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

Purple and Blue text represents the responses from measure developers

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

Brief Measure Information

NQF #: 3532

Corresponding Measures:

De.2. Measure Title: Discouraging the routine use of occupational and/or supervised physical therapy after carpal tunnel release.

Co.1.1. Measure Steward: American Academy of Orthopaedic Surgeons

De.3. Brief Description of Measure: Percentage of patients 18+ with carpal tunnel syndrome who received surgical carpal tunnel release, and who should not routinely be prescribed postoperative physical and/or occupational therapy within 6 weeks after release.

1b.1. Developer Rationale: This measure should discourage (and thus decrease) the routine use of physical therapy or occupational therapy, after carpal tunnel syndrome release procedures, as such post-procedural therapy does not improve outcomes.

S.4. Numerator Statement: Number of patients with carpal tunnel syndrome, who underwent carpal tunnel release, and who did not receive postoperative hand, physical therapy (low, moderate, or high complexity) and/or occupational therapy (low, moderate, or high complexity) within 6 weeks (42 days) of the carpal tunnel release.

S.6. Denominator Statement: Patients 18 years or older, with a diagnosis of carpal tunnel syndrome, undergoing carpal tunnel syndrome release.

S.8. Denominator Exclusions: N/A

De.1. Measure Type: Process

S.17. Data Source: Claims

S.20. Level of Analysis: Clinician : Individual, Facility

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

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

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

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

Preliminary Analysis: New Measure

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *structure, process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary

- This is a new claims-based process appropriate use measure at the individual clinician and facility level which assesses the percentage of patients 18+ with carpal tunnel syndrome who received surgical carpal tunnel release, and who should not routinely be prescribed postoperative physical and/or occupational therapy within 6 weeks after release.
- The developer cites the American Academy of Orthopaedic Surgeons 2016 Management of Carpal Tunnel Syndrome Evidence-Based Clinical Practice Guideline as evidence.
    - The Guideline states that moderate evidence supports that there is no additional benefit to routine supervised therapy over home programs in the immediate postoperative period.
    - Developer notes one high quality study and two moderate quality studies supporting the recommendation.

Exception to evidence

- Not Applicable

Questions for the Committee:

- What is the relationship of this measure to patient outcomes?
- How strong is the evidence for this relationship?
- Is the evidence directly applicable to the process of care being measured?

Guidance from the Evidence Algorithm

Process measure based on systematic review (Box 3)→ QQC present (Box 4)→ Quantity: moderate; Quality: moderate; Consistency: moderate (Box 5b)→Moderate

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

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

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

- The developer provides data from 3 years (2016-2018). The data is listed below:

  Time: Numerator; Denominator; Performance; Facility-level Range

  2016: 7187; 7530; 95.4%; 65.4 – 100.0

  2017: 6317; 7070; 89.3%; 14.7 – 100.0

  2018: 5495; 6213; 88.4%; 8.0 – 100.0

Disparities

- The developer did not provide information on potential disparities.

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?
- If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?

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

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

1a. Evidence to Support Measure Focus: 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.

- Overuse measure. Is there benefit to reduction in use?
- This measure the number of patients undergoing a carpal tunnel release procedure who did not have PT or OT. It is based on guidelines with moderate evidence showing no additional benefit to therapy; developer notes one high quality study and two moderate quality studies supporting the recommendation.
- No new data since 2016. Evidence from systematic review
- Evidence supports the measure.

- Evidence cited indicates that routine PT/OT after CTS release surgery does not provide additional benefit compared to home exercise program alone. However, this measure's specifications do not adequately reflect the evidence cited.
- This is an initial process measure. The evidence review was done for the Management of Carpal Tunnel Syndrome Evidence-Based Clinical Practice Guideline by the American Academy of Orthopedic Surgeons (2016).Two studies were found that addressed supervised therapy vs. home therapy following carpal tunnel syndrome (CTS) surgery. These were not well controlled. Results in one study were hand strength measurements. Results in the other study were measured by a CTS-specific patient reported instrument.
- Evidence moderately supports the measure.
- Evidence is strong and is supported by the American Academy of Orthopedic Surgeons 2016 Management of Carpal Tunnel Syndrome Evidence-Based Clinical Practice Guidelines. There is a strong relationship to patient outcomes. The evidence is directly applicable to the process of care being measured.
- Process Measure, evidence rating is moderate

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?

- Rates moving in wrong direction. Is current rate of 90%ish going to see significant improvement?
- The numerator in this measure is defined by the number that did NOT receive PT/OT. Therefore, higher numerical performance is desirable. The developer has presented performance data showing 95.4% (2016), 89.3% (3017) and 88.4% (2018). Although mean performance is relatively high, it's not clear why performance is falling. The developer has provided a range of performance but it's not clear what the distribution is.; There is no disparity data provided.
- There is worsening performance in this measure over the previous 3 years of data. In 2018, nearly 12% of individuals meeting inclusion had PT/OT/
- Performance data show worsening over time in this process.?
- Yes, gap exists, but is seemingly getting worse
- Performance data is from inpatient and outpatient claims in the Veterans Health Administration (VA) Corporate Data Warehouse (FY; 2016-18). The denominator is patients 18 years or older with a diagnosis of carpal tunnel syndrome undergoing carpal tunnel syndrome release. The numerator is those whose records show no patient encounter for postoperative hand physical therapy or occupational therapy within 42 days of CTS release. The percentages in the three years studied were 95.4%, 89.3%, and 88.4%. The data was not analyzed for disparities.
- Issue with using the historical data as tabulated. Inverse may be a better measure, that is those who do not receive supervised physical therapy.
- No disparities were identified by the measure developer. Not as strong of an argument for a gap in care to address.
- A small performance gap exists, however the data presented suggests that there is a worsening performance against this measure recently

