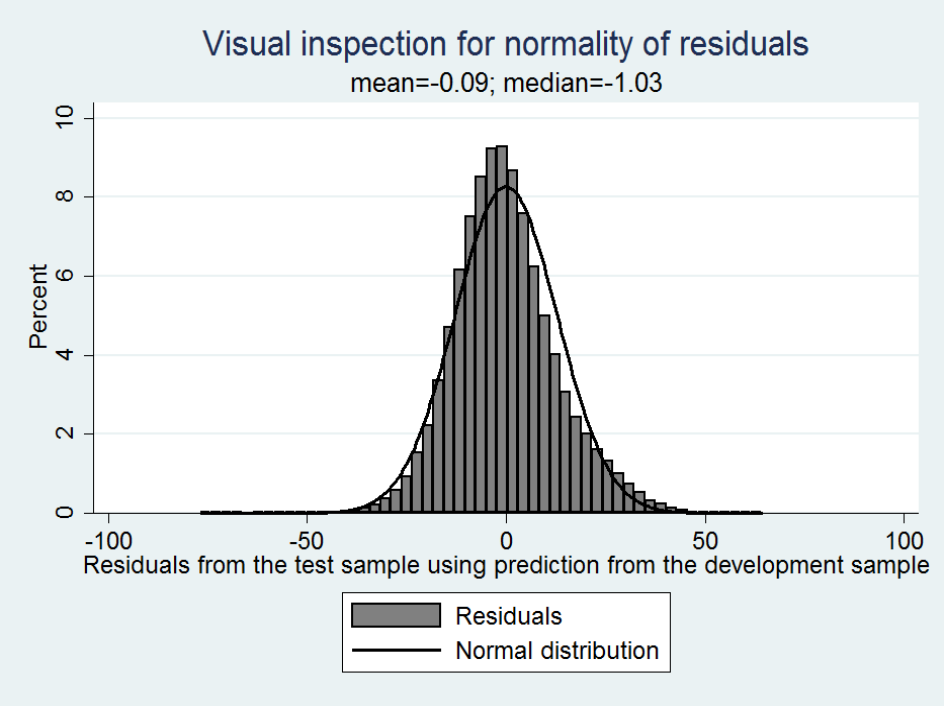
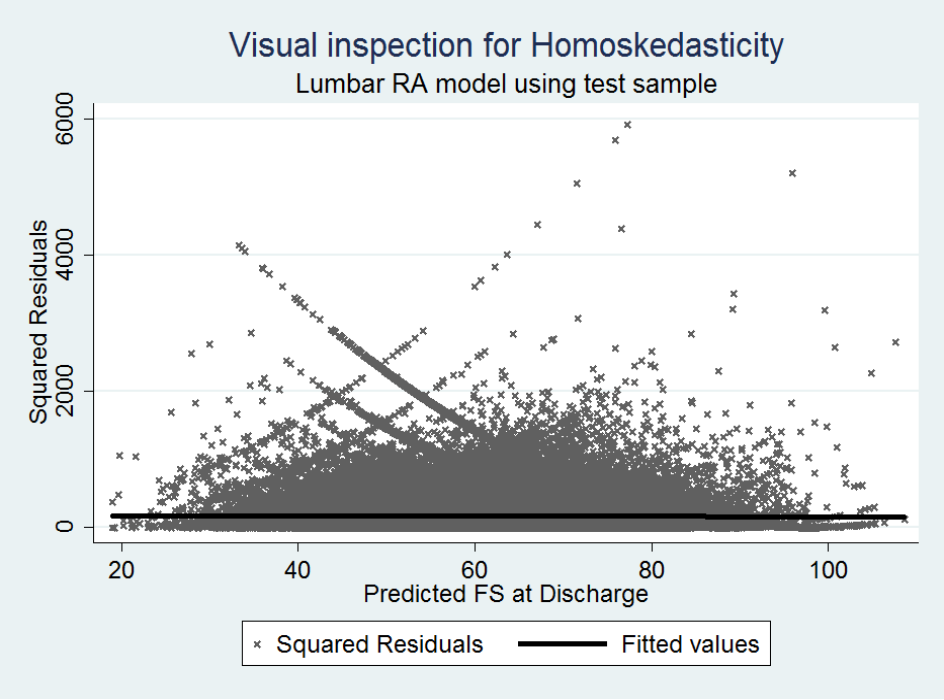


FIGURE 2b3.6ii: Visual Inspection of Homoscedasticity

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*)

As noted above, we are not aware of an agreed upon value for an acceptable level of shrinkage, but we considered a shrinkage of less than 1% to strongly support the model's external validity. Along with the predictive ratio 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.[49] In this method, complete cases are weighted by the inverse of their probability of being a complete case.[51] 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.[49] We compared the coefficients created by the un-weighted and weighted models.

All unstandardized beta coefficients were practically identical when using un-weighted and weighted models, with differences ranging from 0-0.22 on the 0-100 scale range. This result supported that missing data are mostly missing at random and that the risk-adjustment model was not impacted by missing data.

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)*

2b4.1i. Data Elements (patient) Level:

Performance of the Low Back FS-PROM at the patient level was assessed by calculating the patient's risk-adjusted residual score (residual = actual change – predicted change) after risk-adjusted 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, using the full sample of patients that had complete outcomes data (N= 652,675). 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.

To assess OUTCOMES OVER TIME, we compared residual scores by year (2016-2018). Since each consecutive year included new clinicians that used the FOTO system to collect PROM data, to assess whether patient outcomes improved when treated by clinicians that had experience using PROM data over the three-year period, we repeated this comparison only for patients treated by clinicians that contributed data during all three years. A one-way ANOVA was conducted to determine if the mean residuals were different between years, followed by A Tukey post-hoc to assess if differences between years were significant.

2b4.1ii-iii. Clinician & Clinic Performance Score Level:

Patient level residual scores were aggregated to a provider level 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).

To assess PERFORMANCE OVER TIME at the provider levels (clinicians and clinics), we analyzed data from the years 2016 to 2018 to compare aggregated residual scores by year and individual provider, only for providers (clinics or clinicians) that contributed data during all three years. A one-way ANOVA was conducted to determine if the mean residuals were different between years, followed by a Tukey post-hoc to assess if differences between years were significant.

To assess PERFROMANCE GAP for the overall data collection period and by year, providers (clinicians or clinics) were ranked into deciles by their average residual scores. For each decile, the mean

residuals for all clinics categorized within each decile rank was calculated. The overall performance gap is represented by differences in mean residuals between the 1st and 10th decile.

For these analyses, at the clinic level, performance over time 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 over time was evaluated only for clinicians that had a minimum of 10 patients per calendar year.

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*)

2b4.2i Performance at the Patient Level

The mean residual score by year for the overall sample, and for the sample of patients treated by clinicians that contributed data to all three years are presented in TABLE 2b4.2i below. For the full sample, mean residuals were significantly different between years (one-way ANOVA ($F(2,652672) = 208.3$, $p < 0.001$) with a monotonic increased from -0.4 to +0.4 over time. For the sub-sample of patients treated by clinicians that contributed data to all three years, mean residuals were also significantly different between years (one-way ANOVA ($F(2, 346927) = 255.5$, $p < 0.001$) with a monotonic increased from -0.1 to +1.1 over time ($P<0.001$).