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

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

Reliability

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

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

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

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

Composite measures only:

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

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

Evaluators: Staff

Staff Review

Evaluation Summary:

Reliability

- Reliability testing conducted at the measure score level using a signal-to-noise analysis.
- Data was collected from the Veterans Health Administration (VA) Corporate Data Warehouse for three fiscal years (FY; 2016-18). The sample included 111 VA facilities and 1873 surgeons across 3-years (with a subset of 610 surgeons who performed at least 5 carpal tunnel releases).
- The developer states that the results of .99 at the clinician level and .95 at the facility level suggest that the reliability of the proposed quality measure is 'highly reliable'.

Validity

- Validity testing conducted at the measure score level using face validity.
    - Using a modified Delphi voting process, a developer convened 15-member workgroup rated 13 candidate measure on a 9-point scale on a scale ranging from 1 (definitely not valid) to 9 (definitely valid) for face validity on each of the four NQF criteria.

- All 4 NQF domains received a median rating of 8. The developer states that rating indicates that the measure is valid.

Questions for the Committee regarding reliability:

- Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?

Questions for the Committee regarding validity:

- Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?

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

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

2a1. Reliability-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?

- na
- No concerns about reliability specifications.
- No concerns
- Data elements are clear. No concerns.
- No major concerns
- ICD-10 codes for CTS diagnosis and CPT codes for CTS surgical release and for levels of physical therapy and occupational therapy sessions are provided. Cases in which the appropriate diagnostic and surgery codes but no PT/OT CPT codes in the specified postop period were pulled to double check they belong in the numerator. There was no sampling.
- Appears to be well specified for reliability.
- No concerns with the reliability of the measure.
- No concerns

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

- measure score level using a signal-to-noise analysis
- no
- No
- No concerns.
- No major concerns
- Signal to noise ratio reliability testing was performed at the clinician and facility levels and found to be highly reliable.
- The data support high reliability.

- Measure only tested with the VA Administration, so there could be concerns identified if a broader testing were to occur.
- No concerns

2b1. Validity -Testing: Do you have any concerns with the testing results?

- measure score level using face validity
- used a modified Delphi process
- No
- Face validity testing was performed by a workgroup.
- No major concerns
- Face validity was studied using a multidisciplinary clinician work group. No construct validity testing was submitted.
- No concerns about the validity, but could be clearer about home care such as stretching exercises.
- Validity testing results were not clear.
- No concerns

2b2-3. 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?

- na
- no risk adjustment performed
- No concerns for face validity
- No exclusions.
- I feel that for an overuse measure, exclusions must be listed. Otherwise, this may have unintended consequences of curbing appropriate use.
- There were no exclusions or risk adjustments/stratification.
- It would be helpful to hear from the developer on the individuals who would benefit from supervised physical therapy and any indicators to possibility excluded them.
- No concerns with exclusion criteria.
- No concerns

2b4-7. 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?

- na

- no
- No threats to validity noted.
- No obvious threats to validity.
- unclear how you are able to capture "home exercise program prescription"
- Differences among the three study years in the percentage of postop VA patients who did not have PT/OT therapy in the CTS release acute recovery period were small (2016 - 95.4%, 2017 - 89.3%, 2018 - 88.4%). The "compliance rate" declined over this three year period. There is only one set of specifications. There was no missing data analysis.
- At present no specific concerns.
- Data was provided, but would like to see data from across populations or programs to determine validity. No identified threats to validity from the information provided.
- No concerns

Criterion 3. Feasibility

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.

- Coded by someone other than person obtaining original information
- ALL data elements are in defined fields in electronic claims

Questions for the Committee:

- Are the required data elements routinely generated and used during care delivery?
- Are the required data elements available in electronic form, e.g., EHR or other electronic sources?
- Is the data collection strategy ready to be put into operational use?

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?

- data elements are in defined fields in electronic claims
- claims based measure.
- Elements are electronically available/
- Data elements are routinely generated during care delivery. No concerns
- No major concerns
- All data elements are in defined fields in electronic claims.
- Seems very feasible.

- All data elements are routinely generated during care provision. All data elements are administrative/electronic. No concerns with feasibility.
- Process measure, data elements routinely generated during routine care delivery

Criterion 4: Usability and Use

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.

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

Accountability program details

- The developer plans to submit this measure to the Centers for Medicare and Medicaid services for consideration for inclusion in Merit-Based Incentive Payment System.

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

- This measure´s technical report was provided for public comment for a 30-day period.
- A total of 99 respondents provided feedback for this measure, with 94% generally supportive of the measure as written.

***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

- Measure has not been implemented and therefore does not have year-over-year performance data for review.

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

- During the developer's public commenting period, there were multiple respondents that highlighted the need, at times, for therapy for certain patients with stiff proximal interphalangeal joints, generalized arthritis of the digits, or preoperative stiffness. While the developer convened workgroup agreed with this statement, they did not believe building in exclusion criteria would be beneficial as the incidence of these comorbidities requiring formal therapy is uncommon when patients are appropriately educated and counseled on a home program.

Potential harms: None identified

***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

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?

- plans to submit this measure to cms
- Plan to submit for MIPS inclusion; underwent a 30 day public feedback period. 94% were generally supportive.
- Not been implemented yet. Feedback and public comments have been solicited, with 94% supportive.
- Not yet in use. Plan to submit for inclusion in medicare/medicaid merit based incentive payment system
- I feel that for an overuse measure, exclusions must be listed. Otherwise, this may have unintended consequences of curbing appropriate use.

- The measure is not currently being publicly reported. The measure is to be submitted to the Centers for Medicare and Medicaid Services for consideration for inclusion in the Merit-Based Incentive Payment System. This measure´s technical report was provided for public comment for a 30-day period with comments received from 99 respondents. Part of that feedback from multiple respondents highlighted the need, at times, for therapy for certain patients with stiff proximal interphalangeal joints, generalized arthritis of the digits, or preoperative stiffness. The Workgroup was in complete agreement with these concerns and discussed these specific examples during the in person meeting. The Workgroup acknowledged that these comorbidities are uncommon and implementation of the measure would drive utilization towards 0% but it would never reach 0%. The literature does not support the routine use of formal therapy but implementation of the measure will establish a national benchmark of utilization and will identify outliers that routinely use therapy 100% of the time, as we found during the validation study. The Workgroup did not believe building in exclusion criteria would be beneficial as the incidence of these comorbidities requiring formal therapy is uncommon when patients are appropriately educated and counseled on a home program.
- Meets the requirements for Use.
- Not currently in use in a accountability program. Measure developer plans to submit to CMS for use in the Merit-Based Incentive Program.
- Proposed use in MIPS program

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.

- na
- no obvious harms.
- No significant unintended consequences except comments that individuals with severe osteoarthritis in the digits may not achieve best outcomes with PT/OT. Utility of measure lies in healthcare cost reduction in face of interventions that show no benefits.
- Not in use. No harms.
- I feel that for an overuse measure, exclusions must be listed. Otherwise, this may have unintended consequences of curbing appropriate use.
- The developers believe that their analysis demonstrates that, although there is relatively low variation in care,, there is still meaningful variation in postop CTS surgery care and room for improvement in further limiting postop PT and OT treatments. Unintended negative consequences were not discovered in the development of this measure.
- Appears to meet the requirements for Usability.
- Measure has not yet been implemented. Public comments stated that there are many patients that benefit from PT or OT after treatment. Which indicates that the implementation of the measure may result in some patients that would benefit from OT/PT may not receive it in order for improved score. Results could be used to improve high-quality, efficient care.
- No concerns

Criterion 5: Related and Competing Measures

Related or competing measures
None
Harmonization
N/A

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

5. 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?

- na
- none
- None
- No related or competing measures
- None identified
- There are no related and competing measures.
- Doesn't apply as New Measure with no related or competing measures.
- No competing measures were identified.
- NA

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 01/21/2021

- No NQF Members have submitted support/non-support choices as of this date.
- No Public or NQF Member comments submitted as of this date.

Staff Scientific Acceptability Evaluation

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 3532

Measure Title: Discouraging the routine use of occupational and/or supervised physical therapy after carpal tunnel release.

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**

Level of Analysis:

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

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.**

   No concerns

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**

   Submission document: Testing attachment, section 2a2.2

   - Data was collected from the Veterans Health Administration (VA) Corporate Data Warehouse for three fiscal years (FY; 2016-18). The sample included 111 VA facilities and 1873 surgeons across 3-years (with a subset of 610 surgeons who performed at least 5 carpal tunnel releases).
   - The developer conducted signal-to-noise ratio reliability (SNR) testing.
   - The signal, in SNR testing, and within the context of performance measurement, represents the 'true' differences in measure scores among providers, that is, real differences in the quality of care that is delivered. The noise is the measurement error, or the amount of imprecision of the instrument – here, the PM. Contained within the measure score (quality

action) is a 'signal' and 'noise'. Ideally, you want the most signal and the least noise, producing a highly reliable score.

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

   Submission document: Testing attachment, section 2a2.3

   - 132,049 VA patients who were diagnosed with CTS in FY16-18 in the dataset used. Of those, 20,813 VA patients received carpal tunnel release and were eligible for the denominator.

Measures	Overall SNR Score	N	Min value	25th percentile	50th percentile	75th percentile	Max value
Clinician Level SNR Analysis	0.99	1185	0.08	1.00	1.00	1.00	1.00
Facility Level SNR Analysis	0.95	111	0.30	0.91	0.95	0.97	1.00

   - The developer states that the results suggest that the reliability of the proposed quality measure is 'highly reliable'.

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**

   ☐ **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 <u>all</u> 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 <u>not</u> 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.**

    The SNR result suggests a high rate of reliability.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    Submission document: Testing attachment, section 2b2.

    None- No exclusions

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

    Submission document: Testing attachment, section 2b4.

    N/A

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.
    N/A

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

    Submission document: Testing attachment, section 2b6.

    N/A

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**

For cost/resource use measures ONLY:

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

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

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

VALIDITY: TESTING

19. **Validity testing level: ☒ Measure score ☐ Data element ☐ Both**
20. **Method of establishing validity of the measure score:**

    ☒ Face validity

    ☐ Empirical validity testing of the measure score

    ☐ N/A (score-level testing not conducted)
21. **Assess the method(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.2

    Using a modified Delphi voting process, a developer convened 15 member workgroup rated 13 candidate measure on a 9-point scale on a scale ranging from 1 (definitely not valid) to 9 (definitely valid) (Figure 1) for face-validity on each of the four NQF criteria (Table 3).
22. **Assess the results(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.3

    All 4 NQF domains received a median rating of 8. The developer states that rating indicates that the measure is valid.
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.**

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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

    ☐ **High**

    ☐ **Moderate**

    ☐ **Low**

    ☐ **Insufficient**

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

ADDITIONAL RECOMMENDATIONS

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

Developer Submission

Additional evaluations and submission materials attachments…

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_7.1-PY_DONE-636996657653061203.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.