TABLE 2b4.2i: Performance at the patient level

Year	All patients with complete outcomes data during 2016-2018		Patients treated by clinicians that contributed data to all three years	
Year	N (%)	Mean Residuals±SD (95%CI) Minimum-Maximum	N (%)	Mean Residuals±SD (95%CI) Minimum-Maximum
2016	183,113 (28)	-0.4±12.9 (-0.43 to -0.31) -77.2 to 64.3	106,618 (31)	-0.1±12.9 (-0.18 to -0.02) -77.2 to 60.8
2017	217,651 (33)	-0.2±13.1 (-0.21 to -0.10) -77.7 to 69.7	131,701 (38)	0.3±13.0 (0.19 to 0.333) -66.2 to 69.7
2018	251,911 (39)	0.4±13.3 (0 .35 to 0.46) -73.5 to 61.1	108,611 (31)	1.1±13.2 (1.05 to 1.21 -61.5 to 61.1

Year	All patients with complete outcomes data during 2016-2018		Patients treated by clinicians that contributed data to all three years	
Total	652,675	0.0±13.2 (-0.03 to 0.03) -77.7 to 69.7	346,930	0.42±13.0 (0.38 to 0.47 -77.2 to 69.7

Abbreviations: CI=confidence interval

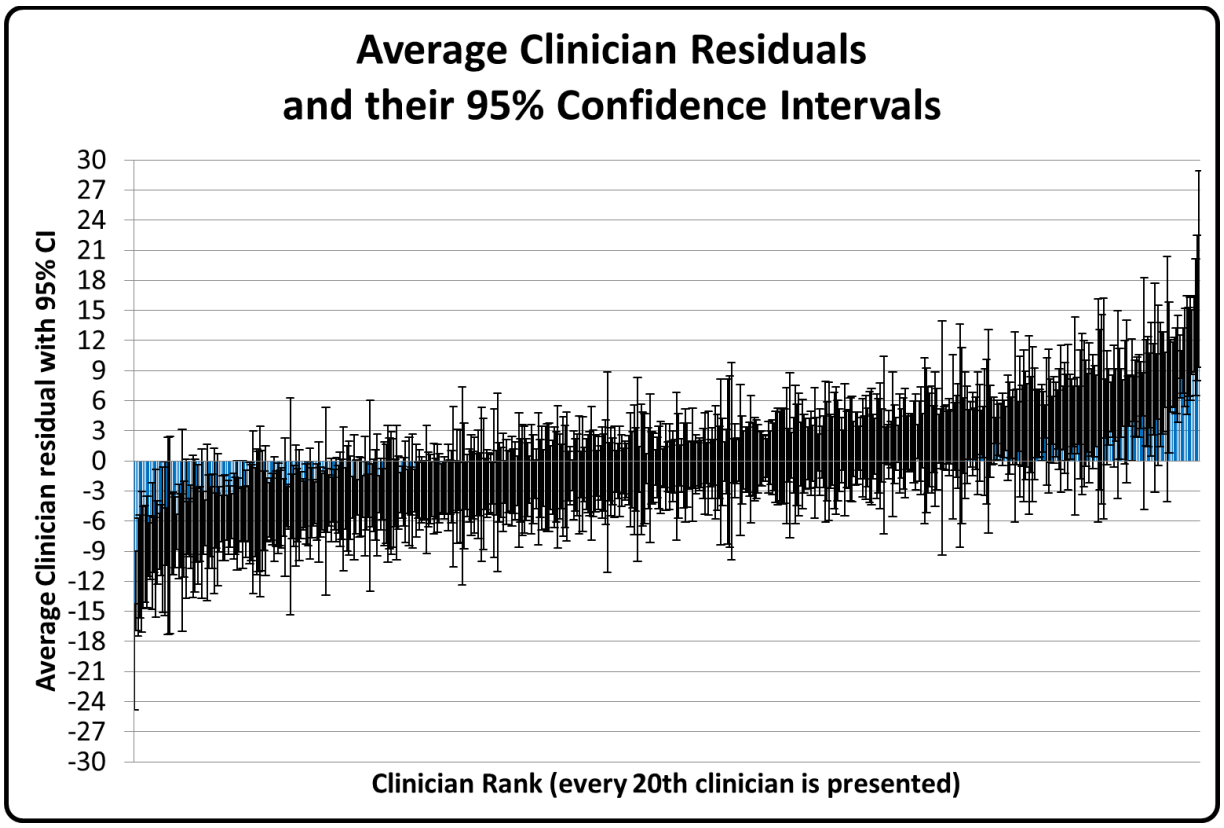
2b4.2ii. Performance Individual Clinician Level

The distribution of clinician performance by 3 distinct quality levels or by deciles of average residuals are presented in the validity testing section above (2b1.3viii-a&b).

Clinician performance based on uncertainty assessments is summarized in TABLE 2b1.3viii-a above and illustrated in FIGURE 2b4.2iia below, with 18%, 68% and 14% of clinicians achieving low, average and high performance, respectively.

Clinician performance based on percentile ranking (deciles) is summarized in TABLE 2b1.3viii-b above and illustrated in FIGURE 2b1.3viii-b above, 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 87% of patients achieving the MCII, leaving room for additional improvement even at that high-performance level.

FIGURE 2b4.2iia: Average Clinician Residual (95%CI)

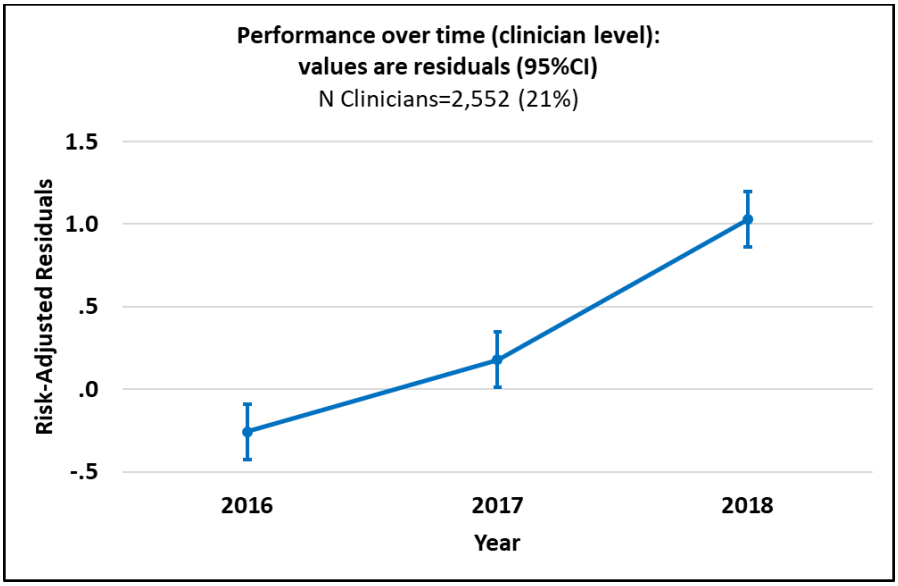


CLINICIAN PERFORMANCE OVER TIME assessed by the mean residual score by year for the sample of clinicians that contributed data to all three years are presented in TABLE 2b4.2iia and FIGURE 2b4.2iib below. There were 21% of all clinicians assessed that contributed data for all 3 years. Mean residuals were different between years (one-way ANOVA ($F(2,7653) = 58.2$, $p < 0.001$) with a monotonic and significant increased from -0.3 to +1.0 over time.

TABLE 2b4.2iia: Performance at the Clinician Level Over Time

Year	Mean Residuals ± SD	(95%CI) Minimum-Maximum
2016	-0.3±4.3	(-0.42 to -0.09) -13.6 to 26.4
2017	0.2±4.2	(0.01 to 0.35) -16.1 to 23.0
2018	1.0±4.5	(0.86 to 1.20) -13.3 to 22.7

FIGURE 2b4.2iib: Performance at the Clinician Level Over Time



CLINICIAN PERFORMANCE GAP is demonstrated in table 2b4.2iib below. Overall, average residual scores by clinic ranks based on deciles of their average residual scores ranged from -7.1 to +7.6 for 1st and 10th decile ranks, respectively. Over the three-year period assessed, performance gap between 1st and 10th decile ranks were from -7.6 to +7.9 in 2016 to -7.0 to +8.8 in 2018.

TABLE 2b4.2iib: Performance Gap at the Clinician Level Over Time

Performance gap over time (years) at the clinician level

Decile ranking by average clinic residuals	2016 (5,772 clinicians)	2017 (6,800 clinicians)	2018 (7,899 clinicians)	Total (12,025 clinicians)
1	-7.6	-7.5	-7.0	-7.1
2	-4.9	-4.7	-4.3	-4.4
3	-3.5	-3.3	-2.8	-3.1
4	-2.3	-2.1	-1.6	-2.0
5	-1.3	-1.0	-0.5	-1.0
6	-0.2	-0.1	0.6	-0.1
7	0.8	1.1	1.7	1.0
8	2.1	2.4	3.1	2.2
9	3.8	4.1	4.8	3.9
10	7.9	8.1	8.8	7.6
Total	-0.5	-0.3	0.3	-0.3

Values are mean residuals by deciles of average clinician residuals.
Residuals represent the difference between actual and predicted outcomes at discharge.

A residual of 0 represents no difference between actual and predicted outcomes.
Higher residuals represent better outcomes.

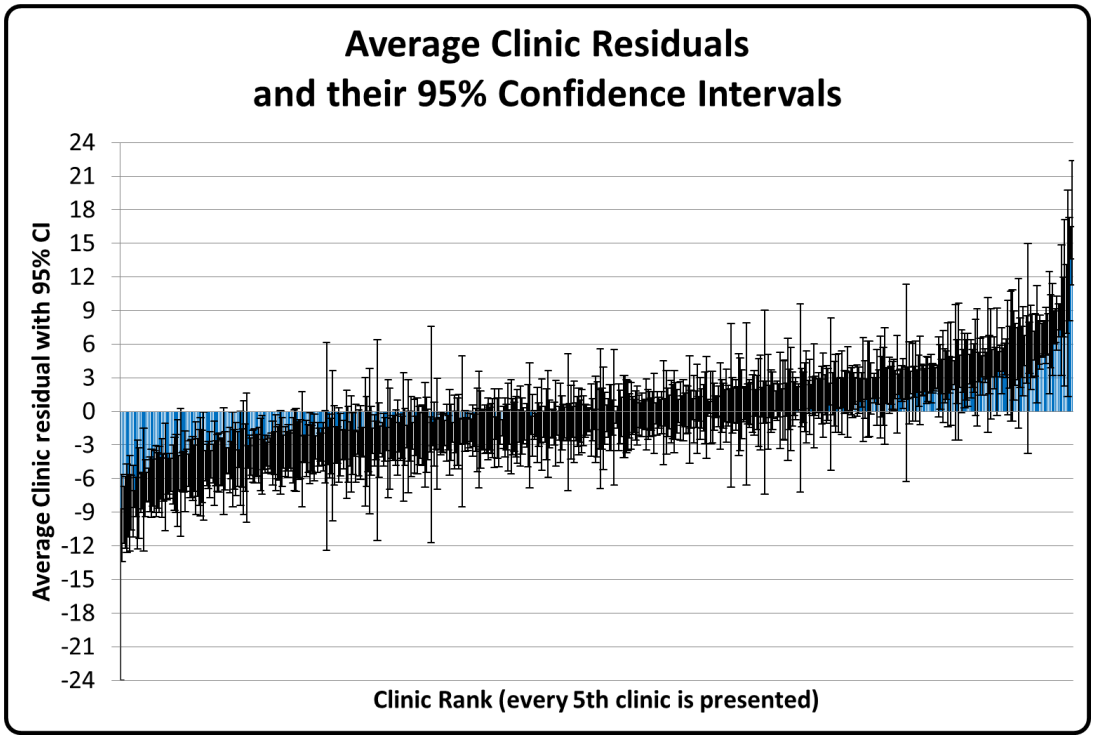
2b4.2iii. Performance Clinic (Group Practice) level

The distribution of clinic performance by 3 distinct quality levels or by deciles of average residuals are presented in the validity testing section above (2b1.3ix-a&b).

Clinic performance based on uncertainty assessments summarized in TABLE 2b1.3ix-a above (Validity Testing section), and illustrated in FIGURE 2b4iiia below, with 29%, 52%, and 19% of clinic achieving low, average, and high performance, respectively.

Clinic performance based on percentile ranking (deciles) is summarized in TABLE 2b1.3ix-b above and illustrated in FIGURE 2b1.3ix-b above, showing monotonic increase between ranks of rates of patients achieving the minimal clinically important improvement (MCII). Also, clinics at the highest performance rank had on average only 85% of patients achieving the MCII, leaving room for improvement even at that high-performance level.

FIGURE 2b4.2iiia: Average Clinic Residual (95%CI)



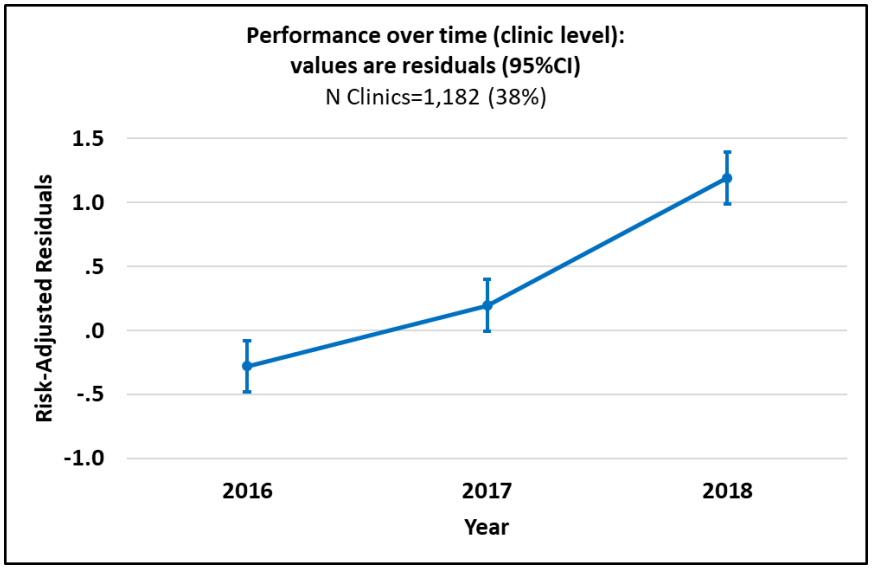
CLINIC PERFORMANCE OVER TIME assessed by the mean residual score by year for the sample of clinics that contributed data to all three years are presented in TABLE 2b4.2iiia and FIGURE 2b4.2iiib below. There were 38% of all clinics assessed that contributed data for all 3 years. Mean residuals were different between years (one-way ANOVA ($F(2,3543) = 53.6$, $p < 0.001$) with a monotonic and significant increased from -0.3 to +1.2 over time.