No

1a. Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: Discouraging the routine use of occupational and/or physical therapy after carpal tunnel release.

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

Date of Submission: 7/1/2019

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

Outcome

☐ Outcome:

☐ Patient-reported outcome (PRO):

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: Physical therapy and/or occupational therapy use, after carpal tunnel syndrome release procedures.

☐ 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.

The American Academy of Orthopaedic Surgeons published a moderate strength clinical practice guideline recommendation finding no additional benefit to routine supervised therapy over home programs in the immediate postoperative period after carpal tunnel syndrome surgical release. This is a process measure discouraging the use of supervised physical therapy or occupational therapy after carpal tunnel syndrome release procedures, in the 6-week post-surgical period.

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.)

N/A; data is not derived from patient report.

****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.

N/A; data is not derived from patient report.

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)

X 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	• Management of Carpal Tunnel Syndrome Evidence-Based Clinical Practice Guideline – Recommendation: Moderate evidence supports no additional benefit to routine supervised therapy over home programs in the immediate postoperative period. No evidence meeting the inclusion criteria was found comparing the potential benefit of exercise versus no exercise after surgery. • The American Academy of Orthopaedic Surgeons. • February 29, 2016. • American Academy of Orthopaedic Surgeons. Management of Carpal Tunnel Syndrome Evidence-Based Clinical Practice Guideline. (P. 684-735) www.aaos.org/ctsguideline. Published February 29, 2016. • www.aaos.org/ctsguideline.

Systematic Review	Evidence
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	Moderate evidence supports no additional benefit to routine supervised therapy over home programs in the immediate postoperative period. No evidence meeting the inclusion criteria was found comparing the potential benefit of exercise versus no exercise after surgery. Rationale: Routine post-operative therapy after carpal tunnel release was examined in 6 high quality studies. From these, two studies (Hochberg 2001 and Jerosch-Herold 2012) addressed interventions not relevant to current core practices of postoperative rehabilitation. The remaining four studies (Alves 2011, Fagan 2004, Pomerance 2007, and Provinciali 2000) addressed the need for supervised therapy in addition to a home program in the early postoperative period, the early use of laser, or the role of sensory reeducation in the later stages of recovery. One high quality study (Alves 2011) evaluated the use of laser administered to the carpal tunnel in 10 daily consecutive sessions at a 3J dosage and found no difference in pain/symptom reoccurrence in comparison to placebo. Two moderate quality studies (Pomerance 2007 and Provinciali 2000) compared in-clinic or therapist supervised exercise programs in addition to a home program to a home program alone. The studies were somewhat limited by an incomplete description of who delivered home programs, exercise/education content and dosage, and treatment progression. Pomerance (2007) compared a two week program directed by a therapist combined with a home program alone and found no additional benefit in terms of grip or pinch strength in comparison to the home program alone. Provinciali (2000) compared one hour sessions over 10 consecutive days of in-clinic physiotherapy comprising a multimodal program with a home program that was progressed in terms of strength/endurance. No benefit was found in outcome when measured by a CTS-specific patient reported instrument.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	High Quality Evidence and Moderate Quality Evidence were both included in this recommendation.
Grade assigned to the **recommendation** with definition of the grade	**Moderate:** Evidence from two or more “Moderate” quality studies with consistent findings, or evidence from a single “High” quality study for recommending for or against the intervention.

Systematic Review	Evidence
Provide all other grades and definitions from the recommendation grading system	**Strong:** Evidence from two or more "High" quality studies with consistent findings for recommending for or against the intervention. **Moderate:** Evidence from two or more "Moderate" quality studies with consistent findings, or evidence from a single "High" quality study for recommending for or against the intervention. **Limited:** Evidence from two or more "Low" quality studies with consistent findings or evidence from a single "Moderate" quality study recommending for against the intervention or diagnostic or the evidence is insufficient or conflicting and does not allow a recommendation for or against the intervention.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	1 high quality study and 2 moderate quality studies.
Estimates of benefit and consistency across studies	*
What harms were identified?	There are no known harms associated with implementing these recommendations.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	No new studies have changed the conclusions from the SR.

*cell intentionally left blank

Provide all other grades and definitions from the evidence grading system

High Quality

Moderate Quality

Low Quality

Very Low Quality

These ratings are determined by how many study design flaws are identified during quality evaluation review.

Randomized Study Design Quality Key:

High Quality Study	<2 flaws
Moderate Quality Study	≥2 and <4 flaws
Low Quality Study	≥4 and <6 flaws
Very Low Quality Study	≥6 flaws

Observational Study Design Quality Key:

High Quality Study	<2 flaws
Moderate Quality Study	≥2 and <4 flaws
Low Quality Study	≥4 and <6 flaws
Very Low Quality Study	≥6 flaws

To see which domains are addressed in these ratings, please visit the AAOS Guideline and Systematic Review Process website.

_____

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. N/A. Measure based off Clinical Practice Guidelines.

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

N/A. Measure based off Clinical Practice Guidelines.

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

N/A. Measure based off Clinical Practice Guidelines.

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.

This measure should discourage (and thus decrease) the routine use of physical therapy or occupational therapy, after carpal tunnel syndrome release procedures, as such post-procedural therapy does not improve outcomes.

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.*

System	Time	Denominator	Numerator	Performance	Facility-level Range
VA	2016	7530	7187	95.4%	65.4 – 100.0
VA	2017	7070	6317	89.3%	14.7 – 100.0
VA	2018	6213	5495	88.4%	8.0 – 100.0
VA	FY16-18	20813	19455	91.3%	32.9 – 100.0

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.*

N/A

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

See above.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

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

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

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

https://www.aaos.org/contentassets/85f5a78e1b054c27ac679adeb3459e81/cts-measures-technical-report-2019-update.pdf

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**:** Data_Dictionary.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.

No, this is not an instrument-based measure **Attachment:**

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.

Not an instrument-based measure

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.

No

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.

N/A

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).

Number of patients with carpal tunnel syndrome, who underwent carpal tunnel release, and who did not receive postoperative hand, physical therapy (low, moderate, or high complexity) and/or occupational therapy (low, moderate, or high complexity) within 6 weeks (42 days) of the carpal tunnel release.

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 encounter for Carpal Tunnel Release (CPT): 64721 or 29848

AND

Diagnosis of Carpal Tunnel Syndrome (ICD-10-CM): G560, G5600, G5601, G5602, G5603

AND

No Patient encounter for postoperative hand, physical therapy (low, moderate, or high complexity) within 6 weeks (42 days) of carpal tunnel release (CPT): 97161, 97162, 97163

OR

No patient encounter for postoperative hand occupational therapy (low, moderate, or high complexity) within 6 weeks (42 days) of carpal tunnel release (CPT): 97165, 97166, 97167.

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

Patients 18 years or older, with a diagnosis of carpal tunnel syndrome, undergoing carpal tunnel syndrome release.

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).

Patient encounter for Carpal Tunnel Release (CPT): 64721 or 29848

AND

Diagnosis of Carpal Tunnel Syndrome (ICD-10-CM): G560, G5600, G5601, G5602, G5603.

Denominator cases must have (1) a CTS diagnosis, and (2) a CTS-R code. The measurement period is 1-year. This is a claims-based measure, and a process/appropriate use measure. Denominator cases that did not undergo supervised physical therapy or occupational therapy (defined by PT/OT evaluation codes), in the 42-day (or 6-week) post-procedural window, will be numerator patients. This is a patient-based, provider-level measure.

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

N/A

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.)*

N/A

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.)*

N/A

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

No risk adjustment or risk stratification

If other:

S.12. Type of score:

Ratio

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 = Lower 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.)*

1) Identify cases with a carpal tunnel syndrome diagnosis code (ICD-10-CM: G560, G5600, G5601, G5602, G5603).
2) Identify those from above with an associated carpal tunnel syndrome release procedural CPT code: 64721 or 29848.
3) Ensure cases pulled are within the age range of > 17, are labeled as denominator patients, did not leave AMA, were not discharged dead, and were not discharged to hospice. Label the date of the CTS-R procedure, so we can identify cases in the post-procedural window.
4) Specify the 42-day post-procedure window. Ensure CTS-R dates are prior to PT/OT dates.
5) Pull those denominator cases that did not have a PT/OT code in the 42-day post-procedure window. Ensure cases did not have a PT/OT CPT code: 97161, 97162, 97163, 97165, 97166, 97167.
6) Label cases as numerator patients.

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.

N/A.

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.

N/A.

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

If other, please describe in S.18.

Claims

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.

N/A

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 : Individual, Facility

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

Inpatient/Hospital, 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_updated_AAOS_Nov_2020.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.

No

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.

No - This measure is not risk-adjusted

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

Measure Number (*if previously endorsed*)**:** 3532
Measure Title: Discouraging the routine use of supervised physical therapy and/or occupational therapy after carpal tunnel release.
Date of Submission: August, 2020
Type of Measure:

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

*cell intentionally left blank

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:	☐ other:

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*).

Main Dataset

Veterans Health Administration (VA) Corporate Data Warehouse for three fiscal years (FY; 2016-18)

1.3. What are the dates of the data used in testing? (FY; 2016-18)

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:	☐ other:

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*)

The 'facility-level range' pertains to the distribution of performance across 111 VA facilities.

The physician level analysis included 1873 surgeons across 3-years, and the subset of 610 surgeons who performed at least 5 carpal tunnel releases.

a. **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*)

Patient Case Mix

132,049 VA patients who were diagnosed with CTS in FY16-18 in the dataset used. Of those, 20,813 VA patients received carpal tunnel release and were eligible for the denominator.

	Overall
n	20813
age (mean (SD))	60.82 (13.43)
gender = Male (%)	18814 (90.4)
race (%)	
Black or African American	2626 (12.6)

Hispanic White or other minority 1375 (6.6)
Non-Hispanic White 15725 (75.6)
Unknown 1087 (5.2)
Marital Status (%)
Single 711 (3.4)
Married 12254 (59.2)
Separated/divorced 5087 (24.6)
Widow/widower 818 (4.0)
Never married 1819 (8.8)

b. **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**.

Veterans Health Administration (VA) Corporate Data Warehouse (FY; 2016-18) was used for all aspects of testing

c. **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.

Differences in race, age, place of primary residence, and insurance status were not analyzed

---

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*)

Signal-to-noise ratio reliability (SNR) is one type of reliability testing and is considered the generally accepted standard method for describing the repeatability or consistency of healthcare performance measures (PMs). The typical process includes fitting regression model(s) to the data to derive estimated population parameters that are either used in the calculation of reliability, or are used to derive variances, which are inputs for the calculation of reliability. Gold standard models for the computation of reliability employ Bayesian estimation and mixed-effects modeling. Bayesian hierarchical modeling should be favored over other types of regression, when possible, as these models are superior for estimating, smoothing, and explaining variation.

The signal, in SNR testing, and within the context of performance measurement, represents the 'true' differences in measure scores among providers, that is, real differences in the quality of care

that is delivered. The noise is the measurement error, or the amount of imprecision of the instrument – here, the PM. Contained within the measure score (quality action) is a 'signal' and 'noise'. Ideally, we want the most signal and the least noise, producing a highly reliable score. The formulas used in these analyses are shown below:

Provider-to-provider variance[1]

$$\sigma^2_{provider-to-provider} = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$$

Statistical error[1]

$$\sigma^2_{error} = \frac{\hat{p}(1-\hat{p})}{n}$$

Reliability[1]

$$reliability = \frac{\sigma^2_{provider-to-provider}}{\sigma^2_{provider-to-provider} + \sigma^2_{binomial}} = \frac{\sigma^2_{provider-to-provider}}{\sigma^2_{provider-to-provider} + \frac{p(1-p)}{n}}$$

To summarize the process, the basic method for SNR calculation used here was to (1) fit a model to the data, (2) use derived populations parameters to compute variance, and (3) use those variances to calculate reliability.

These methods are the same methods commonly used by other Measure Developers, and proposed by the Rand Corporation's John Adam's Paper, and the BETABIN MACRO.

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*)

Measures	Overall SNR Score	N	Min value	25th percentile	50th percentile	75th percentile	Max value
Clinician Level SNR Analysis	0.99	1185	0.08	1.00	1.00	1.00	1.00
Facility Level SNR Analysis	0.95	111	0.30	0.91	0.95	0.97	1.00

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?*)

The results suggest that the reliability of the proposed quality measure is 'highly reliable'.

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)*

Validity testing focuses on systematic errors and bias. It involves testing agreement between the data elements obtained when implementing the measure as specified and data from another source of known accuracy. Validity of computed measure scores involves testing hypotheses of relationships between the computed measure scores as specified and other known measures of quality or conceptually related aspects of quality. A variety of approaches can provide some evidence for validity. The specific terms and definitions used for validity may vary by discipline, including face, content, construct, criterion, concurrent, predictive, convergent, or discriminant validity.

The approach utilized in determining the validity of these measures was face validity, as determined by a multidisciplinary clinician work group. The systematic and transparent methods address the question of whether scores obtained from the measures will provide an accurate reflection of quality and can therefore be used to distinguish good and poor quality.

The workgroup evaluated the validity of the original 13 preliminary candidate measures to address their adherence to the National Quality Forum (NQF) criteria. Using a modified Delphi voting process, workgroup members rated each candidate measure on a 9-point scale on a scale ranging from 1 (definitely not valid) to 9 (definitely valid) (Figure 1) for face-validity on each of the four NQF criteria (Table 3).

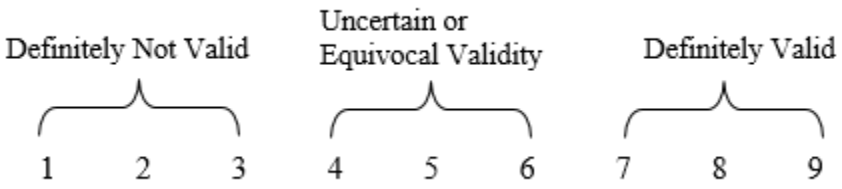

Figure 1. 9-point validity scale

Workgroup members were instructed that a candidate measure would be considered valid and be considered for gap and reliability testing only if the final median rating on all four domains was in the upper tertile for validity (median = 7-9) with agreement amongst the work group (fewer than 4 of the 15 voter ratings were outside the 3-point median range, see Table 2).

Table 1. Agreement Requirements

Panel Size	Disagreement # of ratings between 1-3 or 7-9	Agreement # of ratings outside of appropriateness rating range
8,9,10	≥ 3	≤ 2
11,12,13	≥ 4	≤ 3
14,15,16	≥ 5	≤ 4

Panels beyond these sizes require calculation of IPRAS (Interpercentile Range Adjusted for Symmetry) described in RAM.

The validity rating took place over two rounds. Before the meeting, each panelist used the aforementioned 9-point validity scale to individually rate the candidate measures based on the following criteria: 1) important to measure 2) scientifically acceptable 3) feasible and 4) useable. For the first round of voting, workgroup members were provided with the list of candidate quality measures, measure specifications, supporting literature, and definitions of each of the domains that were being used to determine validity (Table 3).

Table 2. NQF Criteria

Important to measure and report to keep our focus on priority areas, where the evidence is highest that measurement can have a positive impact on healthcare quality.

Scientifically acceptable, so that the measure when implemented will produce consistent (reliable) and credible (valid) results about the quality of care.

Feasible to collect with data that can be readily available for measurement and retrievable without undue burden.

Useable and relevant to ensure that intended users — consumers, purchasers, providers, and policy makers — can understand the results of the measure and are likely to find them useful for quality improvement and decision making.

The first round votes were compiled and presented as a basis for discussion at the in-person meeting. At the in-person meeting, votes were kept anonymous, and workgroup members discussed every preliminary candidate measure at length, specifically addressing how each measure may or may not fulfill the domains “importance for clinical care” “scientific acceptability” “feasibility” and “usability”. As an extension of this discussion, the workgroup was invited to revise the preliminary candidate measures as necessary to increase overall validity across the board.

After discussion and revision, workgroup members were again provided with the measure specifications, supporting literature, and evidence, and asked to vote a second time on the face validity of each revised measure. The candidate measures which received a valid rating with agreement on all four criteria were considered to have met the requirement for face validity and moved on for clinical data testing.

At the conclusion of discussion, revision, and re-voting, the workgroup members voted that a total of 6 measures were considered valid with agreement on all four domains. Three of these measures proved to be possible to test with the dataset used.

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

The results of the validity testing are shown below:

Measure	Voter #1	Voter #2	Voter #3	Voter #4	Voter #5	Voter #6	Voter #7	Voter #8	Voter #9	Voter #10	Voter #11	Voter #12	Voter #13	Voter #14	Voter #15
Importance for clinical care	9	9	7	8	7	7	9	8	9	7	9	8	8	7	9
Scientific Acceptability	9	9	7	8	9	8	9	8	9	8	9	8	7	7	8
Feasibility	9	9	7	8	9	8	9	8	9	8	9	8	7	8	9
Usability	9	9	7	8	9	8	9	8	9	9	9	8	7	8	8

Measure	Range	Median	Agreement Rating
Importance for clinical care	Valid	8	Agreement
Scientific Acceptability	Valid	8	Agreement
Feasibility	Valid	8	Agreement
Usability	Valid	8	Agreement

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?*)

The median rating fell in the "valid" range for all 4 NQF domains with statistical agreement according to the BIOMED criteria for agreement we were using to determine if there was agreement among the panelists. Given this strong response, we interpret the measures as written to have face validity.

_____

2b2. EXCLUSIONS ANALYSIS

NA ☑ no exclusions — *skip to section* 2b3

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*)

N/A.

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*)

N/A.

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*)

N/A.

_____

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* 2b4.**

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

☑ **No risk adjustment or stratification**

☐ **Statistical risk model with risk factors**

☐ **Stratification by risk categories**

☐**Other,**

We did not use risk-adjustment for the calculation of the measure score.

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.

N/A.

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.

N/A.

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?

N/A.

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)**

N/A.

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

N/A.

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.**

N/A.

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*)

N/A.

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

***If stratified, skip to* 2b3.9**

N/A.

2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

N/A.

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

N/A.

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

N/A.

2b3.9. Results of Risk Stratification Analysis:

N/A.

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*)

N/A.

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*)

N/A.

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)*

2b4.2. 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*)

Of 20,813 VA patients who received carpal tunnel release, 1,814 patients received OT/PT in the postoperative 6 weeks after CTR (nationwide Measure 3 performance = 91.3%; Table 4). Twenty-one (21) facilities (18.9%) had lower than 90% performance (min-max 32.9-100%), and 333 surgeons (17.8%) had lower than 90% performance. Restricting the analyses to 610 surgeons who performed at least 5 CTR, 179 (29.4%) had lower than 90% performance. The quality measure version that required that the OT/PT visit include a CTS diagnosis resulted in somewhat better performance - 93.5% overall; 13.5% of facilities <90% performance; 20.6% of surgeons with >5 CTRs had <90% performance. In SHC, only 17 of the 640 patients receiving CTR had OT/PT within 6 weeks of their first CTR (Measure 3 performance = 97.3%).

The median facility- level reliability for Measure 3 was 0.95 overall (min-max: 0.30 – 1.00, facility N = 111); 0.95 (0.61-1.00, n = 105) for facilities with at least 20 cases; and 0.95 (0.62 -1.00, n = 99) for facilities with at least 50 cases. Median surgeon-level reliability for Measure 3 was 0.99 overall (min-max: 0.08 – 1.00,

surgeon N = 1185); 0.96 (0.18-1.00, n = 630) for surgeons with at least 5 cases; and 0.95 (0.64-1.00, n = 189) for surgeons with at least 30 cases.

Table 4: Discouraging routine use of in-clinic occupational and/or physical therapy after CTR

System	Time	Denominator	Numerator	Performance	Facility-level Range
VA	2016	7530	7187	95.4%	65.4 – 100.0
VA	2017	7070	6317	89.3%	14.7 – 100.0
VA	2018	6213	5495	88.4%	8.0 – 100.0
VA	FY16-18	20813	19455	91.3%	32.9 – 100.0
SHC	2014-16	640	623	97.3%	NA

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?*)

Our analyses demonstrate that although there is relatively low variation in care, there is still meaningful variation in care and room for improvement. This highlights the importance of this measure for reducing unnecessary healthcare spending for carpal tunnel syndrome release post-procedural care.

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

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

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*)

N/A.

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*)

N/A.

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*)

N/A.

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 non-responders) 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 have no reason to expect that the extent and nature of non-VA care varies between accountable entities.

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*)

N/A

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 non-responders) 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*)

N/A

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.

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

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**.

ALL data elements are in defined fields in electronic claims

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).

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.

This measure requires accurate coding using CPT and ICD-9 or ICD-10 codes.

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)*.

No additional cost beyond accurate claims reporting should be incurred by entities who choose to utilize this measure.

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)
Payment Program Quality Improvement (external benchmarking to organizations) Quality Improvement (Internal to the specific organization)	*

*cell intentionally left blank

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

N/A.

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?*)

We plan to submit this measure to the Centers for Medicare and Medicaid services for consideration for inclusion in Merit-Based Incentive Payment System.

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.)

We plan to submit this measure to the Centers for Medicare and Medicaid services for consideration for inclusion in Merit-Based Incentive Payment System.

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.

This measure´s technical report was provided for public comment for a 30-day period. This was publicized by both the American Academy of Orthopaedic Surgeons (AAOS) and The American Society for Surgery of the Hand (ASSH). The technical report included measure descriptions, specifications, analysis of reliability, validity, and performance variance.

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.

A total of 99 respondents provided feedback for this measure, with 94% generally supportive of the measure as written. The full public comment report is available at:

https://www.aaos.org/uploadedFiles/PreProduction/Quality/Measures/CTS%20PM%20Public%20Comment%20Report.pdf.

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.

The full public comment report is available at:

https://www.aaos.org/uploadedFiles/PreProduction/Quality/Measures/CTS%20PM%20Public%20Comment%20Report.pdf.

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

The full public comment report is available at:

https://www.aaos.org/uploadedFiles/PreProduction/Quality/Measures/CTS%20PM%20Public%20Comment%20Report.pdf.

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

The full public comment report is available at:

https://www.aaos.org/uploadedFiles/PreProduction/Quality/Measures/CTS%20PM%20Public%20Comment%20Report.pdf.

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.

There was a robust response and support to the proposed Measure #3 addressing the routine use of formal therapy after carpal tunnel release. I reviewed all of the written responses, and in collaboration with representatives of the ASSH and the AAOS, we discussed the findings. The workgroup developed this measure based on an AAOS Clinical Practice Guideline with Moderate Evidence for no additional benefit to routine supervised therapy over home programs in the immediate postoperative period. The Workgroup determined that nudging surgeons away from the routine use of formal therapy would improve quality of care by minimizing unnecessary interventions. There were multiple respondents that highlighted the need, at times, for therapy for certain patients with stiff proximal interphalangeal joints, generalized arthritis of the digits, or preoperative stiffness. The Workgroup was in complete agreement with these concerns and discussed these specific examples during the in person meeting. The Workgroup acknowledged that these comorbidities are

uncommon and implementation of the measure would drive utilization towards 0% but it would never reach 0%. The literature does not support the routine use of formal therapy but implementation of the measure will establish a national benchmark of utilization and will identify outliers that routinely use therapy 100% of the time, as we found during our validation study. The workgroup did not believe building in exclusion criteria would be beneficial as the incidence of these comorbidities requiring formal therapy is uncommon when patients are appropriately educated and counseled on a home program.

Robin Kamal, MD

Chair, Carpal Tunnel Quality Measures Workgroup

Vice-Chair, ASSH Quality Metrics Committee

Improvement

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

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

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

The literature does not support the routine use of formal therapy but implementation of the measure will establish a national benchmark of utilization and will identify outliers that routinely use therapy 100% of the time, as we found during our validation study.

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.

N/A.

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

N/A.

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?

Yes

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

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.)

There are no competing measures.

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.

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

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): American Academy of Orthopaedic Surgeons

Co.2 Point of Contact: Ryan, Pezold, pezold@aaos.org, 847-384-4311-

Co.3 Measure Developer if different from Measure Steward: American Academy of Orthopaedic Surgeons

Co.4 Point of Contact: Ryan, Pezold, pezold@aaos.org, 847-384-4311-

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.

1. Steve McCollam, MD

Oversight Chair

2. Robin Kamal, MD

Chair

3. Philip Blazar, MD

American Society for Surgery of the Hand/American Academy of Orthopaedic Surgeons

4. Mia Erickson, PT, EdD, CHT, ACT

American Society of Hand Therapists

5. Brent Graham, MD, MSc

American Society for Surgery of the Hand/American Academy of Orthopaedic Surgeons

6. Andy Gurman, MD

American Society for Surgery of the Hand/American Academy of Orthopaedic Surgeons

7. Peter Jebson, MD

American Society for Surgery of the Hand/American Academy of Orthopaedic Surgeons

8. William Jones, MD

American Academy of Physical Medicine and Rehabilitation

9. John Kincaid, MD

American Academy of Neurology

10. David Ring, MD, PhD

American Association for Hand Surgery

11. John Seiler, MD

American Society for Surgery of the Hand/American Academy of Orthopaedic Surgeons

12. Alex Sox-Harris, PhD, MS

Hand Surgery Quality Consortium

13. John Stephenson, MD

American Society for Surgery of the Hand/American Academy of Orthopaedic Surgeons

14. Jennifer Waljee, MD

American Society of Plastic Surgeons

15. Daniel Wessell, MD, PhD

American College of Radiology

16. Hayes Wilson, MD

American Society for Surgery of the Hand/American Academy of Orthopaedic Surgeons

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2017

Ad.3 Month and Year of most recent revision: 01, 2020

Ad.4 What is your frequency for review/update of this measure? 1 to 3 years

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

The Measures, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for

commercial gain. Commercial uses of the Measures require a license agreement between the user and the American Academy of Orthopaedic Surgeons (AAOS).

The American Society for Surgery of the Hand's significant past efforts and contributions

to the development and updating of the Measures is acknowledged. AAOS is solely responsible for the review and enhancement ("Maintenance") of the Measures as of publication. AAOS encourages use of the Measures by other health care professionals, where appropriate.

Ad.7 Disclaimers: N/A.

Ad.8 Additional Information/Comments: The American Society for Surgery of the Hand should be listed as an additional developer, with AAOS as the sole measure steward.