TABLE 2b4.2iiia: Performance at the Clinic Level Over Time

Clinic performance (n=1,182)

Year	Mean Residuals±SD	(95%CI) Minimum-Maximum
2016	-0.3±3.56	(-0.48 to -0.08) -11.5 to 21.1
2017	0.2±3.40	(-0.01 to 0.40) -9.2 to 20.6
2018	1.2±3.62	(0.99 to 1.39) -10.2 to 17.7

FIGURE 2b4.2iiib: Performance at the Clinic Level Over Time



CLINIC PERFORMANCE GAP is demonstrated in table 2b4.2iiib below. Overall, average residual scores by clinic ranks based on deciles of their average residual scores ranged from -6.2 to +6.3 for 1st and 10th decile ranks, respectively. Over the three-year period assessed, performance gap between 1st and 10th decile ranks were from -6.8 to +6.6 in 2016 to -5.8 to +7.6 in 2018.

TABLE 2b4.2iiib: Performance Gap at the Clinic Level Over Time

Performance gap over time (years) at the clinic level

Decile ranking by average clinic residuals	2016 (1,757 clinics)	2017 (2,029 clinics)	2018 (2,440 clinics)	Total (3,098 clinics)
1	-6.8	-6.3	-5.8	-6.2
2	-4.3	-3.9	-3.5	-3.8
3	-3.1	-2.6	-2.3	-2.7

Decile ranking by average clinic residuals	2016 (1,757 clinics)	2017 (2,029 clinics)	2018 (2,440 clinics)	Total (3,098 clinics)
4	-2.1	-1.8	-1.3	-1.8
5	-1.2	-0.9	-0.4	-1.1
6	-0.4	-0.1	0.4	-0.2
7	0.5	0.8	1.3	0.7
8	1.5	1.8	2.5	1.7
9	2.9	3.2	4.1	3.0
10	6.6	6.6	7.6	6.3
Total	-0.6	-0.3	0.3	-0.4

Values are mean residuals by deciles of average clinic residuals.
Residuals represent the difference between actual and predicted outcomes at discharge.
A residual of 0 represents no difference between actual and predicted outcomes.
Higher residuals represent better outcomes.

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 NQF 0425 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. This interpretation was also supported by the results of the analyses of performance over time. Findings demonstrated significant improvement over the three-year period assessed, at both the patient and provider (clinicians and clinics) levels.

Additionally, analyses of performance gap demonstrated important differences in clinic risk-adjusted residual scores between lower and higher ranked providers (clinicians and clinics), providing additional support for the ability of NQF 0425 to serve as a useful measure for quality improvement initiatives.

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

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

NA, only one set of measure specifications

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*)

We addressed the assessment of potential bias due to missing outcomes data using 3 methods: 1) comparing patients with or without complete outcomes, 2) assessing correlations between clinician and clinic residuals and their completion rates, and 3), assessing average residuals at the clinician or clinic levels by completion rate categories with or without an adjustment using inverse probability weighting. These methods are described below.

2b6.1i. 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 initial evaluation 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 standpoint, 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.[12-14] If no specific trends are identified, the assumption of missing data largely at random is supported, reducing concern that systematic patient selection bias exists.

Thus, if a systematic patient selection bias at discharge existed, we expected that patients with complete PROM 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 (initial evaluation only) and complete (initial evaluation 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 initial evaluation because they are known to be the strongest positively associated predictor of outcomes, i.e., higher FS at initial evaluation 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 initial evaluation for the treatment of the Low Back pain (yes/no); and having received previous treatment for Low Back pain (yes/no), both associated with lower outcomes.[11, 12, 14, 22]

2b6.1ii-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 providers systematically selected “good outcome” patients to complete surveys at discharge and avoided having “poor outcome” patients complete surveys at discharge. We examined Pearson Correlations between clinician and clinic completion rate and their average residual scores. Only providers that passed the threshold for inclusion in the FOTO benchmarking process were included in this analysis (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).

2b6.1iv-v. Average Residuals at the Clinician or Clinic Levels by Completion Rate Categories With or Without the Use of Inverse Probability Weighting

For this method, we tested the impact of a weighted adjustment for missing data using inverse probability weighting (IPW).[49] In this method, complete cases are weighted by the inverse of their probability of being a complete case.[51] Hence, patients less likely to have complete FS data are given more weight in the analyses of interest than those who are more likely to have complete data.[49] We examined 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). For this, we grouped clinicians and clinics into 10 completion rate categories. Only providers that passed the threshold for inclusion in the FOTO benchmarking process were included in this analysis (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).

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*)

The frequency of missing data was 33% as shown below in TABLE 2b6.2i. This rate of missing data is typical for data collected from real-life outpatient settings in which patients often stop attending treatment visits before an episode of care is completed.[12-14]

Distribution of missing data across providers is demonstrated below in tables 2b6.2iv-v as number of providers (clinicians or clinics) by completion rate categories.

2b6.2i. Comparing Patients With or Without Complete Outcomes

The comparison of patients with complete and incomplete FS outcomes data is presented in TABLE 2b62i. Due to the extremely large patient group included, our interpretation was clinically rather than statistically driven. No important differences between groups were identified for initial evaluation FS, and sex. Compared to patients with incomplete outcomes data, patients with complete outcomes data were 4 years older, had more comorbidities, had a slightly higher rate of surgical history and previous treatment for Low Back impairment, not supporting a systematic patient selection bias that would have positively biased outcomes since these differences are known to be associated with lower outcomes. However, these patients also had a slightly lower rate of chronicity and Medicaid payer, had a higher rate of Medicare Part B for ages 65 or above, exercised slightly more and used less medications related to their Low Back pain at initial evaluation, which might have biased outcomes in favor of this patient group. Overall, these analyses were inconclusive and did not support a systematic patient selection bias.

TABLE 2b6.2i: Health and Demographic Patient Characteristics of Those With Complete or Incomplete FS Outcomes Data*

Patient characteristics	Complete (n= 652,675; 67%)	Incomplete (n= 324,480; 33%)	p-value†
FS score at initial evaluation: Mean ± SD (Min to Max)	48.8±13.1 (0-98)	48.4±13.7 (0-98)	<0.001‡
Age (years): Mean ± SD (Min to Max)	56.4±17.7 (14-89)	52.2±17.5 (14-89)	<0.001‡
Number of comorbidities: Mean ± SD (Median, Range)	5.0±3.2 (5, 0 to 27)	4.9±3.3 (4, 0 to 28)	<0.001‡
Sex: Female	59.9	60.2	0.003
Acuity:			<0.001

Patient characteristics	Complete (n= 652,675; 67%)	Incomplete (n= 324,480; 33%)	p-value[†]
0-7 days	3.9	4.3	
8-14 days	6.4	6.3	
15-21 days	8.0	7.8	
22-90 days	23.5	22.2	
91 days to 6 months	12.5	12.3	
Over 6 months	45.6	47.1	
Payer:			<0.001
Indemnity insurance	3.0	4.3	
Medicaid	5.4	8.7	
Medicare A	1.6	1.3	
Medicare B Under Age 65	3.6	4.4	
Medicare B Age 65 or above	26.7	17.0	
Patient	0.5	0.8	
Workers compensation	5.7	4.4	
Other (Litigation, Medicare C, School, No charge, Early Intervention, Commercial Insurance)	9.7	11.2	
No fault, Auto insurance	1.5	1.2	
HMO, PPO	42.3	46.7	
Surgical history:			<0.001
No related surgery	81.4	83.8	
1 related surgery	11.9	10.1	
2 related surgeries	3.8	3.4	
3 or more related surgeries	2.9	2.7	
Exercise history:			<0.001
At least 3x/week	38.2	37.0	
1-2x/week	24.6	24.8	
Seldom or Never	37.1	38.1	
Medication use at initial evaluation	52.1	54.4	<0.001
Previous treatment	49.6	47.6	<0.001

Abbreviations: FS, functional status; HMO, health maintenance organization; PPO, preferred provider organization

[*]Patient characteristics for patient with functional status data at initial evaluation and discharge (Complete) and patient with functional status data at initial evaluation 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.

2b6.2ii-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 0.003, respectively.

2b6.2iv-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 pattern, with no impact of IPW on the results.

TABLE 2b6.2iv: 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 clinicians	Residual without IPW	Residual with IPW
0-10	21	2	2.2	1.8
10-20	159	11	-1.2	-1.1
20-30	1,322	78	-0.6	-0.6
30-40	9,005	347	-0.3	-0.3
40-50	30,196	912	0.0	0.0
50-60	77,329	1,801	-0.1	0.0
60-70	134,607	2,746	-0.1	0.0
70-80	172,826	3,111	0.0	0.0
80-90	122,956	2,102	0.1	0.1
90-100	36,936	915	0.3	0.3
Total	585,357	12,025	0.0	0.0

TABLE 2b6.2v: 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
0-10	NA	NA	NA	NA
10-20	94	4	-2.2	-2.3
20-30	882	13	-0.2	-0.1
30-40	5,765	73	-0.1	-0.1
40-50	26,198	203	0.0	0.0
50-60	99,754	471	0.0	0.0
60-70	161,363	797	0.0	0.0
70-80	197,065	860	0.0	0.0
80-90	114,101	547	0.1	0.1
90-100	13,250	130	0.2	0.2
Total	618,472	3098	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 NQF 0425 data. However, we acknowledge that a potential selection bias may still exist based on factors not available in our dataset. 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. Adams JL. *The Reliability of Provider Profiling: A Tutorial*. RAND Corporation; 2009.
2. Andrich D. A rating formulation for ordered response categories. *Psychometrika*. 1978;43:561-573.
3. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med*. 2004;66:411-421.
4. Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. *J Clin Epidemiol*. 2001;54:1204-1217.
5. Bland JM, Altman DG. Cronbach's alpha. *BMJ*. 1997;314:572.
6. Brodke DS, Goz V, Lawrence BD, Spiker WR, Neese A, Hung M. Oswestry Disability Index: a psychometric analysis with 1,610 patients. *Spine J*. 2017;17:321-327.
7. Clement RC, Welander A, Stowell C, et al. A proposed set of metrics for standardized outcome reporting in the management of low back pain. *Acta Orthop*. 2015;86:523-533.
8. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. *Measurement in Medicine: A Practical Guide*. Cambridge: Cambridge University Press; 2011.
9. Delitto A, George SZ, Van Dillen L, et al. Low back pain. *J Orthop Sports Phys Ther*. 2012;42:A1-57.
10. Deutscher D, Cook KF, Kallen MA, et al. Clinical Interpretation of the Neck Functional Status Computer Adaptive Test. *J Orthop Sports Phys Ther*. 2019;1-34.
11. 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;91:1072-1084.
12. 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;90:1349-1363.
13. Deutscher D, Werneke MW, Gottlieb D, Fritz JM, Resnik L. Physical therapists' level of McKenzie education, functional outcomes, and utilization in patients with low back pain. *J Orthop Sports Phys Ther*. 2014;44:925-936.
14. 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.
15. Ender PB. regvalidate. Available at: http://www.philender.com/courses/linearmodels/notes2/cross.html. Accessed 2010.
16. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. *Physiotherapy*. 1980;66:271-273.
17. Fritz JM, Irrgang JJ. A comparison of a modified Oswestry Low Back Pain Disability Questionnaire and the Quebec Back Pain Disability Scale. *Phys Ther*. 2001;81:776-788.
18. 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;20:561-565.
19. Gozalo PL, Resnik LJ, Silver B. Benchmarking Outpatient Rehabilitation Clinics Using Functional Status Outcomes. *Health Serv Res*. 2016;51:768-789.
20. Hambleton RK. Good practices for identifying differential item functioning. *Med Care*. 2006;44:S182-188.
21. Hart DL. Assessment of unidimensionality of physical functioning in patients receiving therapy in acute, orthopedic outpatient centers. *J Outcome Meas*. 2000;4:413-430.
22. Hart DL, Connolly JB. *Pay-for-Performance for Physical Therapy and Occupational Therapy: Medicare Part B Services. Grant #18-P-93066/9-01*. Health & Human Services/Centers for Medicare & Medicaid Services.; 2006.
23. Hart DL, Mioduski JE, Werneke MW, Stratford PW. Simulated computerized adaptive test for patients with lumbar spine impairments was efficient and produced valid measures of function. *J Clin Epidemiol*. 2006;59:947-956.

24. Hart DL, Stratford PW, Werneke MW, Deutscher D, Wang YC. Lumbar computerized adaptive test and Modified Oswestry Low Back Pain Disability Questionnaire: relative validity and important change. *J Orthop Sports Phys Ther*. 2012;42:541-551.
25. 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;90:928-938.
26. 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;17:1081-1091.
27. 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;89:2129-2139.
28. 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;61:1113-1124.
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;35:2157-2164.
30. Hart DL, Wright BD. Development of an index of physical functional health status in rehabilitation. *Arch Phys Med Rehabil*. 2002;83:655-665.
31. Hinkle DE. *Applied statistics for the behavioral sciences*. 5th ed. Boston: Houghton Mifflin; 2003.
32. Holland PW, Wainer H. *Differential Item Functioning*. Hillsdale. NJ: Lawrence Erlbaum; 1993.
33. 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.
34. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials*. 1989;10:407-415.
35. Jette DU, Jette AM. Physical therapy and health outcomes in patients with spinal impairments. *Phys Ther*. 1996;76:930-941; discussion 942-935.
36. Kautter J, Ingber M, Pope GC, Freeman S. Improvements in Medicare Part D risk adjustment: beneficiary access and payment accuracy. *Med Care*. 2012;50:1102-1108.
37. 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.
38. Linacre JM. Detecting multidimensionality: which residual data-type works best? *J Outcome Meas*. 1998;2:266-283.
39. Linacre JM. *A User's Guide to WINSTEPS*. Chicago, IL: MESA Press; 2008.
40. Lord FM. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associated; 1980.
41. Maxwell SE. Sample size and multiple regression analysis. *Psychol Methods*. 2000;5:434-458.
42. Nunnally JC. *Psychometric theory*. New York,: McGraw-Hill; 1967.
43. Nunnally JC, Bernstein IH. *Psychometric theory*. 3rd. New York: McGraw-Hill; 1994.
44. 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;33:90-94.
45. 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:S22-31.
46. Resnik L, Gozalo P, Hart DL. Weighted index explained more variance in physical function than an additively scored functional comorbidity scale. *J Clin Epidemiol*. 2011;64:320-330.
47. Resnik L, Hart DL. Using clinical outcomes to identify expert physical therapists. *Phys Ther*. 2003;83:990-1002.
48. Riddle DL, Stratford PW, Binkley JM. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 2. *Phys Ther*. 1998;78:1197-1207.

49. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11:550-560.
50. Schmitt JS, Di Fabio RP. Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *Journal of Clinical Epidemiology*. 2004;57:1008-1018.
51. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res*. 2013;22:278-295.
52. 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;54:774-781.
53. 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;76:359-365; discussion 366-358.
54. Stratford PW, Binkley JM. A comparison study of the back pain functional scale and Roland Morris Questionnaire. North American Orthopaedic Rehabilitation Research Network. *J Rheumatol*. 2000;27:1928-1936.
55. Stratford PW, Binkley JM, Riddle DL. Development and initial validation of the back pain functional scale. *Spine*. 2000;25:2095-2102.
56. Wang YC, Hart DL, Cook KF, Mioduski JE. Translating shoulder computerized adaptive testing generated outcome measures into clinical practice. *J Hand Ther*. 2010;23:372-382; quiz 383.
57. Wang YC, Hart DL, Stratford PW, Mioduski JE. Baseline dependency of minimal clinically important improvement. *Phys Ther*. 2011;91:675-688.
58. Wang YC, Hart DL, Stratford PW, Mioduski JE. Clinical interpretation of a lower-extremity functional scale-derived computerized adaptive test. *Phys Ther*. 2009;89:957-968.
59. 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;39:753-764.
60. 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;90:1340-1348.
61. 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.
62. Ware J, Jr., Snow KK, Kosinksi M, Gandek B. *SF-36 Health Survey: Manual and Interpretation Guide*. Boston: The Health Institute, New England Medical Center; 1993.
63. White LJ, Velozo CA. The use of Rasch measurement to improve the Oswestry classification scheme. *Arch Phys Med Rehabil*. 2002;83:822-831.
64. Wright BD, Linacre JM. Reasonable meansquare fit values. *Rasch Meas Trans*. 1994;8:370.
65. Wright BD, Masters GN. *Rating Scale Analyses*. Chicago, IL: MESA Press; 1982.

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.

We have learned that the web based IRT guided CAT survey methodology is most efficient, in terms of accessibility and efficiency. The web based platform is accessible from the patient´s home or office. This accessibility has increased participation levels and completion percentages. This results in reduction of missing data and improved coordination of timely data collection. The patient, the provider and the manager all benefit from these factors. ADD: PATIENT CONFIDENTIALITY / TIME ESTIMATES TO COMPLETE SURVEYS. LIST TRANSLATIONS TO SERVE POPULATIONS.DATA COLLECTION

For patients (i.e., those providing the data), patients respond to, on average, 7 questions from the Low Back FS PROM when administered via computer adaptive testing, and this takes an average of 1 minute, 40 seconds. The PROM survey is followed by 10 questions pertaining to risk adjustment. The typical amount of time needed to complete the entire survey assessment, i.e., the PROM and risk adjustment questions, is 5 minutes.

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

For providers (i.e., those being measured), the amount of time needed to administer the PROM and risk adjustment questions to the patient, complete scoring and risk adjustment calculations, and compile/report data for clinical use is reduced if the FOTO system is used. In that case, 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. Alternatively, if the FOTO system is not used, the data points are entered in the public access version of the measure.

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 ensure that clinic operational processes support strong rates of completed episodes. That is, ensuring that each patient completes an assessment at Intake 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); alternatively, this can be accomplished automatically by discharging the patient in the EHR only, with the needed data points sent automatically from the EHR.

TIMING AND FREQUENCY OF DATA COLLECTION

The assessments are to be completed, at a minimum, at the time of Intake and at least one additional time at or near the time of discharge from the episode of care.

SAMPLING

Sampling is NA. All patients with low back 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.

LANGUAGES

The Low Back PRO-PM is available in English and Spanish.

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)*.

The short form lumbar CAT is publically available at http://www.fotoinc.com/science-of-foto/NQF0425.html. The lumbar specific survey module is contained in a comprehensive outcome data system designed to survey patients with a wide range of orthopedic, neurological and musculoskeletal impairments. A fee of $15 per provider per month provides access to the web based survey platform and the patient specific report. A fee of $25 per provider per month is levied for the survey system and all patient specific and aggregated comparative reports. All measure processes and calculation and reporting is included in these fees.Providers have 3 options for use of the Low Back PRO-PM:

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 a large library of PROMs including the Low Back 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.

Both the public access version, OM Lite, and OM options are feasible for producing measure scores in an efficient manner. The feasibility (affordability) of the costs for OM and OM Lite, which provide further services for meaningful use of outcomes data, 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.

As a further illustration of cost feasibility, a small practice of 4 clinicians would equate to an ongoing cost of $20 per clinician per month for OM Lite and $38 per clinician per month for OM. By using the reporting results to improve and communicate quality of care to referral sources, these may costs become off-set by generating just 1 or 2 new patient referrals for a typical private practice or hospital-based outpatient rehab clinic.

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 https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/physician-compare-initiative/ Payment Program CMS MIPS program as a Clinical Quality Measure https://qpp.cms.gov/about/resource-library Programs with private payers are in use in AZ, LA, MN, WI, and CA, NA AIM Specialty Health w/BCBS/Anthem https://aimspecialtyhealth.com/ CMS MIPS program as a Clinical Quality Measure https://qpp.cms.gov/about/resource-library Programs with private payers are in use in AZ, LA, MN, WI, and CA, NA AIM Specialty Health w/BCBS/Anthem https://aimspecialtyhealth.com/ Regulatory and Accreditation Programs PQRS includes a measure of data collection for lumbar FS Measure 220 that uses the previously endorsed NQF measure . In 2011, 180 providers submitted 7,514 completed episodes, in 2012, 625 providers submitted 15, 488 completed patient episodes and in 2013 https://pqrspro.com/Functional_Deficit_Change_in_Risk-Adjusted_Functional_Status_for_Patients_with_Lumbar_Spine_Impairments 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 has been in CMS payment programs since 2009 including PQRS 2009-2016 and MIPS since 2017. It is a MIPS Clinical Quality Measure, applicable for all patients seen by MIPS Eligible Clinicians, particularly in outpatient settings. CMS provides public reporting results from MIPS measures on Physician Compare. Provider level of measurement, primarily used in outpatient settings.

CURRENT USE: Programs with private payers in Minnesota, Wisconsin, Louisiana, Arizona, and California; provider level use in primarily outpatient settings

The PRO-PM measures in the FOTO system, including the Low Back Functional Status (FS) PRO-PM (NQF measure 0425), 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 Low Back 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 use of FOTO risk adjusted outcome measures of functional status and number of visits, including the Low Back 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 Low Back 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 October 31, 2019
2. https://therapypartners.com/foto-outcomes/. Accessed October 31, 2019.
3. https://cdn2.hubspot.net/hubfs/442011/docs/P4P/TPI%20Statement%20for%20Ways%20and%20Means.pdf?t=1531375320446. Accessed October 31, 2019.

PLANNED USE (Dec 2019): AIM Specialty Health w/Anthem/BCBS; patient-level use in primarily outpatient settings

AIM Specialty Health provides a program for BCBS to encourage use of evidence-based clinical guidelines, including patient-reported outcome measures, for rehabilitation providers and their patients. Their initiative applies to providers who see Anthem Blue Cross Blue Shield (BCBS) beneficiaries in the following states: CT, ME, NH, IN, KY, MO, OH, WI, GA, NY, CA, CO, NV. The FOTO Low Back FS PROM is one of the outcome measures that rehabilitation providers are encouraged to collect and report. Among other clinically relevant data points, a patient's progress as measured by the FOTO Low Back FS PROM will be considered in the context of a provider's requests for authorization of additional therapy visits beginning mid-December of 2019 according to the most recent update from AIM.

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 Low Back 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.

SPECIFIC EXAMPLE OF ORGANIZATION THAT USED THE DATA IN A QI INITIATIVE AND MEASURED RESULTS.
This measure is commonly used for internal quality improvement activities, as described above under CURRENT USE: Quality improvement (internal to the specific organization). Another example comes from a large, multi-state organization of private practice outpatient physical therapy clinics. They applied processes similar to those described above as part of a formal initiative to improve completion rates, functional status improvement (effectiveness), and a measure of amount of improvement relative to the number of therapy visits (utilization). They shared their measured results in the slide which we have pasted in below:
(For the slide, please see the Appendix A.1 for the attachment Addendum to Accountability Transparency. The slide demonstrates improve completion rate from 55 to 68%, effectiveness ranking from 70 to 80%, and utilization rate from 45 to 52%.)4.1.a:Early provider service agreements mandate that FOTO maintain confidentiality of provider participation. This policy has been revised and the confidentiality of current/new providers is not required. FOTO has planned and soon will implement a smart phone application and web site widget that each recognize each of the following provider categories: 1. Participation; 2. Satisfactory episode Completion Threshold; 3. A 3 level percentile ranking for each of Effectiveness, Utilization (Effectiveness and Efficiency and Patient Satisfaction. This program will be voluntary participation, but include a provider release to publish on NQF directed website.
4.1.c.: FOTO has agreements to provide provider outcomes data to three proprietary pilot programs: 1. Therapy Partner Provider Network and Health Partners in Minnesota; 2. Physical Therapy Provider Network of Louisiana and other providers with Blue Cross, Blue Shield of Louisiana; and 3. Therapy Partners with ValueNet of New York.
4.1.d.: NQF Measure 0425 is also the PQRS approved measure #220.
4.1.f:Outcomes Manager by Focus On Therapeutic Outcomes Inc.
The purpose is to provide risk adjusted benchmark efficiency and effectiveness data based on the aggregated data submitted by patients of participating outpatient rehabilitation providers.
The aggregated data was submitted on YYYY patients in the care of XXXX providers from all 50 states. The denominator for the number of comparable entities and patients is unknown.
4.1.g.: The majority of 6,716 providers subscribed to FOTO´s Outcome Manager services and utilizing measure 0425 are utilizing the risk adjusted benchmark comparative reports in quality improvement initiatives.

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?*)
N/A

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.*)

N/A

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.

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 Nov 1, 2019.

HOW MANY AND TYPES OF MEASURED ENTITIES

In the a recent 12-month period ending 9/30/2019 adjusted functional status outcomes (Low Back PRO-PM) data was captured in the FOTO system for 303,243 completed episodes for patients with low back impairments. The patient episodes were incurred by 15,253 clinicians in 4109 clinics. All patients with low back 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 other members of the science team 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 low back 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 additional factors/variables. In particular, smoking and pregnancy are 2 of the most common factors requested. In response to those requests, FOTO plans to add data collection fields for smoking and pregnancy.

Additionally, we interviewed 5 clinicians who use the Low Back PRO-PM in their own clinical care and/or oversee the use of FOTO measures across large, multi-site facilities, gaining feedback from large numbers of clinicians users based on the clinicians' experience as well as patient feedback and experiences. The following is a summary of the responses from each of the 5 clinicians:

- CJ oversees a large, multi-site physical therapy practice organization (states of MN and WI) consisting of 36 clinics and approximately 115 clinicians. CJ is responsible for clinical use of FOTO measures across all sites and gains feedback from large numbers of clinician users and their patients on a regular basis. About the FOTO Low Back PRO-PM, CJ said "In terms of users understanding the questions, there have been no negative comments at all. [It has been] solid."

- BK is the Director of Outcomes for a large private practice physical therapy company with approximately 2000 clinicians across 475-500 clinics across multiple states. As part of his role, BK receives large amount of feedback from clinicians using the FOTO Low Back PRO-PM with their patients. In following with this, BK had one main concern. He said that due to the nature of the IRT/CAT-administration, patients get to the end of the survey without being asked about particular problems that are important to them with respect to their low back problem. Particular questions in the Low Back item bank that BK would like to see asked of all patients are the questions about bending/stooping, lift/carry groceries, changing position, putting on shoes/socks, standing for one hour.

- MY is a practicing physical therapist and uses the FOTO Low Back PRO-PM with her own patients. She also oversees use of FOTO measures for her facilities in the states of VA and WVa. In this role she receives feedback about use of FOTO measures from approximately 100 outpatient clinicians across multiple hospital-based clinics

o Their only concern with the FOTO Low Back PRO-PM has been with respect to the item about "…performing your usual hobbies, recreational or sporting activities?" Sometimes a patient has a hard time answering the question because the patient feels he/she doesn't have any hobbies and/or doesn't do any recreational or sporting activities. While MY understands that the computer adaptive testing functionality accounts for varying patient responses, and that the measure scores are still valid and reliable, the concern is the experience of the patients and clinicians in the clinic. That is, when patients express frustration with the question or difficulty selecting a response, that suggests a negative experience with the survey. MY and her colleagues prefer functional questions that refer to more specific activities or tasks.

o MY said, “Overall [the Low Back PRO-PM] gives a really accurate interpretation of where your patient is, or at least where the patient perceives themselves to be. If you just address the activities list here, you are going to get a better outcome, generally speaking. When you are not seeing the patient achieving what is predicted, you should be looking at the Intake and Status reports and what the patient is answering [for the functional questions].”

o MY went on to say that the scoring results from the Low Back PRO-PM often “cues me to ask the patient [about] things we might not otherwise discuss. It’s like have a computerized detective.” For example, MY illustrated how sometimes the hobbies/recreational/sporting question actually becomes useful in the detective work because sometimes at a Status (progress/interim) point a patient may respond to that question in a manner she did not expect. When this happens, MY asks the patient, “What were you thinking about when you answered that?” She said a patient may then tell her the patient was thinking about his/her “hobby” of watching television, and that the patient has been unable to sit long enough to watch his/her favorite show all the way through. And this might further relate to the patient being unable to sit through their full shift at work. In this manner, MY said, she might discover that sitting tolerance is still a problem for the patient when otherwise the patient has verbally told her he/she is doing well overall. That is, “Patients more often tell us they are progressing well, but the scores tell a different story….It is often that last little piece [of information that is] needed to get the best result.”

o MY also described how this “computerized detective” work may lead to (appropriately) discharging the patient from physical therapy sooner than later, thus promoting more efficient care. “I have statistics telling me what’s a likely outcome for this patient….We may talk sooner about other treatment options,” or start discharge planning sooner.

- GH is a practicing physical therapist seeing outpatients in the state of New Jersey. Patients with neck and back pain comprise a large proportion of his caseload. His comments about the Low Back PRO-PM included:

o “This works better most of the time than a straightforward ODI [Oswestry] or Roland Morris [patient-reported outcome measures specific to low back pain] because the questions cover a whole wide spectrum of what the person could do. Because of the computer adaptive [nature], the questions will changes based on the person….vs the ODI or Roland Morris some are appropriate and some are not….I think the computer adaptive, especially for spinal patients, makes more sens to me than the standard paper version such as ODI or Roland Morris.”

o He also mentioned a limitation in that some patients, such as older adults, “sometimes find it challenging using the ipad,..but in the end they report less confusion and less questioning when I compare the FOTO system to the older paper system we used. Again, I think the adaptive questioning system generates aa more accurate profile of the patient than the old ODI and Roland Morris [that we used to us]. I’ve been a fan of this for quite a long time.”

o GH also discussed the value of using the Low Back PRO-PM in conjunction with other patient-reported measures for assessing psychosocial factors for patients with low back pain. In addition to the FOTO Low Back PROM, he administers the STaRT Back screening tool as well as measures for fear avoidance and self-efficacy.

o GH is passionate about his work with patients with low back pain. He cited the World Health Organization when stating that “low back pain is the #1 cause of disability worldwide, and it needs to be better managed. We need to do the best we possibly can for data management, and I think [the Low Back PRO-PM] does a lot to help with better management.”

- JS oversees the outcomes program for a large private practice organization of approximately 170 outpatient clinics across multiple states with approximately 430 physical therapists

o “I think they are all really good questions. I don’t see anything significant that jumps out at me [as an] inappropriate question.”

o JS said that in general they prefer functional questions that refer to specific activities rather than general activities. Two of the Low Back items, for heavy activities and hobbies/recreational/sporting, sometimes seem to cause confusion for patients and cause therapists to have to spend more time discussing the functional question/patient response with the patient when using the data in clinical care processes. JS said, "The more the questions explain it, the less interaction the therapist has to have with the patient" to explain or discuss a question. Furthermore, JS said his therapists tend to feel more confident about the accuracy of a patient's response to a more specific question.

o For the risk adjustment component, JS said that the collective feedback from his therapists are that they feel it is important to consider smoking and pregnancy as potential risk adjustment factors.

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

The feedback suggests that the Low Back PRO-PM is functioning well, including the 3 new items, with respect to the experience of the providers being measured and their patients. Providers efficiently incorporate the PRO-PM operational processes into their daily workflows. Providers' feedback supports that the risk adjusted PROM data is being used 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. We continue to hear a desire for ongoing risk-adjustment model development with consideration of more variables/constructs, particularly smoking and pregnancy. Functional questions that ask about specific activities are preferred, at least by some.

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. The AIM/BCBS program will soon include the Low Back PRO-PM scores in their utilization decision-making with respect to weighing quality of care/patient improvement relative to authorized visits, and we look forward to gaining their feedback in the future.

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.

Clinician feedback is and will continue to be a primary driver in our decisions to collect and analyze data related to the new risk adjustment model changes as well as ongoing evaluation of item banks. The addition of the 3 new items to the Low Back item bank is one concrete example of clinician feedback as an important consideration for measure improvement. While FOTO has done limited data collection and analyses in the past for smoking, we plan to revamp data collection for smoking and add pregnancy.

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.

As described in the Testing Form in section 2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE, providers (clinicians and clinics) demonstrated significant improvement in their performance over time, based on monotonic and significant increases in mean residual scores over a 3 year period when examining only the same providers who contributed data across the 3 year

timeframe (n = 346,930 patients, n=2,552 clinicians, n = 1,182 clinics. Geographic area was all 50 states plus District of Columbia).

For clinics, mean residuals were different between years (one-way ANOVA ($F(2,3543) = 53.6$, $p < 0.001$) with a monotonic and significant increased from -0.3 to +1.2 over time. For clinicians, mean residuals were different between years (one-way ANOVA ($F(2,7653) = 58.2$, $p < 0.001$) with a monotonic and significant increased from -0.3 to +1.0 over time.

These findings support that providers may better learn and gain skills over time for using risk-adjustment PROM data in the context of everyday data-driven clinical decision making with the patient at the center. Providers may improve their ability to use the data to enhance their communication with the patient, promote patient engagement. Using risk-adjusted patient-reported outcome measures (PROMs) of function promotes a focus on patient-perceived function and encourages meaningful discussions about goals and expectations for the results of the care episode.

The performance results of the Low Back PRO-PM (NQF measure 0425) will be further evaluated by CMS after data collection/submission is completed for the 2019 MIPS performance year. As of Nov 6, 2019, 1531 clinicians across 72 organizations participating in the 2019 MIPS performance year were registered as reporting MIPS measures through the FOTO QCDR. Given these strong provider sample sizes and the high prevalence of patients presenting for rehabilitation care for low back impairments, we anticipate CMS will be able to determine meaningful performance results and set the stage for CMS to be able to reward improvement over time, which is the basis for one of the MIPS bonus criteria.

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 continually gather data for the sake of future risk adjustment analyses. One such data point has been for "Hepatitis, HIV, or AIDS." We recently learned that a patient was denied his/her application for a military position due to the presence of these words in the medical record. While the problem with the application was eventually corrected, we are planning to either modify or remove this item.

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

We are unaware of any unexpected benefits.

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?

No

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

N/A

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.)

The two other measures that surface in a NQF/QPS search are #0514: MRI of Lumbar Spine, and #0739: Radiation dose for Lumbar Xray. Neither measure change in function of lumbar spine or grades clinician and/or facility quality based on risk-adjusted functional status change for patients with low back pain.N/A

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.

Attachment **Attachment:** Addendum_to_Accountability_Transparency_section_in_Use_and_Usability.docx

Contact Information

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

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

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

Co.4 Point of Contact: Deanna, Hayes, deanna.hayes@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 was the original developer of this measure in 2008. Dr. Hart died in 2012. The contributors to continued development, analysis, maintenance and re-submission since that time have included Karon F. Cook, PhD; Daniel Deutscher, PT, PhD; Julie Fritz, PT, PhD, ATC; Linda Resnik, PT, PhD; Ying-Chih "Inga" Wang, OTR/L, PhD; Mark Werneke, PT, MS, Dip MDT; Michael Kallen, PhD; and Deanna Hayes, PT, DPT, MS.Dennis Hart, PhD was the original developer of this measure in 2008. Dr. Hart died in 2012. The expert panel assembled for continued development, analysis, maintenance and re-submission, include: Karon F. Cook, PhD, *Daniel Deutscher, PT, PhD, Julie Fritz, PT, PhD, ATC, *Linda Resnik,PT, PhD,Ying-Chih "Inga" Wang, OTR/L, PhD. *Mark Werneke, PT, DPT, MS, SCS, OCS, CSCS. Those names preceded by an asterisk are specifically involved in the preparation and submission for endorsement renewal.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2008

Ad.3 Month and Year of most recent revision: 07, 2012

Ad.4 What is your frequency for review/update of this measure? as required

Ad.5 When is the next scheduled review/update for this measure? 12, 2019

Ad.6 Copyright statement: See NQF document

Ad.7 Disclaimers:

Ad.8 Additional Information/Comments: