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 text represents the responses from measure developers.

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

Brief Measure Information

NQF #: 3621

Corresponding Measures:

De.2. Measure Title: Composite weighted average for 3 CT Exam Types: Overall Percent of CT exams for which Dose Length Product is at or below the size-specific diagnostic reference level (for CT Abdomen-pelvis with contrast/single phase scan, CT Chest without contrast/single

Co.1.1. Measure Steward: American College of Radiology

De.3. Brief Description of Measure: Measure title continued: Composite weighted average for 3 CT Exam Types: Overall Percent of CT exams for which Dose Length Product is at or below the size-specific diagnostic reference level (for CT Abdomen-pelvis with contrast/single phase scan, CT Chest without contrast/single phase scan and CT Head/Brain without contrast/single phase scan)

Description: Weighted average of 3 CT Exam Types: Overall Percent of CT exams for which Dose Length Product is at or below the size-specific diagnostic reference level (for CT Abdomen-pelvis with contrast/single phase scan, CT Chest without contrast/single phase scan and CT Head/Brain without contrast/single phase scan)

1b.1. Developer Rationale:

S.4. Numerator Statement: Number of CT Abdomen-Pelvis exams with contrast (single phase scan), CT Chest exams without contrast (single phase scan), and CT Head/Brain exams without contrast (single phase scan) for which Dose Length Product is at or below the size-specific exam-specific diagnostic reference level

S.6. Denominator Statement: Number of CT Abdomen-pelvis exams with contrast (single phase scans), CT Chest exams without contrast (single phase scans), and CT Head/Brain (single phase scans)

Target population: all patients regardless of age.

S.8. Denominator Exclusions: No denominator exclusions

De.1. Measure Type: Composite

S.17. Data Source: Registry Data

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

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

- The developer drew the evidence for this intermediate clinical outcome measure from a systematic review (SR) of 56 studies that examined CT diagnostic reference levels for brain, chest, and abdominal examinations. Garba, I., Zarb, F., McEntee, M. F., & Fabri, S. G. (2020). Computed tomography diagnostic reference levels for adult brain, chest and abdominal examinations: A systematic review. Radiography, S1078817420301723. https://doi.org/10.1016/j.radi.2020.08.011
- The study noted a 2-3 fold variation in diagnostic reference levels (DRLs) between studies for the same procedure. The causes of variation are reported and include study design, scanner technology and the use of different dose indices.
- Studies in the SR were of moderate quality mostly (54) and two of low quality.
- Several additional new studies since the publication of the systematic review were listed and show similar findings.

Questions for the Committee:

- For structure, process, and intermediate outcome measures:
  - As an intermediate clinical outcome, the Standing Committee should consider if a relationship exists between this measure to other observable patient outcomes?
  - How strong is the evidence for this relationship?

Guidance from the Evidence Algorithm

(Box 1) -> No -> Intermediate Clinical Outcome (Box 3) -> Yes -> Systematic Review Provided (Box 4) -> Yes -> Quantity: High, Consistency: Moderate (Box 5b) -> Moderate

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

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

Maintenance measures – increased emphasis on gap and variation

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

2017: Performance Rate: 79.93, Mean: 80.17, # of patients: 1698254, # of groups: 173, Min: 11.01, Max: 100, Std Deviation: 16.82, Interquartile Range: 20.69

2018: Performance Rate: 78.37, Mean: 78.61, # of patients: 1317898, # of groups: 189, Min: 11.01, Max: 100, Std. Deviation: 18.04, Interquartile Range: 22.87

2019: Performance Rate: 79.86, Mean: 78.41, # of patients: 2832268, # of groups: 208, Min: 13.59, Max: 100, Std. Deviation: 18.74, Interquartile Range: 24.34

2020: Performance Rate: 78.32, Mean: 78.47, # of patients: 2832268, # of groups: 205, Min: 13.60, Max: 100, Std. Deviation: 18.85, Interquartile Range: 21.73

CMS recently provided preliminary historical benchmark data for this measure based on reporting for 2019. The measure average performance rate was 80.3% with a range of performance by decile.

Decile 3: 28.83 - 60.42
Decile 4: 60.43 - 73.28
Decile 5: 73.29 - 82.24
Decile 6: 82.25 - 87.25
Decile 7: 87.26 - 89.15
Decile 8: 89.16 - 94.27
Decile 9: 94.28 - 95.13
Decile 10: >= 95.14

Disparities

- No disparities data were provided by the developer, nor did the developer provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement.

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?
- The Standing Committee should discuss whether there is evidence showing that disparities exist in this area of healthcare?

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

1c. Composite – Quality Construct and Rationale

Maintenance measures – same emphasis on quality construct and rationale as for new measures.

1c. Composite Quality Construct and Rationale. The quality construct and rationale should be explicitly articulated and logical; a description of how the aggregation and weighting of the components is consistent with the quality construct and rationale also should be explicitly articulated and logical.

- The measure goal is to decrease preventable harm through effective optimization of computed tomography (CT) protocols and resulting reduction in radiation dose to patients.
- This is a composite weighted average for three CT exam types. The overall score is the percent of CT exams for which Dose Length Product (DLP) is at or below the size-specific diagnostic reference level benchmarks (for CT Abdomen-pelvis with contrast/single phase scan, CT Chest without contrast/single phase scan and CT Head/Brain without contrast/single phase scan).
- This measure will be calculated using the weighted average of three performance rates:
    - Rate 1: Percent of CT Abdomen-pelvis exams with contrast (single phase scan) for which Dose Length Product is at or below the size-specific diagnostic reference level
    - Rate 2: Percent of CT Chest exams without contrast (single phase scan) for which Dose Length Product is at or below the size-specific diagnostic reference level
    - Rate 3: Percent of CT Head/brain exams without contrast (single phase scan) for which Dose Length Product is at or below the size-specific diagnostic reference level

Questions for the Committee:

- Are the quality construct and a rationale for the composite explicitly stated and logical?
- Is the method for aggregation and weighting of the components explicitly stated and logical?

Preliminary rating for composite quality construct and rationale:

☒ **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.

- Evidence is appropriate, drawn from recent literature.
- Moderate quality evidence from systematic reviews and additional citations.
- Measure is based on a systematic review and grading of the body of empirical evidence. The evidence shows 2-3 fold variation in diagnostic reference levels for the same procedure.
- Evidence is Pass.
- moderate
- Reasonable evidence
- composite measure

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

- There is a performance gap evident. It appears there has been little progress on addressing the gap. No disparities noted.
- Significant spread in the data without mention of disparities
- A performance gap is noted however disparities were not provided by the developer.
- High - existing performance gap.
- moderate
- reasonable gap
- moderate evidence - no disparities data provided

1c. Composite Performance Measure - Quality Construct (if applicable): Are the following stated and logical: overall quality construct, component performance measures, and their relationships; rationale and distinctive and additive value; and aggregation and weighting rules?

- Yes
- High rating for quality and rationale of the measure construct.
- Yes
- I'm not sure why these specific three types of CT were chosen?
- high
- Adequate
- High construct rating

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

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

2c. For composite measures: empirical analysis support composite approach

Reliability

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

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

Validity

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

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

Composite measures only:

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

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

Evaluators: NQF Scientific Methods Panel Subgroup

Methods Panel Review (Combined)

Methods Panel Evaluation Summary:

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

Reliability

- Developer calculated a signal-to-noise ratio (SNR) using a Beta-Binomial model (as the event is pass/fail - DLP below benchmark), but calculated the testing only for physician groups, not facilities.
- The reliability score was above .997 for all types of CT's and the composite weighted average. Confidence intervals included the same high reliability.
- SNR analysis on eight million patients in 237 entities.

Validity

- Face validity was conducted as this is a new measure for both group- and facility-level of analysis. The developer reports that:
    - 95% of the panel (20 members) agreed that monitoring radiation dose indices from clinical CT exams is a good and worthwhile activity for advancing or maintaining safety and quality

    - 71% of the panel (15 members) agreed that the measure components as described is a reasonable and appropriate way to assess performance quality of a facility or practice with regards to dose optimization
    - 62% of the panel (13 members) agreed that the scores obtained from the measure would differentiate clinical performance across providers
- Some SMP members questioned the testing methods, stating that additional validity testing could have been conducted with the large available sample asking for split sample testing, specifically because reliability results were very high and generally higher than other composites. One SMP member stated that unusually high reliability results may indicate validity concerns.
- SMP members further questioned the level of analysis (clinician group versus facility), specifically whether face validity was conducted at the clinician group or facility level of analysis or both levels and why stratification was conducted at the clinical group level. The developer has clarified this within their submission that face validity was conducted at both levels of analysis.
- According to one SMP member: "The developers rely on the measure's current use with CMS and its alignment with expert guidelines as demonstration of its face validity. NQF has typically looked for a more formal process."
- The developer also described the measure's validity through consensus documents from a wide range of professional, advisory, and regulatory organizations. Additionally, the use of this measure has significantly increased over the past few years, indicating wider acceptance of this measure by clinicians.
- Lastly, the developer reports that the risk stratification analysis is only performed at the level of the facility and not group, stating that "groups are generally aggregations of facilities – a group supports one or more facilities. Any findings for patient size stratification applicable at the facility level formulation of the measure applies to group level as well."
- The SMP voting result was consensus not reached on validity.

Composite

- Developer demonstrated that performance on one of the component measures has little relationship on other measures, so each component measure does add something "new". The developer all demonstrated that a weighted average (current measure) produces similar results to a straight average.

Questions for the Committee regarding reliability:

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

Questions for the Committee regarding validity:

- Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?
- The Scientific Methods Panel did not reach consensus on validity. The Standing Committee should revote on validity and consider the threats to validity (i.e., exclusions, risk stratification)?

Questions for the Committee regarding composite construction:

- Do you have any concerns regarding the composite construction approach (e.g., do the component measures fit the quality construct and add value to the overall composite? Are the aggregation and weighting rules consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible?)?

- The Scientific Methods Panel is satisfied with the composite construction. Does the Standing Committee think there is a need to discuss and/or vote on the composite construction?

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

Preliminary rating for validity: ☐ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient** – **Consensus was not reached by the SMP (Vote: M-4, L-2, I-2)**

Preliminary rating for composite construction: ☐ **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?

- Reliability is very high. As stated in the document, would like to see breakdown.
- Very high reliability in MD groups but not for facilities.
- The reliability score was high based on method used which was calculated only for physician groups and not facilities.
- no concerns.
- no concerns
- Reasonable
- panel review - moderate reliability

2a2. Reliability - Testing: Do you have any concerns about the reliability of the measure; reliability testing and results for the measure?

- see above
- Very high reliability for construct, especially given weighted average of 3 components
- Yes, based on feedback from SMP members, there was no consensus reached.
- No concerns
- no concerns
- no
- moderate reliability with some concerns

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

- Since consensus was not reached on validity, I would like to hear more from the developer and the committee members
- Face validity is strong but it is unclear whether additional testing is needed such as split sampling. Missed opportunity to explore heterogeneity by aggregating as groups.
- Some SMP members questioned the testing methods, stating that additional validity testing could have been conducted with the large available sample. Unusually high reliability results may indicate validity concerns.
- I would consider Moderate in terms of face validity
- no concerns
- no
- Lack of consensus by panel

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?

- no concerns
- A lot of holes in data presented. Probably need clarification from developers on measurement and validation approach
- Multiple concerns regarding validity testing were raised by the SMP.
- No concerns.
- no concerns
- none

- lack of consensus by panel

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?

- There is no risk-adjustment or exclusions
- No disparity data presented
- Social risk factors were not addressed.
- no concerns
- no concerns
- no issues
- lack of consensus by panel

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

- yes
- It appears composite measures unique aspects of care most agree are important (minimizing radiation doses)
- The composite analysis raises several questions of concern.
- no concerns
- no concerns
- yes
- rated as moderate by panel

Criterion 3. Feasibility

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

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

- ALL data elements are in defined fields in a combination of electronic sources
- The initial setup for submitting data requires the site to have staff resources for installing data collection software. It is a small amount of time to set up the CT equipment to transmit the dose information and to map the site exam names to standardized DIR names for comparison. Occasionally, if done incorrectly, this can require a site to review the set-up and standardized formatting.
- Minimal participation fee to participate in the DIR, which is based on facility size, number of facilities and number of radiologists in each practice. The fee is typically about $500-$1000 per year. The primary purpose of participating sites in DIR is quality improvement, but an additional benefit of this specific measure is the accountability purpose.
    - NRDR and Participation Fees: https://nrdrsupport.acr.org/support/solutions/articles/11000029012-registration-and-participation-fees

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?

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?

- no concerns
- Extractions from registry data; no concerns outside the fee needed to be paid and some concerns with data adjudication
- Requires an initial setup for submitting data which requires staff resources and there is also a fee based on facility size and number of radiologists in the practice.
- No concerns.
- moderate
- feasible
- moderate feasibility, data coming from electronic sources

Criterion 4: Usability and Use

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

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

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

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

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

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

Accountability program details

- Payment Program Merit-based Incentive Payment System qpp.cms.gov
- Quality Improvement (Internal to the specific organization) ACR Dose Index Registry https://www.acr.org/Practice-Management-Quality-Informatics/Registries/Dose-Index-Registry

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

- Feedback is provided to all DIR participants reporting this quality measure daily. Feedback is based on registry benchmarks. ACR educational webinars are conducted bimonthly to explain measure requirements and interpretation of performance results.

Additional Feedback:

- Feedback is obtained through email, the ACR help desk, the CMS quality help desk, and CMS contractor QMMS.
- Feedback on this measure is positive. Facilities are able to evaluate when their CT exam protocols should be reviewed and/or updated to optimize radiation dose exposure to patients.

***Questions for the Committee**:*

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

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

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

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

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

Improvement results

- The developer reports that the "performance has remained steady in the 79-80% for this measure. There hasn't been a significant performance improvement, which demonstrates that there is still a gap in care for optimizing radiation dose to patients. Improving performance in this measure would demonstrate that a facility is adjusting radiation dose protocols."

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

Unexpected findings (positive or negative) during implementation

- The developer reports that they are not aware of any unintended consequences.

Potential harms

- The developer reports that they are not aware of any unintended consequences.

Additional Feedback:

- N/A

***Questions for the Committee**:*

- How can the performance results be used to further the goal of high-quality, efficient healthcare?
- Do the benefits of the measure outweigh any potential unintended consequences?

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

Committee Pre-evaluation Comments:

Criteria 4: Usability and Use

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

- not publicly reported, but used for accountability.
- Participants in registry get feedback on data. Facilities appear able to use data to improve outcomes.
- Feedback on this measure is positive and obtained through multiple sources.
- No concerns
- moderate
- Not publicly reported but used.
- not being publicly reported, use for accountability - state laws?

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.

- no unintended consequences evident
- No reports of unintended consequences. Minimal changes over time suggesting measure may not be responsive though would like to hear from developers.
- The developer reports that performance has been steady in the 79-80% range and that there is a gap in care of optimizing radiation dose to patients.
- No concerns
- benefits > harms
- usable
- rated moderate

Criterion 5: Related and Competing Measures

Related or competing measures

2820 : Pediatric Computed Tomography (CT) Radiation Dose

Harmonization

Our measure, NQF #3621, evaluates the whole population and is not limited to pediatric patients as for NQF #2820. In NQF #3621 performance for facilities and groups is calculated comparing dose indices to published benchmarks. NQF #2820, "provides a simple framework for how facilities can assess their dose, compare their doses to published benchmarks (Smith-Bindman, Radiology, 2015) and identify opportunities to improve if their doses are higher than the benchmarks". Measure users thus are self-calculating results against one of three published benchmarks themselves using one of three benchmarks published benchmarks for both levels of measurement (group and facility). NQF #3621 uses data published in the ACR 2017 study, U.S. Diagnostic Reference Levels and Achievable

Doses for 10 Adult CT Examinations, identifying DRLs and Achievable Doses (ADs) for the 10 most common CT adult examinations performed in the United States. It represents the first time that national adult DRLs and ADs have been developed as a function of patient size, a milestone in optimizing radiation dose to patients. NQF #3621 has eight years of performance data for each measure component, as well as four years of data for the composite. Using electronic data sources, NQF #3621 has high feasibility and low collection burden, which minimizes missing data bias. NQF #3621 provides greater consistency and level of comparison across facilities and groups, providing more validity and reliability for use in quality improvement and specifically for accountability programs.

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?

- no competing measures
- No overlaps with adult patients
- no
- No concerns.
- 2820
- one competing measure for pediatric population

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 06/03/2021

- Comment by: **University of California, San Francisco**

The American College of Radiology (ACR) has proposed measure #3621 titled "Multi-strata weighted average for 3 CT Exam Types: Overall Percent of CT exams for which Dose Length Product is at or below the size-specific diagnostic reference level" for the purpose of measuring the radiation doses used for computed tomography (CT). A quality measure that can inform clinicians about how they can safely lower radiation doses used for diagnostic CT scanning while maintaining the quality of images needed for diagnosis can greatly improve the health and safety of patients. However, the ACR's proposed measure is inadequate for this purpose and, if adopted, could undermine the broad application of more effective ways of using quality measures to achieve this goal. I therefore strongly recommend that National Quality Forum not endorse the proposed measure as it will not reduce the unintended harm of radiation in diagnostic imaging.

The radiation doses used for CT examinations are highly variable across hospitals and imaging facilities for patients imaged for the same indications, are frequently far higher than needed for diagnosis, and are in the range known to be carcinogenic. More than patient or machine characteristics, the most important predictor of radiation dose is the choice the radiologist makes as to what protocol to use (e.g. single-phase scan or double-phase scan). Protocols with more phases deliver proportionally more radiation, yet for most indications, there is no evidence suggesting the higher phase protocol provides better diagnostic utility.

The measure that the ACR has proposed will evaluate radiation doses that are used for three specific protocols: a single-phase head, single-phase chest, and single-phase abdomen. The measure will assess doses in these three groups against benchmarks only after the primary decision of protocol selection is made. This limited assessment of dose within these stratified groups ignores the primary factor determining the patient's dose, i.e. which protocol to use, which is almost entirely at the discretion of the imaging physician. The measure will assess only the relatively smaller variation in technical parameters within single-phase head, chest, or abdomen protocols, but will leave unassessed the variation that occurs due to the choice of protocol. The unnecessary variation in protocol selection is the critical factor, but the ACR measure over-adjusts for this factor by stratifying based on the protocol. The ACR

defines the target population for the measure as "all patients who require either a CT Abdomen-pelvis exam with contrast (single-phase scans), a CT Chest exam without contrast (single-phase scans), and/or a CT Head/Brain (single-phase scans) exam." However, the measure fails to identify patients who require these exams based on their clinical need, but who instead received much higher doses through multi-phase exams, when the single-phase study would have been appropriate. In the University of California, San Francisco International CT Dose Registry, which includes over 8 million CT scans from 162 hospitals and image facilities, these three CT exam types together make up 39% of exams overall across the registry. However, they account for 1% to 83% of exams across the different hospitals and imaging facilities, suggesting the denominator for this measure does not reflect a patient population who require these exams, but rather reflects the varying decisions of radiologists to assign patients to different protocols.
The only way to accurately judge physicians in their use of radiation for CT is to evaluate how they use radiation in a population of patients where their selection of imaging protocol is included in the assessment. Radiation doses need to be assessed based on the intent and clinical question of the provider ordering the scan, not on the radiologist's subjective choice of protocol, which is too often driven more by preference than clinical need. The measurement of the dose within the ACR's narrowly defined groups will only camouflage the large variation in practice that exists and will not serve to improve practice.
The University of California, San Francisco was contracted by CMS to develop a quality measure for CT for use in the MIPS payment program. The measure was submitted to the CMS MUC list in May 2021 and will be submitted to NQF in August. This measure assesses radiation doses among adult patients who undergo diagnostic CT based on the diagnoses and clinical questions generated at the time of the test order, and therefore is not undermined by the concern raised in the ACR measure.
Rebecca Smith-Bindman, MD
University of California, San Francisco

- No NQF Members have submitted support/non-support choices as of this date.

Combined Methods Panel Scientific Acceptability Evaluation

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 3621

Measure Title: Composite weighted average for 3 CT Exam Types: Overall Percent of CT exams for which Dose Length Product is at or below the size-specific diagnostic reference level (for CT Abdomen-pelvis with contrast/single phase scan, CT Chest without contrast/single

RELIABILITY: SPECIFICATIONS

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

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

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

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

   Panel Member 1: No concerns.

   Panel Member 2: This measure is to ensure that the dose is not unusually high for a specified procedure. This may be a more common issue but it is not clear how to guard against the occurrence of unusually low does for a specified procedure. This measure is specified as Dose Length Product at or below the size-specific reference level.

   Panel Member 3: No concerns

Panel Member 4: Fairly well defined on the MIF but definitions were not consistent e.g., for the last category "Head/Brain" did not consistently indicate whether the exam was without contrast. I am assuming that it is without contrast.

Panel Member 5: None

Panel Member 6: None

Panel Member 8: None

RELIABILITY: TESTING

Type of measure:

☐ **Outcome (including PRO-PM)** ☐ **Intermediate Clinical Outcome** ☒ **Process**

☐ **Structure** ☒ **Composite** ☐ **Cost/Resource Use** ☐ **Efficiency**

Data Source:

☐ **Abstracted from Paper Records** ☐ **Claims** ☒ **Registry**
☐ **Abstracted from Electronic Health Record (EHR)** ☐ **eMeasure (HQMF) implemented in EHRs** ☐ **Instrument-Based Data** ☐ **Enrollment Data** ☐ **Other (please specify)**

Level of Analysis:

☐ **Individual Clinician** ☒ **Group/Practice** ☒ **Hospital/Facility/Agency** ☐ **Health Plan**
☐ **Population: Regional, State, Community, County or City** ☐ **Accountable Care Organization**
☐ **Integrated Delivery System** ☐ **Other (please specify)**

Measure is:

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

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

   Panel Member 1: Developer calculated a signal-to-noise ratio using a Beta-Binomial model (as the event is pass/fail - DLP below benchmark), but calculated the testing only for physician groups, not facilities.

   Panel Member 2: The developer calculated signal to noise reliability for clinician group using beta-binomial model. The reliability scores were extremely high in part due to high group level sample size.

   Panel Member 3: No concerns.

   Panel Member 4: Appropriate

Panel Member 5: A signal-to-noise ratio (SNR) analysis test on the performance data for reliability. In SNR analysis, reliability is the measure of confidence in differentiating performance between physicians or other providers; in this case, radiology groups.

Panel Member 6: Over 8 million patients reported across the 3 CT types. Did not see exact breakdown of group/ practice versus hospital/facility/agency. I assume the term "radiology groups" can also mean facility. Signal to noise reliability testing was performed using a beta binomial methodology. Testing was limited to N greater than or equal to 10 patients during the timeframe.

Panel Member 7: SNR

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

   Submission document: Testing attachment, section 2a2.3

   Panel Member 1: Found the reliability testing results to be quite high (median value of 0.9999), raising the question if an appropriate method was used.

   Panel Member 2: The reliability scores were extremely high in part due to high group level sample size.

   Panel Member 3: Results of the STN reliability analysis are practically perfect, which seems to be related to the groups sample size. Therefore, patient sample size by group and year would be a useful addition to table 1. Also, the range/IQR of reliability scores per group would be important to see to get a better understanding of their distribution.

   Panel Member 4: Adequate

   Panel Member 6: Amazingly, the reliability score was above .997 for all types of CT's and the composite weighted average. Confidence intervals included the same high reliability.

   Panel Member 7: >0.98; This is a process measure.

   Panel Member 8: SNR analysis on 8 million patients in 237 entities.

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 **all** testing results):

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

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

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

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

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

Panel Member 1: The developer indicated that the measure is specified for a facility, but did not provide any facility-level reliability testing. In addition, the median SNR values are extremely high, making me question whether their approach was appropriate for this situation.

Panel Member 4: Based on the testing and the clarity of the measure description (Numerator and Denominator).

Panel Member 5: Mean reliability scores of 0.9999 for the composite measure

Panel Member 6: Result above .997 for all types of CT and the composite for this group of patients.

Panel Member 7: Would a SME have input on this?

Panel Member 8: Near perfect reliability as measured by the SNR. Especially with the very large sample size, a split sample reliability analysis and a 'stability of classification' analysis would have been illuminating.

VALIDITY: TESTING

12. **Validity testing level:** ☒ **Measure score** ☒ **Data element** ☐ **Both**
13. **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)

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

☒ **Face validity**

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

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

15. **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)

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

Submission document: Testing attachment, section 2b2.2

Panel Member 1: For face validity testing, NQF typically looks for a systematic and transparent process whereby independent experts are explicitly asked if the measure can distinguish high quality from poor quality care. It does not appear as if the developers engaged outside experts in this evaluation.

Panel Member 2: The developer primarily established the measure validity by consensus documents by a wide range of professional, advisory, and regulatory organizations. Additionally, the use of this measure has significantly increased over the past few years, indicating wider acceptance of this measure by clinicians.

Panel Member 3: No concerns

Panel Member 4: The developer uses approval by CMS and their contractors as evidence of validity of the measure. Here I am assuming that they mean the composite measure score although it was not clear whether it's

the composite score of the individual component scores of the measures within the composite in the testing document submitted. While they reference Table 6 (actually Table 5) as indicating consensus agreement that the use of diagnostic reference levels is a good indicator of quality and dose optimization which was demonstrated by quotes from various organizations. It appears they did not convene a panel to establish face validity but provided access to various reports. I'm not sure of the methods they used to collect, review and evaluate the literature they did review.

Panel Member 5: Face validity only it appears that multiple entities have been reporting their performance through QCDR. I question why they did not analyze the results available to them.

Panel Member 6: Face validity was used by a consensus process. A number of professional societies, governmental agencies, and various committees and commissions, both national and international have indicated their support of the measure. Evidence was provided of increased usage of this measure by providers since 2017 as indicative of support for a valid measure. Decile performance rates are provided and range from 28-60 in decile 3 to greater than 95 for decile 10.

Panel Member 7: Consensus statements (many)

Panel Member 8: Face validity that the components and composite are important and well operationalized.

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

Submission document: Testing attachment, section 2b2.3

Panel Member 1: The developers rely on the measure's current use with CMS and its alignment with expert guidelines as demonstration of its face validity. NQF has typically looked for a more formal process.

Panel Member 2: Although there is no specific empirical validity testing at measure score level, the developer made a convincing case why this measure is conceptually valid with endorsing documents from relevant organizations. Increasing adoption of this measure by clinician groups also lend support to this measure.

Panel Member 3: No concerns

Panel Member 4: The developer uses approval by CMS and their contractors as evidence of validity of the measure. Here I am assuming that they mean the composite measure score although it was not clear whether it's the composite score of the individual component scores of the measures within the composite in the testing document submitted. While they reference Table 6 (actually Table 5) as indicating consensus agreement that the use of diagnostic reference levels is a good indicator of quality and dose optimization which was demonstrated by quotes from various organizations It appears they did not convene a panel to establish face validity but provided access to various reports. I'm not sure of the methods they used to collect, review and evaluate the literature they did review.

Panel Member 5: Demonstrated acceptance of the measure among multiple stakeholders

Panel Member 6: Validity is established by expert opinion and increased utilization since its inception in 2017.

Panel Member 7: As a face validity concept - very good.

Panel Member 8: Fine

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

Submission document: Testing attachment, section 2b2.

Panel Member 1: No exclusions.

Panel Member 2: No concern

Panel Member 3: NA

Panel Member 4: No exclusions.

Panel Member 5: No exclusions noted

Panel Member 6: There are no exclusions from those who submitted data.

Panel Member 7: None.

19. **Risk Adjustment**

Submission Document: Testing attachment, section 2b3

19a. **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☒ **Stratification**

Panel Member 4: 14 risk categories

19b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

☒ Yes ☐ No ☒ Not applicable

19c. **Social risk adjustment:**

19c.1 Are social risk factors included in risk model? ☐ Yes ☒ No ☒ Not applicable

19c.2 Conceptual rationale for social risk factors included? ☐ Yes ☒ No

19c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☐ Yes ☒ No

19d. **Risk adjustment summary:**

19d.1 All of the risk-adjustment variables present at the start of care? ☐ Yes ☐ No

19d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☐ Yes ☐ No

19d.3 Is the risk adjustment approach appropriately developed and assessed? ☐ Yes ☒ No

19d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☐ Yes ☒ No

19d.5. Appropriate risk-adjustment strategy included in the measure? ☒ Yes ☒ No

19e**. Assess the risk-adjustment approach**

Panel Member 1: The measure should be stratified by patient size, so each stratum is compared to size-based DRLs. This seems like a logical step.

Panel Member 2: Given the nature of this measure and measure outcome is already procedure and size specific, risk adjustment is not necessary.

Panel Member 4: The developer provided minimal information with regard to stratification e.g., two articles found on Google Scholar. "Interpretation of the comparison of stratified and unstratified measures will be provided with final submission." ???

Panel Member 5: No justification for not risk adjusting provided

Panel Member 6: The measure is risk adjusted by type of CT and the size of the patient.

Panel Member 7: I did not see final reporting on this (which ideally should be done before SNR reliability analysis...)

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

Submission document: Testing attachment, section 2b4.

Panel Member 1: The measure developer should provide results by the levels of specification - clinician group and facility - not aggregated results.

Panel Member 2: The performance score range is quite wide, decile 3 is from 28.83 - 60.42% and decile 10 is greater than 95.14%.

Panel Member 3: No concerns

Panel Member 4: Minimal data provided: Descriptive statistics and Student's t-test.

Panel Member 6: Mean, median, standard deviation, and interquartile ranges are provided for performance across the various types of CT's. The paper delineates variation by the size contribution as a variable.

Panel Member 7: Cannot - I do not see the risk adjustment reporting.

Panel Member 8: Good variation in practice

21. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**
Submission document: Testing attachment, section 2b5.

Panel Member 1: Not applicable.

Panel Member 2: No concern

Panel Member 5: Section 2b4.2 demonstrates statistical differences in performance rates

Panel Member 6: None

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

Submission document: Testing attachment, section 2b6.

Panel Member 1: No missing data, as the data are generated from the scanner.

Panel Member 2: No concern

Panel Member 3: There are no missing data for this measure

Panel Member 4: There is no double check on whether the ACR NRDR Dir participants check on the quality of the data e.g., what does the developer do to ensure data accuracy? The developer indicates that "no missing data was found through testing, nor would missing data be expected to occur in the future." No testing results presented.

Panel Member 5: None

Panel Member 6: No missing data is technically possible given the direct submission from the software in the treating machines to the registry, for participating providers.

For cost/resource use measures ONLY:

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

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

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

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

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

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

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

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

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

Panel Member 1: Face validity was not systematically assessed by recognized independent experts. Developers relied on its current use and alignment with national guidelines as proof of its face validity.

Panel Member 2: No score-level testing was performed but face validity seems compelling.

Panel Member 3: Score level testing not conducted.

Panel Member 4: There is no double check on whether the ACR NRDR Dir participants check on the quality of the data e.g., what does the developer do to ensure data accuracy? The developer indicates that "no missing data was found through testing, nor would missing data be expected to occur in the future." No testing results presented.

Panel Member 5: Could not identify score level testing or data element testing for validity as required per NQF instructions

Panel Member 6: Face validity is strong. Size and type of CT are the major/only measured sources of variation provided.

Panel Member 7: I did not see the risk adjustment work.

Panel Member 8: Reasonable evidence of face validity

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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

    ☒ **High**

    ☒ **Moderate**

    ☐ **Low**

    ☒ **Insufficient**

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

    Panel Member 1: Developer demonstrated that performance on one of the component measures has little relationship on other measures, so each component measure does add something "new". The developer all demonstrated that a weighted average (current measure) produces similar results to a straight average.

    Panel Member 2: This is more like an opportunities based composite and the way it is constructed is sound.

    Panel Member 3: Strong face validity of the composite.

    Panel Member 4: Developer did not provide enough data to make a judgement.

    Panel Member 5: Stewards provided adequate analysis that there was no significant difference between composite score and individual measure scores.

    Panel Member 7: Negative correlation between some measures. weighting seems not important.

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.**

    Panel Member 1: The developer should provide both reliability and validity testing for each level they are seeking endorsement (facility, clinician group). They also should take a more formal approach to their face validity testing.

    Panel Member 2: It will be important to guard against unusually low dose given this measure is about at or below the size-specific diagnostic reference level.

Panel Member 5: It appears measure stewards have access to perform measure score validity testing. I would recommend that they do this analysis and resubmit

Panel Member 6: Interesting measure that has high reliability, performance means in the mid 70's, with 25th percentiles in the 70's and increased utilization over the last 4 years. Statistical testing of validity is not the strongest but probably acceptable.

Developer Submission

NQF #: 3621

Corresponding Measures:

De.2. Measure Title: Composite weighted average for 3 CT Exam Types: Overall Percent of CT exams for which Dose Length Product is at or below the size-specific diagnostic reference level (for CT Abdomen-pelvis with contrast/single phase scan, CT Chest without contrast/single

Co.1.1. Measure Steward: American College of Radiology

De.3. Brief Description of Measure: Measure title continued: Composite weighted average for 3 CT Exam Types: Overall Percent of CT exams for which Dose Length Product is at or below the size-specific diagnostic reference level (for CT Abdomen-pelvis with contrast/single phase scan, CT Chest without contrast/single phase scan and CT Head/Brain without contrast/single phase scan)

Description: Weighted average of 3 CT Exam Types: Overall Percent of CT exams for which Dose Length Product is at or below the size-specific diagnostic reference level (for CT Abdomen-pelvis with contrast/single phase scan, CT Chest without contrast/single phase scan and CT Head/Brain without contrast/single phase scan)

1b.1. Developer Rationale:

S.4. Numerator Statement: Number of CT Abdomen-Pelvis exams with contrast (single phase scan), CT Chest exams without contrast (single phase scan), and CT Head/Brain exams without contrast (single phase scan) for which Dose Length Product is at or below the size-specific exam-specific diagnostic reference level

S.6. Denominator Statement: Number of CT Abdomen-pelvis exams with contrast (single phase scans), CT Chest exams without contrast (single phase scans), and CT Head/Brain (single phase scans)

Target population: all patients regardless of age.

S.8. Denominator Exclusions: No denominator exclusions

De.1. Measure Type: Composite

S.17. Data Source: Registry Data

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

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_3621_Evidence_Attachment.docx

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

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

1a. Evidence (subcriterion 1a)

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

Measure Title: Composite weighted average for 3 CT Exam Types: Overall Percent of CT exams for which Dose Length Product is at or below the size-specific diagnostic reference level (for CT Abdomen-pelvis with contrast/single phase scan, CT Chest without contrast/single phase scan and CT Head/Brain without contrast/single phase scan).

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

Date of Submission: 4/2/2021

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*): Each measure captures how well radiation exposure from the scanner is adjusted for patient size, using size-specific exam-level diagnostic reference levels and how well total radiation exposure from an exam is optimized based on the CT dose index dose-length product (DLP). A single composite performance measure consisting of these three indicators allows physicians and facilities to accurately view which body area exam may require further improvement on dose protocols.

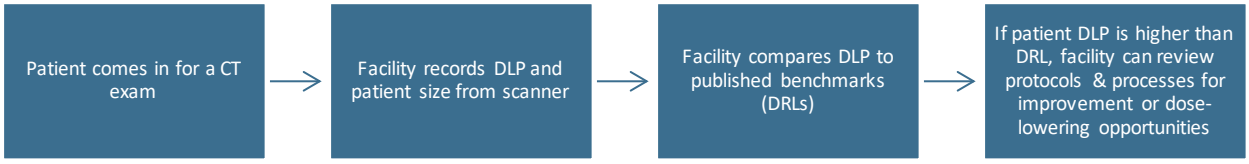
☐ Process:

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

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



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.)

****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.

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

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

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

☐ Other

Systematic Review	Evidence
Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	Computed tomography diagnostic reference levels for adult brain, chest and abdominal examinations: A systematic review Garba, I., Zarb, F., McEntee, M. F., & Fabri, S. G. September 15, 2020

Systematic Review	Evidence
	Garba, I., Zarb, F., McEntee, M. F., & Fabri, S. G. (2020). Computed tomography diagnostic reference levels for adult brain, chest and abdominal examinations: A systematic review. *Radiography*, S1078817420301723. https://doi.org/10.1016/j.radi.2020.08.011 PubMed: https://pubmed.ncbi.nlm.nih.gov/32948454/
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.	*C Conclusions*: "The study noted a 2-3 fold variation in DRLs between studies for the same procedure. The causes of variation are reported and include study design, scanner technology and the use of different dose indices." Kanal et al (Radiology 2017), referenced in our NQF submission form, was the only study in the systematic review to report size-based radiation dose indices. *Implications for practice:* "There is a need for standardization of CT DRLs in line with recommendations from the International Commission on Radiological Protection (ICRP) to reduce dose variation and facilitate dose comparison." Future DRLs should include size-based recommendations.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Of the 56 studies included in the systematic review, two were graded as weak and not included in the discussion. The remaining 54 studies were graded as moderate. The authors state the following: "The quality of each of the included articles was assessed by the primary reviewer using the Effective Public Health Practise Project (EPHPP) tool for quantitative studies. Each article was graded as weak, moderate or strong using the quality assessment scale provided in the EPHPP tool."
Grade assigned to the **recommendation** with definition of the grade	Recommendation(s) not assigned grade(s).
Provide all other grades and definitions from the recommendation grading system	N/A

Systematic Review	Evidence
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	A total of 56 studies were included in the systematic review. Of those 56, two were rated as weak and were not included in the discussion so as to avoid poor studies biasing the overall findings. The remaining 54 studies were rated as moderate, in large part because conducting a randomized trial or other study with no selection bias is incredibly difficult—likely impossible—at this time due to technical limitations.
Estimates of benefit and consistency across studies	In fact, it's the lack of consistency across the studies which indicates the quality gap in clinical practice. There was variation in radiation dose in the human studies up to a factor three, and a factor of two in the phantom studies. Further, the phantom studies may not faithfully represent clinical practice; there were eight studies included in the review that reported exclusively phantom results, and one study that reported phantom and human results.
What harms were identified?	Inconsistent results across the published literature indicate that there is lack of standardization in clinical practice, and lack of reliable benchmarks—DRLs—for radiology practices to use as comparison. Using size-based DRLs based on large sample sizes would a major step forward for clinical practice.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	None of the following new studies change the conclusions of the systematic review: Abuzaid, M. M., Elshami, W., El Serafi, A., Hussien, T., McConnell, J. R., & Tekin, H. O. (2020). Toward national CT diagnostic reference levels in the united Arab emirates: A multicenter review of CT dose index and dose length product. *Radiation Protection Dosimetry*, *190*(3), 243–249. https://doi.org/10.1093/rpd/ncaa100 AlNaemi, H., Tsapaki, V., Omar, A. J., AlKuwari, M., AlObadli, A., Alkhazzam, S., Aly, A., & Kharita, M. H. (2020). Towards establishment of diagnostic reference levels based on clinical indication in the state of Qatar. *European Journal of Radiology Open*, *7*, 100282. https://doi.org/10.1016/j.ejro.2020.100282 Benmessaoud, M., Dadouch, A., Talbi, M., Tahiri, M., & El-ouardi, Y. (2020). Diagnostic reference levels for paediatric head computed tomography in morocco: A nationwide

Systematic Review	Evidence
	survey. *Radiation Protection Dosimetry*, *191*(4), 400–408. https://doi.org/10.1093/raddos/ncaa170 Compagnone, G., Padovani, R., D'Ercole, L., Orlacchio, A., Bernardi, G., D'Avanzo, M. A., Grande, S., Palma, A., Campanella, F., & Rosi, A. (2021). Provision of Italian diagnostic reference levels for diagnostic and interventional radiology. *La Radiologia Medica*, *126*(1), 99–105. https://doi.org/10.1007/s11547-020-01165-3 Joseph Zira, D., Haruna Yahaya, T., Umar, M. S., Nkubli B, F., Chukwuemeka, N. C., Sidi, M., Emmanuel, R., Ibrahim, F. Z., Laushugno, S. S., & Ogenyi, A. P. (2020). Clinical indication-based diagnostic reference levels for paediatric head computed tomography examinations in Kano Metropolis, northwestern Nigeria. *Radiography*, S1078817420302509. https://doi.org/10.1016/j.radi.2020.11.021 Khelassi-Toutaoui, N., Merad, A., Tsapaki, V., Meddad, F., Sakhri-Brahimi, Z., Guedioura, D., & Saadi, S. (2020). Adult CT examinations in Algeria: Towards updating national diagnostic reference levels. *Radiation Protection Dosimetry*, *190*(4), 364–371. https://doi.org/10.1093/rpd/ncaa116 Ploussi, A., Syrgiamiotis, V., Makri, T., Hatzigiorgi, C., & Efstathopoulos, E. P. (2020). Local diagnostic reference levels in pediatric CT examinations: A survey at the largest children's hospital in Greece. *The British Journal of Radiology*, *93*(1116), 20190358. https://doi.org/10.1259/bjr.20190358 Yurt, A., Özsoykal, İ., Kandemir, R., & Ada, E. (2020). Local study of diagnostic reference levels for computed tomography examinations of adult patients in Izmir, turkey. *Radiation Protection Dosimetry*, *190*(4), 446–451. https://doi.org/10.1093/rpd/ncaa121

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.

There are several publications that are practice guidelines and/or expert consensus recommendations based on synthesis of available evidence.

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

Practices should use DRLs as guidance when reviewing their own clinical CT performance. DRLs are the first step in the optimization process to manage patient dose commensurate with the medical purpose of the procedure. Quality assurance programs should include ongoing monitoring and comparison of clinical dose metrics to published normative data including reference levels and achievable levels. As part of a quality assurance program emphasizing radiation management, practices should monitor doses to patients and check the facility doses against DRLs, where available. The DRL is an essential tool in the optimization process, especially as dose limits are not relevant in the medical exposure of patients. The DRL has proven to be an effective tool that aids in optimization of protection in the medical exposure of patients for diagnostic and interventional procedures. Comparing clinical performance to DRLs enables practices to understand whether, in routine conditions, the patient dose from a specified procedure is unusually high or low for that procedure. All examinations resulting in high collective doses should have DRLs—CT delivers the highest collective dose of all medical imaging procedures. The application of DRLs should be the responsibility of all providers of X-ray imaging. This means that DRLs should also be applied to imaging performed outside the radiology department, including cardiology, orthopedic surgery, gastroenterology, intensive care (line placement), neurology, vascular surgery, etc. Specific considerations may also be appropriate for imaging associated with radiation therapy where the purpose and scope of imaging can be different. Practices should compare CT exposures to Diagnostic Reference Levels (DRLs) to compare their protocols to regional and national values. Pediatric facilities should establish DRLs and compare their routine clinical dose index data to them. Adult DRLs can be used to establish pediatric DRLs by using physics principles to account for smaller patient size.

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

The literature synthesized above and cited below was curated by ACR Senior Advisor for Medical Physics, Dustin Gress. Mr. Gress is a diagnostic and nuclear medical physicist, board certified by the American Board of Radiology and the American Board of Science in Nuclear Medicine. He has approximately 14 years of clinical experience, spending roughly half in private practice—supporting upwards of 200 client facilities ranging from academic hospitals to rural standalone imaging clinics—and the other half in a high-volume academic cancer hospital. Mr. Gress selected the submitted literature based on his experience as a medical physicist, in order to demonstrate broad, both national and international, expert consensus support for medical imaging practices to monitor their use of radiation dose in patient imaging and compare their performance to available benchmarks. The organizations whose documents are referenced are the standard bearers in their space and are widely followed. ICRP guidance is followed by EU nations and others around the world for national policymaking and clinical practice guidance; the NCRP is similarly regarded in the US. ACR and AAPM are professional organizations that define standards of care for medical physics and the radiological professions in the US.

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

NCRP Report No. 172, *Reference Levels and Achievable Doses in Medical and Dental Imaging: Recommendations for the United States*

Report of AAPM Task Group 232, *Current state of practice regarding digital radiography exposure indicators and deviation indices*

U.S. FDA, *Medical X-ray Imaging*

ICRP Publication 135, *Diagnostic reference levels in medical imaging*

UNSCEAR 2013 Report to the General Assembly, *Sources, Effects, and Risks of Ionizing Radiation*, Volume II: Scientific Annex B

European Commission, Radiation Protection No. 185, *European guidelines on diagnostic reference levels for paediatric imaging*

ACR Computed Tomography Quality Control Manual

ACR–AAPM–SPR PRACTICE PARAMETER FOR DIAGNOSTIC REFERENCE LEVELS AND ACHIEVABLE DOSES IN MEDICAL X-RAY IMAGING

International Atomic Energy Agency

IAEA Safety Standards, Specific Safety Guide No. SSG-46, *Radiation Protection and Safety in Medical Uses of Ionizing Radiation*

Image Wisely, *CT Protocol Design*

Image Gently (instructions)

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.*

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.*

2017: Performance Rate: 79.93, Mean: 80.17, # of patients: 1698254, # of groups: 173, Min: 11.01, Max: 100, Std Deviation: 16.82, Interquartile Range: 20.69

2018: Performance Rate: 78.37, Mean: 78.61, # of patients: 1317898, # of groups: 189, Min: 11.01, Max: 100, Std. Deviation: 18.04, Interquartile Range: 22.87

2019: Performance Rate: 79.86, Mean: 78.41, # of patients: 2832268, # of groups: 208, Min: 13.59, Max: 100, Std. Deviation: 18.74, Interquartile Range: 24.34

2020: Performance Rate: 78.32, Mean: 78.47, # of patients: 2832268, # of groups: 205, Min: 13.60, Max: 100, Std. Deviation: 18.85, Interquartile Range: 21.73

CMS recently provided preliminary historical benchmark data for this measure based on reporting for 2019. The measure average performance rate was 80.3% with a range of performance by decile.

Decile 3: 28.83 - 60.42

Decile 4: 60.43 - 73.28

Decile 5: 73.29 - 82.24

Decile 6: 82.25 - 87.25

Decile 7: 87.26 - 89.15

Decile 8: 89.16 - 94.27

Decile 9: 94.28 - 95.13

Decile 10: >= 95.14

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.

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.*

We are unable to provide disparities by population group, but we can provide facility characteristics for this measure. In particular, the gap between the metropolitan facilities and the rural facilities show how many more patients are being seen in metropolitan communities. We hope this gap continues to improve with continued use of the measure.

Facility category	# of facilities	# of patients
Academic	173	4,014,721
Community hospital	1,277	17,776,843
Multi-specialty clinic	119	412,793
Freestanding center	623	1,450,846

Children´s hospital	33	92,927
Other	108	320,303

Facility location	# of facilities	# of patients
Metropolitan	1,011	13,351,998
Suburban	837	7,751,000
Rural	438	2,965,435

Census region	# of facilities	# of patients
Northeast	473	5,588,555
Midwest	537	4,708,684
South	890	10,224,294

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

1c. Composite Quality Construct and Rationale

1c.1. A composite performance measure is a combination of two or more component measures, each of which individually reflects quality of care, into a single performance measure with a single score.

For purposes of NQF measure submission, evaluation, and endorsement, the following will be considered composites:

- Measures with two or more individual performance measure scores combined into one score for an accountable entity.
- Measures with two or more individual component measures assessed separately for each patient and then aggregated into one score for an accountable entity:
  - all-or-none measures (e.g., all essential care processes received, or outcomes experienced, by each patient);

1c.1. Please identify the composite measure construction: two or more individual performance measure scores combined into one score

1c.2. Describe the quality construct, including:

- the overall area of quality
- included component measures and
- the relationship of the component measures to the overall composite and to each other.

The measure goal is to decrease preventable harm through effective optimization of computed tomography (CT) protocols and resulting reduction in radiation dose to patients.

This is a composite weighted average for 3 computed tomography (CT) exam types. The overall score is the percent of CT exams for which Dose Length Product (DLP) is at or below the size-specific diagnostic reference level benchmarks (for CT Abdomen-pelvis with contrast/single phase scan, CT Chest without contrast/single phase scan and CT Head/Brain without contrast/single phase scan).

This measure will be calculated using the weighted average of three performance rates:

Rate 1: Percent of CT Abdomen-pelvis exams with contrast (single phase scan) for which Dose Length Product is at or below the size-specific diagnostic reference level

Rate 2: Percent of CT Chest exams without contrast (single phase scan) for which Dose

Length Product is at or below the size-specific diagnostic reference level

Rate 3: Percent of CT Head/brain exams without contrast (single phase scan) for which Dose

Length Product is at or below the size-specific diagnostic reference level

Dose Length Product (DLP) is a standardized parameter to measure computed tomography (CT) scanner radiation output to a patient and is a useful index to compare protocol-based outputs across different practices and scanners. Providing comparative performance data across CT exam types (e.g., head, chest, and abdomen) to a physician or site will help identify where imaging protocols may need adjustment in order to obtain diagnostic images using the lowest reasonable dose. While DLP itself is not a measure or estimate of actual patient radiation dose, it is closely related to doses received by patients. DLPs cover scan length, which is important in terms of capturing radiation exposure to patients. Physicians can see DLP on their PACS for each exam, which allows for feedback and care coordination between the physician, technologist, and medical physicist in improving scan lengths.

Diagnostic reference levels (DRLs) are used as benchmarks for radiation protection and optimization of patient imaging. The intended use of DRLs is as a simple test for identifying situations where the levels of patient dose are unusually high and provide a means for facilities and clinicians to optimize dose to a lower level than a DRL. In 2017, the American College of Radiology (ACR) published a study, U.S. Diagnostic Reference Levels and Achievable Doses for 10 Adult CT Examinations, identifying DRLs and Achievable Doses (ADs) for the 10 most common CT adult examinations performed in the United States, The study used 2014 data submitted to the National Radiology Data Registry – Dose Index Registry (DIR) for 1,310,727 CT head, neck and body exams. It represents the first time that national adult DRLs and ADs have been developed as a function of patient size. This data enables facilities to effectively compare their patient doses with national benchmarks and to optimize their CT protocols, resulting in lower doses at the appropriate image quality. DRLs should be used to determine if a facility's dose indexes are unusually high; they should not be used as target doses. Both ADs and DRLs are provided to encourage facilities to optimize dose to a lower level than that indicated by the DRL.

This measure and its components measures the DLP of CT exam for a particular aspect of the body (abdomen/pelvis, chest, and head/brain). It is imperative to measure each body area separately since they all have a different DLP requirement and using a weighted average for the three different exam types ensures that physician performance is accurately captured.

There are several potentially justified reasons for variations in dose exposure, such as indication for exam and patient size. We define the exams fairly narrowly for each component measure which narrows variability driven by indication. We stratify records by patient size and compare each record to a size specific DRL to ensure unbiased comparison across patient populations.

Reference:

Kanal KM, Butler PF, Sengupta D, Bhargavan-Chatfield M, Coombs LP, Morin RL. U.S. Diagnostic Reference Levels and Achievable Doses for 10 Adult CT Examinations. Radiology. 2017 Jul;284(1):120-133. doi: 10.1148/radiol.2017161911. Epub 2017 Feb 21. PMID: 28221093.

1c.3. Describe the rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually.

This performance measure was initially developed as three individual quality measures. The ACR combined the three into a composite performance measure in 2019 to consolidate the concept of radiation safety for CT exams to a single measure for optimal radiation dose. Each measure captures how well radiation exposure from the scanner is adjusted for patient size, using size-specific exam-level diagnostic reference levels and how well total radiation

exposure from an exam is optimized based on the CT dose index dose-length product (DLP). A single composite performance measure consisting of these three indicators allows physicians and facilities to accurately view which body area exam may require further improvement on dose protocols.

1c.4. Describe how the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale.

This measure is calculated using the weighted average of three performance rates:

- Rate 1: Percent of CT Abdomen-pelvis exams with contrast (single phase scan) for which Dose Length Product is at or below the size-specific diagnostic reference level
- Rate 2: Percent of CT Chest exams without contrast (single phase scan) for which Dose Length Product is at or below the size-specific diagnostic reference level
- Rate 3: Percent of CT Head/brain exams without contrast (single phase scan) for which Dose Length Product is at or below the size-specific diagnostic reference level

Composite score:

Each component measure percentile score is weighted by the denominator count. The weighted scores are summed then divided by the sum of weights of all 3. Alternatively, the numerator and denominator counts for each measure can be totaled then averaged by 3.

Example:

	Numerator	Denominator	Rate
Head	3000	8000	38%
Abdomen/Pelvis	5000	10000	50%
Chest	2000	5000	40%
All	10000	23000	43%
Weighted average			43%

Weighted average = (Weight Head x Rate Head) + (Weight Abdomen/Pelvis x Rate Abdomen/Pelvis) + (Weight Chest x Rate Chest)))/Sum of weights of all 3

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.acr.org/-/media/ACR/Files/Registries/QCDR/2021-QCDR-Measure-Specification-Details.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 **:** ACRad_34_-_Multistrata_weighted_average_of_three_CT_exam_types.pdf

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.

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.

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 CT Abdomen-Pelvis exams with contrast (single phase scan), CT Chest exams without contrast (single phase scan), and CT Head/Brain exams without contrast (single phase scan) for which Dose Length Product is at or below the size-specific exam-specific diagnostic reference level

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).*

Dose length product; CTDIw Phantom Type; Effective Diameter (calculated from localizer image); size specific exam-specific diagnostic reference level.

These components capture how well radiation exposure from the scanner is adjusted for patient size, using size-specific exam-level diagnostic reference levels and how well total radiation exposure to a patient from an exam is optimized based on the CT dose index dose-length product (DLP).

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

Number of CT Abdomen-pelvis exams with contrast (single phase scans), CT Chest exams without contrast (single phase scans), and CT Head/Brain (single phase scans)

Target population: all patients regardless of age.

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).*

Study description; Exam date; Acquisition protocol

Target population: all patients who require either a CT Abdomen-pelvis exam with contrast (single phase scans), a CT Chest exam without contrast (single phase scans), and/or a CT Head/Brain (single phase scans) exam regardless of age.

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

No denominator exclusions

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.)*

No denominator exclusions

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

The measure calculation is stratified by patient size. The results are not reported separately by the stratification variable.

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

Stratification by risk category/subgroup

If other:

S.12. Type of score:

Rate/proportion

If other:

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

Better quality = Higher score

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

Target population is all patients regardless of age.

To calculate the denominator for each of the measures we include all exams that are mapped to a standardized exam name/study description that corresponds to one of the three exam types used for measures, has a localizer image to permit size assessment, and has non-zero values for dose indices.

To calculate the numerator:

Head exams are categorized using lateral thickness (size) from scout images submitted by facilities. Body exams (chest and abdomen/pelvis) are categorized using the effective diameter (size) that ACR calculates from scout

images. The numerator consists of the total number of exams among the denominator that are at or below the size specific DRL.

To calculate the performance rate, the numerator (Total number of exams among the denominator that are at or below the size specific DRL) is divided by the denominator (submitted eligible records) and multiplied by 100 to indicate the percentage. Physician groups/facilities may compare their performance to other facilities using aggregate registry level benchmarks.

Step 1: Denominator: Total number of exams that were mapped to one of the 3 exam names, had a non-zero DLP and a non-zero CTDIvol, CTDIvol<DLP, age was not missing, and patient size is available

Step 2: Numerator: Total number of exams among the denominator that are at or below the size specific DRL

Step 3: Percentage at or below size-specific DRL for each body part: (Numerator/Denominator)*100

Step 4: Percentage of all exams at or below size-specific DRL. Alternately, calculate weighted average of component measures, where weight is number of records for each body part.

Composite score:

Each component measure percentile score is weighted by the denominator count. The weighted scores are summed then divided by the sum of weights of all 3. Alternatively, the numerator and denominator counts for each measure can be totaled then averaged by 3.

Example:

	Numerator	Denominator	Rate
Head	3000	8000	38%
Abdomen/Pelvis	5000	10000	50%
Chest	2000	5000	40%
All	10000	23000	43%
Weighted average			43%

Weighted average = (Weight Head x Rate Head) + (Weight Abdomen/Pelvis x Rate Abdomen/Pelvis) + (Weight Chest x Rate Chest)))/Sum of weights of all 3

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.

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.

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

If other, please describe in S.18.

Registry Data

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

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

Clinical data registry (ACR National Radiology Data Registry - Dose Index Registry)

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

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

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

Clinician : Group/Practice, Facility

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

Emergency Department and Services, Inpatient/Hospital, Other, Outpatient Services

If other: Dialysis Facility

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.)*

This measure will be calculated using the weighted average of three performance rates:

• Rate 1: Percent of CT Abdomen-pelvis exams with contrast (single phase scan) for which Dose Length Product is at or below the size-specific diagnostic reference level

• Rate 2: Percent of CT Chest exams without contrast (single phase scan) for which Dose Length Product is at or below the size-specific diagnostic reference level

• Rate 3: Percent of CT Head/brain exams without contrast (single phase scan) for which Dose Length Product is at or below the size-specific diagnostic reference level

2. Validity – See attached Measure Testing Submission Form

NQF_3621_Composite_Testing_Form.docx

2.1 For maintenance of endorsement

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

2.2 For maintenance of endorsement

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

2.3 For maintenance of endorsement

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

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

Measure Number (*if previously endorsed*)**:**
Composite Measure Title: Composite weighted average of 3 CT Exam Types: Overall Percent of CT exams for which Dose Length Product is at or below the size-specific diagnostic reference level (for CT Abdomen-pelvis with contrast/single phase scan, CT Chest without contrast/single phase scan and CT Head/Brain without contrast/single phase scan)

Date of Submission: 1/1/2021

Composite Construction:

☒Two or more individual performance measure scores combined into one score

☐ All-or-none measures (e.g., all essential care processes received or outcomes experienced by each patient)

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. ***If there are differences by aspect of testing****, (e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.*

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for* ***all*** *the sources of data specified and intended for measure implementation.* ***If different data sources are used for different components in the composite, indicate the component after the checkbox. 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*).

The American College of Radiology (ACR) used data from their National Radiology Data Registry (NRDR) Dose Index Registry (DIR). The primary participants ("target population") are hospital radiology departments (inpatient/outpatient), radiology groups and free-standing imaging centers.

1.3. What are the dates of the data used in testing? January 1, 2017 – December 1, 2020.

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*)

Group and Facility Structure in the ACR National Radiology Data Registry (NRDR) Dose Index Registry (DIR)

Groups are generally aggregations of facilities – a group supports one or more facilities.

GROUP-LEVEL ANALYSIS:

The testing sample comprised all groups that submitted data to ACR NRDR DIR for this measure. The sample consisted of 237 radiology groups. The eligible population for this measure (i.e., the denominator) includes all submitted eligible records (CT Abdomen-pelvis with contrast/single phase scan, CT Chest without contrast/single phase scan and CT Head/Brain without contrast/single phase scan). There are no exclusions with this measure.

Table 1. Number of groups that submitted data for this measure.

Measures	*Composite Weighted Average of all 3*	CT Abdomen-pelvis with contrast/single phase scan	CT Chest without contrast/single phase scan	CT Head/Brain without contrast/single phase scan
All Years	237	229	233	225
2017	69	63	60	61
2018	88	80	84	82
2019	212	195	206	201
2020	212	198	206	197

FACILITY-LEVEL ANALYSIS

The testing sample comprised all facilities that submitted data to ACR NRDR DIR for this measure. The sample consisted of 2,863 hospitals/imaging facilities. The eligible population for this measure (i.e., the denominator) includes all submitted eligible records (CT Abdomen-pelvis with contrast/single phase scan, CT Chest without

contrast/single phase scan and CT Head/Brain without contrast/single phase scan). There are no exclusions with this measure.

Table 2. Number of facilities that submitted data for this measure.

Measures	*Composite Weighted Average of all 3*	CT Abdomen-pelvis with contrast/single phase scan	CT Chest without contrast/single phase scan	CT Head/Brain without contrast/single phase scan
All Years	2,893	2,721	2,782	2,743
2017	2,148	1,916	1,937	1,929
2018	2,390	2,141	2,182	2,132
2019	2,428	2,105	2,220	2,112
2020	2,386	2,090	2,204	2,079

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

GROUP-LEVEL ANALYSIS

A total of 8,714,236 patients were eligible to be included in this testing. *Reported patients* are the number of patients reported to CMS for accountability purposes. Patients included both male and female of all ages with various indications for the exams in each measure exam category. The registry categorizes data by study description and covers any indication that may be associated with the procedure.

Table 3. Eligible patients and reported patients for group-level testing.

	Composite Weighted Average of all 3		CT Abdomen-pelvis without contrast/ single phase scan		CT Chest without contrast/single phase scan		CT Head/Brain without contrast/single phase scan	
Measures	# of Patients Eligible	# of Patients Reported	# of Patients Eligible	# of Patients Reported	# of Patients Eligible	# of Patients Reported	# of Patients Eligible	# of Patients Reported
All Years	8,714,236	8,443,932	2,155,567	2,079,947	1,037,546	1,007,143	2,777,749	2,642,407
2017	1,871,405	1,698,254	664,572	614,113	263,099	240,986	943,521	842,942
2018	1,386,112	1,317,898	487,792	462,631	221,925	213,635	676,234	641,471
2019	2,861,207	2,832,268	514,155	514,155	280,877	280,877	621,021	621,021
2020	2,595,512	2,595,512	489,048	489,048	271,645	271,645	536,973	536,973

FACILITY-LEVEL ANALYSIS

A total of 50,356,186 patients were eligible to be included in this testing. *Reported patients* are the number of patients reported to CMS for accountability purposes. Patients included both male and female of all ages with various indications for the exams in each measure exam category. The registry categorizes data by study description and covers any indication that may be associated with the procedure.

Table 4. Eligible patients and reported patients for facility-level testing.

	Composite Weighted Average of all 3		CT Abdomen-pelvis with contrast/ single phase scan		CT Chest without contrast/single phase scan		CT Head/Brain without contrast/single phase scan	
Measures	# of Patients Eligible	# of Patients Reported	# of Patients Eligible	# of Patients Reported	# of Patients Eligible	# of Patients Reported	# of Patients Eligible	# of Patients Reported
All Years	50,356,186	50,096,936	17,229,385	17,132,224	7,397,579	7,364,059	25,723,543	25,594,982
2017	10,546,008	10,525,572	3,514,958	3,509,303	1,383,514	1,380,782	5,646,213	5,634,170
2018	12,845,656	12,785,646	4,353,122	4,331,936	1,781,313	1,774,451	6,709,691	6,677,730
2019	14,174,013	14,087,765	4,873,574	4,838,446	2,157,910	2,146,430	7,140,965	7,101,326
2020	12,790,509	12,697,953	4,487,731	4,452,539	2,074,842	2,062,396	6,226,674	6,181,756

Additionally, the ACR has provided **facility characteristics** below in **Table 5**.

Table 5. Characteristics of facilities, all years combined.

	Composite Weighted Average of all 3		CT Abdomen-pelvis without contrast/single phase scan		CT Chest without contrast/single phase scan		CT Head/Brain without contrast/single phase scan	
Measures	**# facilities**	**# patients**	**# facilities**	**# patients**	**# facilities**	**# patients**	**# facilities**	**# patients**
Facility category: *Academic*	173	4,014,721	181	609,716	185	863,040	183	2,682,561
Facility category: *Community hospital*	1,277	17,776,843	1,304	3,805,276	1,379	2,576,163	1,379	12,332,303
Facility category: *Multi-specialty clinic*	119	412,793	127	86,026	148	220,337	138	178,477
Facility category: *Freestanding center*	623	1,450,846	664	370,445	717	752,413	687	462,647
Facility category: *Children's hospital*	33	92,927	34	4,173	34	6,601	37	94,450
Facility category: *Other*	108	320,303	113	50,291	117	73,513	117	215,593
Facility location: *Metropolitan*	1,011	13,351,998	1,048	2,491,055	1,090	2,339,031	1,080	9,270,621
Facility location: *Suburban*	837	7,751,000	868	1,701,127	925	1,643,274	909	4,830,736
Facility location: *Rural*	438	2,965,435	461	733,745	513	509,762	500	1,864,674
Census region: *Northeast*	473	5,588,555	501	1,076,655	549	1,303,431	533	3,595,003

Measures	# facilities	# patients	# facilities	# patients	# facilities	# patients	# facilities	# patients
Census region: *Midwest*	537	4,708,684	555	942,294	589	955,688	575	3,113,661
Census region: *South*	890	10,224,294	923	2,182,726	976	1,666,166	972	6,854,523
Census region: *West*	381	3,546,900	393	724,252	410	566,782	404	2,402,844

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

There are no differences in the data or sample used for different aspects of testing for both the group-level and facility-level data. The ACR used the same data for both analyses because the data sample was obtained from the ACR dose registry.

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

No social risk factors are available for this measure. Social risk factors are not relevant for this measure. Patient size is the most important variable for this measure; risk stratification by patient size is provided in this document.

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*)

Note*: Current guidance for composite measure evaluation states that reliability must be demonstrated for the composite performance measure score.*

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

2a2.2. 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*)

ACR performed a signal-to-noise ratio (SNR) analysis test on the performance data for reliability. In SNR analysis, reliability is the measure of confidence in differentiating performance between physicians or other providers; in this case, radiology groups for group-level analysis and facilities for facility-level analysis. The signal is the variability in measured performance that can be explained by real differences in performance and the noise is the total variability in measured performance.

A reliability score equal to zero implies that all the variability in a measure is attributable to measurement error. A reliability score equal to one implies that all the variability is attributable to real differences in physician performance. A reliability score of 0.70 is generally considered the minimum threshold for reliability and 0.80 is generally considered very good reliability.

SNR reliability testing is performed using the Beta-Binomial Model, which assumes that the performance scores are a binomial random variable conditional on the radiology groups' true value derived from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta are considered intermediate calculations used to establish the variance estimates.

ACR testing protocol limited the analysis to physician groups with at least 10 patients reporting for group-level analysis and facilities with at least 10 patients reporting for group-level analysis. Limiting the reliability analysis to groups with a minimum number of events reduces bias introduced by the inclusion of groups without a significant number of events.

Registry data, aggregated by TIN-year, for the component measures and overall composite measure, was used for the relevant group-level information. Registry data, aggregated by year, for the component measures and overall composite measure, were used for the relevant facility-level data.

2a2.3. 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*)

GROUP-LEVEL ANALYSIS

Using the parameter estimates from the beta-binomial model, we computed and aggregated reliability scores for each year. Please see **Table 6** for the results.

Table 6. Reliability score statistics by year by component for group-level testing.

Component 1: CT Average of all 3

Year	*Number of Groups*	*25th percentile*	*Reliability median*	*75th percentile*	*Reliability mean*	*Lower Confidence Limit (minimum)*	*Upper Confidence Limit (maximum)*
2019	207	.99999	1.00000	1.00000	.99998	.99997	.99999
2020	205	.99999	1.00000	1.00000	.99996	.99994	.99999
ALL	412	.99999	1.00000	1.00000	.99997	.99996	.99998

Component 1: CT Abdomen-pelvis

Year	Number of Groups	25th percentile	Reliability median	75th percentile	Reliability mean	Lower Confidence Limit (minimum)	Upper Confidence Limit (maximum)
2017	304	.99997	.99999	1.00000	.99993	.99991	.99995
2018	171	.99996	.99999	1.00000	.99992	.99988	.99995
2019	182	.99998	1.00000	1.00000	.99995	.99993	.99997
2020	183	.99999	1.00000	1.00000	.99997	.99996	.99999
ALL	840	.99998	1.00000	1.00000	.99994	.99993	.99995

Component 2: CT Chest

Year	Number of Groups	25th percentile	Reliability median	75th percentile	Reliability mean	Lower Confidence Limit (minimum)	Upper Confidence Limit (maximum)
2017	297	.99993	.99998	1.00000	.99991	.99989	.99994
2018	175	.99998	.99999	1.00000	.99996	.99995	.99998
2019	199	.99998	.99999	1.00000	.99997	.99997	.99998
2020	197	.99999	.99999	1.00000	.99998	.99997	.99999
ALL	868	.99997	.99999	1.00000	.99995	.99994	.99996

Component 3: CT Head/Brain

Year	Number of Groups	25th percentile	Reliability median	75th percentile	Reliability mean	Lower Confidence Limit (minimum)	Upper Confidence Limit (maximum)
2017	305	.99939	.99990	.99998	.99815	.99759	.99870
2018	171	.99943	.99992	.99999	.99890	.99832	.99948
2019	189	.99914	.99985	.99998	.99782	.99668	.99895
2020	182	.99863	.99977	.99996	.99645	.99489	.99802
ALL	847	.99928	.99987	.99998	.99786	.99738	.99834

FACILITY LEVEL ANALYSIS

Using the parameter estimates from the beta-binomial model, we computed and aggregated reliability scores for each year. Please see **Table 7** for the results.

Table 7. Reliability score statistics by year by component for facility-level testing.

Composite Weighted Average of All 3

Year	Number of Facilities	25th percentile	Reliability median	75th percentile	Reliability mean	Lower Confidence Limit (minimum)	Upper Confidence Limit (maximum)
2017	2150	.99997	.99999	1.00000	.99994	.99994	.99995
2018	2390	.99998	.99999	1.00000	.99996	.99996	.99997
2019	2430	.99997	.99999	1.00000	.99996	.99995	.99996
2020	2386	.99998	.99999	1.00000	.99995	.99995	.99996
ALL	2,893	.99998	.99999	1.00000	.99995	.99995	.99996

Component 1: CT Abdomen-pelvis

Year	Number of Facilities	25th percentile	Reliability median	75th percentile	Reliability mean	Lower Confidence Limit (minimum)	Upper Confidence Limit (maximum)
2017	1920	.99984	.99996	.99999	.99974	.99971	.99977
2018	2146	.99985	.99996	.99999	.99977	.99974	.99979
2019	2116	.99987	.99997	.99999	.99979	.99976	.99981
2020	2099	.99989	.99997	.99999	.99983	.99980	.99985
ALL	2,090	.99986	.99996	.99999	.99978	.99977	.99979

Component 2: CT Chest

Year	Number of Facilities	25th percentile	Reliability median	75th percentile	Reliability mean	Lower Confidence Limit (minimum)	Upper Confidence Limit (maximum)
2017	1939	.99986	.99996	.99999	.99983	.99981	.99984
2018	2184	.99990	.99997	.99999	.99987	.99986	.99988
2019	2224	.99992	.99997	.99999	.99989	.99988	.99990
2020	2205	.99991	.99997	.99999	.99988	.99986	.99989
ALL	2,204	.99990	.99997	.99999	.99987	.99986	.99987

Component 3: CT Head/Brain

Year	Number of Facilities	25th percentile	Reliability median	75th percentile	Reliability mean	Lower Confidence Limit (minimum)	Upper Confidence Limit (maximum)
2017	1953	.99936	.99987	.99996	.99869	.99852	.99886
2018	2162	.99941	.99988	.99996	.99888	.99873	.99902
2019	2147	.99913	.99984	.99995	.99845	.99825	.99864
2020	2122	.99894	.99977	.99993	.99813	.99790	.99837
ALL	2,079	.99922	.99984	.99995	.99853	.99844	.99863

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?*)

A reliability score of 0.7 is considered a reasonable minimum threshold for reliability. Based on the mean reliability scores of 0.9999 for the composite measure at the group-level and the 0.9995 for the facility-level, this measure is considered reliable. The measure is producing consistent and accurate results for each of the component measures, and for the composite measure using the current weighting algorithm. The measures as defined reliably identify variability in performance across providers.

2b1. VALIDITY TESTING

***Note**: Current guidance for composite measure evaluation states that validity should be demonstrated for the composite performance measure score. If not feasible for initial endorsement, acceptable alternatives include*

assessment of content or face validity of the composite OR demonstration of validity for each component. Empirical validity testing of the composite measure score is expected by the time of endorsement maintenance.

2b1.1. What level of validity testing was conducted?

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

☐ **Composite 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.

☐ **Validity testing for component measures** *(check all that apply)*

***Note**: applies to ALL component measures, unless already endorsed or are being submitted for individual endorsement.*

☐ **Endorsed (or submitted) as individual performance measures**

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

☐ **Empirical validity testing of the component measure score(s)**

☒ **Systematic assessment of face validity of component measure score(s) 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*)

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)*

Method A. The ACR convened a formal panel that recently completed a face validity survey for this measure that evaluated the composite measure components at **both the group-level and facility-level**. The expert panel members consisted of medical physicists, radiologists, a value-based purchasing surveyor, and a patient. These experts were contacted based on their expertise in dose optimization, measurement science, and/or radiological practice.

- Missy Danforth (Value-based Purchasing Surveyor) Washington, DC
- Demetrios Giannikopoulos (Patient) Ellicott City, MD
- Chad Dillon (Medical Physicist) Hoover, AL
- Kyle Jones (Medical Physicist) Houston, TX
- Alexander Towbin, MD (Physician) Cincinnati, OH
- David Jordan (Medical Physicist) Cleveland, OH
- Doug Kitchin, MD (Physicist) Middleton, WI
- Olga Brook, MD (Physician) Boston, MA
- Kimberly Applegate, MD (Physician) Zionsville, IN
- Randell Kruger, PhD (Medical Physicist) Marshfield, WI
- Beth Schueler, MD (Physician) Rochester, MN
- Loretta Johnson (Medical Physicist) Birmingham, AL
- Nadja Kadom, MD (Physician) Atlanta, GA
- Donald Frush, MD (Physician) Durham, NC
- William Breeden (Medical Physicist) Indianapolis, IN
- James Tomlinson (Medical Physicist) Ann Arbor, MI
- Tyler Fisher (Medical Physicist) Signal Hill, CA
- Clinton Jokerst, MD (Physician) Scottsdale, AZ
- Tony Seibert, MD (Physician) Sacramento, CA
- Eric Rubin, MD (Physician) Upland, CA

- David Seidenwurm, MD (Physician) Sacramento, CA

The panel was provided a survey using Survey Monkey. The survey began with the following explanation of the survey:

*The purpose of the following survey is to assess whether subject matter experts (you), think that a **particular measure and its components** of accountability accomplishes its intended purpose. The accountability measure is one that compares site radiation dose indices from clinical CT exams to national benchmarks. The ACR is asking for your expert opinion to include anonymized response data in an application to the National Quality Forum (NQF) for measure endorsement.*

The ACR recognizes that monitoring radiation dose indices are only one element of a quality assurance program and that ongoing assessment of exam quality is equally important. While the radiology community continues to develop evidence and consensus for assessing exam quality, we are able to compare clinical radiation dose indices to national benchmarks through the Dose Index Registry (DIR). Since 2014 the ACR has had a CMS-approved accountability measure in its Qualified Clinical Data Registry (QCDR), which assists physicians and practices in reporting their Merit-based Incentive Payment System (MIPS) performance data to CMS.

The DIR accountability measure uses clinical Dose Length Product (DLP) for abdomen-pelvis (with IV contrast, single phase), chest (without IV contrast, single phase), and head/brain (without IV contrast, single phase) CT exams submitted to the DIR; these exams were chosen because they are performed in very high volumes and their structured report outputs are simplest to handle. The measure compares a site's clinical data to national benchmarks that are size-specific and developed using the ICRP 135 methodology. The following survey questions are simply asking whether the accountability measure accomplishes its focused intended purpose.

The panel was asked three questions for **each of the composite measure components**:

1. Do you think that monitoring radiation dose indices from clinical CT exams is a good and worthwhile activity for advancing or maintaining safety and quality?
2. Is this measure and its components as described a reasonable and appropriate way to assess performance quality of a facility or practice with regards to dose optimization?
3. Will the scores obtained from the measure and its components as specified reasonably differentiate clinical performance across providers, and separate the high performers from the low performers?

Method B. Additionally, we have also provided evidence of face validity using consensus documents from a wide range of professional, advisory and regulatory organizations that have endorsed the importance and use of Diagnostic Reference Levels as tool for optimization of patient imaging and to categorize and assess quality in **Table 10**.

Method C. The use of this measure from 2017 to 2020 increased by 207%, indicating that measure users feel the measure is a valid assessment of quality in their practice. We used a percent difference formula to analyze the percent increase. The number of groups reflected in the table below are unique groups (TINs).

Table 8. Total number of unique groups (TINs) that reported the measure.

Measures	*Total Unique Number of Groups Reporting Measure*
Percent Change	207%
2017	69
2018	88
2019	212
2020	212

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

Method A. Table 9 describes the overall results of the face validity survey. Discussion and comments are provided after **Table 9.**

Table 9. Results from face validity survey on the measure and its components.

Question	Percent in Agreement
Do you think that monitoring radiation dose indices from clinical CT exams is a good and worthwhile activity for advancing or maintaining safety and quality?	95%
Is the measure and its components as described a reasonable and appropriate way to assess performance quality of a facility or practice with regards to dose optimization?	71%
Will the scores obtained from the measure and its components as specified reasonably differentiate clinical performance across providers, and separate the high performers from the low performers?	62%
Percent overall	**76%**

1. The survey asked the following question: Do you think that monitoring radiation dose indices from clinical CT exams is a good and worthwhile activity for advancing or maintaining safety and quality?

   95% of the panel (20 members) **agreed** that monitoring radiation dose indices from clinical CT exams is a good and worthwhile activity for advancing or maintaining safety and quality. Some of the feedback included:

   - radiation dose indices are particularly important when associated with adequate image quality.
   - radiation dose indices are a well-recognized method to compare site CT patient dose indices to a national benchmark.
   - patient size must be considered.

The ACR strongly agrees with the panel comments. This measure does take patient size into consideration and is one of the data elements collected automatically during data transmission to the registry. The one panel member (5%) that did not agree with this statement did not leave a comment for their response.

2. The survey asked the following question: Is this measure as described a reasonable and appropriate way to assess performance quality of a facility or practice with regards to dose optimization?

    - 71% of the panel (15 members) **agreed** that the measure components as described is a reasonable and appropriate way to assess performance quality of a facility or practice with regards to dose optimization.

    - 29% (6 members) while not specifically stating that the measure components were not reasonable or appropriate, did not agree that measure components are the *best* way to assess performance quality. Some feedback included a desire for CTDI or size-specific dose estimate (SSDE) metric, since scan lengths can vary, as well as a suggestion to ensure image quality is sufficient to make an accurate diagnosis.

    The ACR previously included metrics assessing dose indices of DLP, CTDIvol and SSDE in the ACR Dose Index Registry (DIR) for accountability purposes when reporting to CMS payment programs/MIPS. However, because of CMS' goal of measure parsimony, these metrics were removed and replaced by the composite measure NQF #3621. The ACR DIR continues to include these metrics for quality improvement purposes.

    We believe the panel's feedback is reflective of the general desire for advances in quantitative image quality assessment, which is beyond the scope of this measure. Image quality is vital for an accurate diagnosis. **Unfortunately, there are no standards for quantifying image quality at this time.** Measuring size-specific exam level DLPs accommodates this concern to some extent as it allows for dose index differentials to obtain diagnostic quality images across patients of different sizes. **We do not disagree that SSDE is a better metric of size-specific dose; however, most scanners do not report SSDE and all scanners report DLP.** Our use of diagnostic reference levels based on size-specific DLPs is a compromise allowing all facilities to be able to use this measure and improve their performance.

3. The last survey question was: Will the scores obtained from the measure as specified reasonably differentiate clinical performance across providers, and separate the high performers from the low performers?

    62% of the panel (13 members) **agreed** that the scores obtained from the measure would differentiate clinical performance across providers. The panelists noted that DRLs was a positive step forward in this effort.

    38% of the panel (8 members) did not agree. Some feedback included:

- the age of the CT scanner as an important variable in quality; older machines will need higher doses to obtain the same image quality, therefore low performers may be more related to how the old the equipment is.
    - The ACR thinks this is an important point to stress. Patients want the best quality possible on their exams, and if **providers are subjecting patients to higher doses because of old equipment, it is vital to capture this information.** Therefore, it is important to provide sites with comparative performance feedback, which the ACR DIR does at the scanner level.
- DRLs were not meant to differentiate performance.
    - **The ACR agrees that using DRLs alone is not an appropriate way to calculate performance, as it's not a measure or estimate of actual patient radiation dose, but it is closely related to the doses received by patients.** DLP is a measure of radiation output received and experienced by patients and not simply documentation of whether DLP was recorded. **The ACR also collects patient size information so dose estimates can be adjusted accordingly.** Providing comparative data across exam types to a physician or site will help adjust imaging protocols to obtain diagnostic images using the lowest reasonable radiation dose. This measure collects the CT scanner radiation output specific to a patient and exam and compares the actual dose indices to benchmarks for similarly sized patients and similar exam types.

Method B. Table 10 describes several expert opinions on the use of diagnostic reference levels to assess performance. The literature provided was curated by ACR Senior Advisor for Medical Physics.

A literature search was performed using PubMed and the search terms "diagnostic reference level", "CT" and "DRL" with a time frame between 2010 – 2020. This turned up 1907 studies; 17 studies were eliminated by abstract and 79 studies were kept based on inclusion/exclusion criteria. The literature was refined to the final 13 articles in **Table 10** because they demonstrate broad, national and international, expert consensus support for medical imaging practices to monitor their use of radiation dose in patient imaging and compare their performance to available benchmarks. The organizations whose documents are referenced are the standard bearers in their space and are widely followed. ICRP guidance is followed by EU nations and others around the world for national policymaking and clinical practice guidance; the NCRP is similarly regarded in the US. ACR and AAPM are professional organizations that define standards of care for medical physics and the radiological professions in the US.

Table 10. Consensus group data on use of diagnostic reference levels to assess performance.

Document/Report	Consensus Body/Organization	Number of people/ organizations	Summary
Reference CT Protocols	The Alliance for Quality Computed Tomography	29 individuals, AAPM, ACR, Public Health England, FDA, 6 manufacturers	Practices should use DRLs as guidance when reviewing their own clinical CT performance.
NCRP Report No. 172, *Reference Levels and Achievable Doses in Medical and Dental Imaging: Recommendations for the United States*	National Council on Radiation Protection and Measurements	14 committee members & consultants, 100 council members, 13 members of Board of Directors	DRLs are the first step in the optimization process to manage patient dose commensurate with the medical purpose of the procedure.
Report of AAPM Task Group 232, *Current state of practice regarding digital radiography exposure indicators and deviation indices*	AAPM (9k+ members)	16 Task Group members, approved by AAPM Board	Quality assurance programs should include ongoing monitoring and comparison of clinical dose metrics to published normative data including reference levels and achievable levels.
U.S. FDA, *Medical X-ray Imaging*	U.S. Food & Drug Administration		As part of a quality assurance program emphasizing radiation management, practices should monitor doses to patients and check the facility doses against diagnostic reference levels, where available.
ICRP Publication 135, *Diagnostic reference levels in medical imaging*	International Commission on Radiological Protection		The DRL is an essential tool in the optimization process, especially as dose limits are not relevant in the medical exposure of patients. In surveys performed to acquire dose information for different procedures, it is important to identify radiation doses that are too low as well as too high, as both may have consequences for the patient. The DRL has proven to be an effective tool that aids in optimization of protection in the medical exposure of patients for diagnostic and interventional procedures.
UNSCEAR 2013 Report to the General Assembly,	United Nations Scientific	Committee reps for 27 countries	Comparing clinical performance to DRLs enables practices to understand

Document/Report	Consensus Body/Organization	Number of people/ organizations	Summary
Sources, Effects, and Risks of Ionizing Radiation, Volume II: Scientific Annex B	Committee on the Effects of Atomic Radiation		whether, in routine conditions, the patient dose from a specified procedure is unusually high or low for that procedure.
European Commission, Radiation Protection No. 185, *European guidelines on diagnostic reference levels for paediatric imaging*	European Union	16 contributors, plus Expert Advisory Panel reps from CIRSE, IAEA, ICRP, NCRP, PHE, WHO	All examinations resulting in high collective doses should have DRLs. This can include both the most common low dose examinations and the less common high dose examinations. It is acknowledged that other common very low dose procedures (e.g., dental) should also be optimized. The application of DRLs should be the responsibility of all providers of X-ray imaging. This means that DRLs should also be applied to imaging performed outside the radiology department, including cardiology, orthopedic surgery, gastroenterology, intensive care (line placement), neurology, vascular surgery, etc. Specific considerations may also be appropriate for imaging associated with radiation therapy where the purpose and scope of imaging can be different.
ACR Computed Tomography Quality Control Manual (2017)	ACR	39k members	Facilities should explicitly review dose indices. For the limited set of protocols where reference values are available, clinical dose index values should be compared to the reference values of the ACR CT Accreditation Program, AAPM CT Protocols, or other available reference values for the appropriate protocols.
ACR–AAPM–SPR PRACTICE PARAMETER FOR DIAGNOSTIC REFERENCE LEVELS AND ACHIEVABLE DOSES IN MEDICAL X-RAY IMAGING	ACR, AAPM, SPR	ACR (39k), AAPM (9k), SPR	DRLs are suggested action levels above which a facility should review its methods and determine if acceptable image quality can be achieved at lower doses. DRLs and Ads (achievable doses) are part of the optimization process. It is essential to ensure that image quality appropriate for the diagnostic purpose is achieved when changing

Document/Report	Consensus Body/Organization	Number of people/ organizations	Summary
			patient doses. Optimization must balance image quality and patient dose, i.e., image quality must be maintained at an appropriate level as radiation doses are decreased.
International Atomic Energy Agency	IAEA		Diagnostic reference levels (DRLs) are a practical tool to promote optimization. DRLs are one of the steps in the overall process of optimization. DRLs have proved useful as a tool in support of dose audit and practice review for promoting improvements in patient protection.
IAEA Safety Standards, Specific Safety Guide No. SSG-46, *Radiation Protection and Safety in Medical Uses of Ionizing Radiation*	International Atomic Energy Agency (IAEA), International Labour Office (ILO), Pan American Health Organization (PAHO), World Health Organization (WHO)		DRLs are an important tool and should be used for optimization of protection and safety for diagnostic medical exposure.
Image Wisely, *CT Protocol Design*	Image Wisely – ACR, RSNA, AAPM, ASRT	15 Executive Committee members	Practices should compare CT exposures to Diagnostic Reference Levels (DRLs) to compare their protocols to regional and national values.
Image Gently (instructions)	Image Gently	26 Steering Committee members	Pediatric facilities should establish DRLs and compare their routine clinical dose index data to them. Adult DRLs can be used to establish pediatric DRLs by using physics principles to account for smaller patient size.

Method C. This measure has been in the CMS MIPS program as a QCDR measure since 2019. The components of the composite measure have been in the program since 2017. The overall score is the percent of CT exams for which Dose Length Product (DLP) is at or below the size-specific diagnostic reference level benchmarks (for CT Abdomen-pelvis with contrast/single phase scan, CT Chest without contrast/single phase scan and CT Head/Brain without contrast/single phase scan). CMS recently provided preliminary historical benchmark data for this measure based on

reporting for 2019. The measure average performance rate was 80.3% with a range of performance by decile as shown below.

Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
28.83 - 60.42	60.43 - 73.28	73.29 - 82.24	82.25 - 87.25	87.26 - 89.15	89.16 - 94.27	94.28 - 95.13	>= 95.14

This measure and its components measure the DLP of CT exam for a particular aspect of the body (abdomen/pelvis, chest, and head/brain). It is imperative to measure each body area separately since they all have a different DLP requirement. A weighted average is used for the three different exam types to ensure that physician performance is accurately captured based on volume per exam type/component measure.

The underlying data elements of the measure, the median Dose Length Product was a measure at the TIN and individual levels between 2014 and 2016, preceding and informing the current measure construct.

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?*)

Method A. The face validity survey, while having mixed responses on the use of size-specific dose estimates in the last question, does have at least 70% of the panel agreeing that this measure and its components are an appropriate way to differentiate quality across facilities and across groups in dose optimization.

Method B. The expert opinions in **Table 10** indicate a consensus agreement that the use of diagnostic reference levels is a good indicator of quality and dose optimization across facilities and across groups.

Method C. The acceptance of the measure, as constructed, in the CMS accountability programs following review by CMS and by their contractors, is one evidence of validity. Based on their annual measure assessment and approval, CMS experts and their measure development contractors considered the measure to be an appropriate indicator of quality and able to distinguish between levels of quality at the group-level.

2b2. EXCLUSIONS ANALYSIS

Note: *Applies to the composite performance measure, as well all component measures unless they are already endorsed or are being submitted for individual endorsement.*

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

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

Note: *Applies to all outcome or resource use component measures, unless already endorsed or are being submitted for individual endorsement.*

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? *(check all that apply)*

☐ **Endorsed (or submitted) as individual performance measures**

☐ **No risk adjustment or stratification**

☐ **Statistical risk model with _ risk factors**

☒ **Stratification by 17 (5 for head, 6 each for abdomen-pelvis and chest) clinical risk categories**

☐ **Other,**

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

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.

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?

Criteria for selection of clinical risk factor:

We do not apply any social risk factors. Patient size is the predominant justification for variation on dose indices thus was used as the single risk factor in stratification. Radiation dose must increase incrementally with patient size to maintain image quality sufficient to characterize findings. Protocols for imaging are built around size categories, such as large adult, child, etc. Technologists select scanner protocols based on their assessment of patient size in the context of these categories. Size-based diagnostic reference levels (DRLs)* allow facilities to optimize protocols so that the resultant dose is commensurate with the size of the patient, thus avoiding unnecessary radiation exposure to the patient.

Clinical indication for the CT exam may be another justification for variation in dose, however the component measures are defined for a narrowly defined set of procedures (abdomen/pelvis, chest, head) where there should not be much variation driven by indication. These categories of exams are "routine" thus the reasons, or indications, for the exam does not require a special consideration for radiation dose with regard to image quality.

Statistical method for stratification by patient size:

The measure stratifies patients by size into 17 size bands (6 for abdomen/pelvis, 6 for chest, 5 for head). Patient size is expressed in terms of effective diameter for abdomen/pelvis and chest exams, and lateral thickness for head exams. Radiation exposure per exam, for the dose index Dose Length Product (DLP)*, is compared to size-specific DRLs.**

Definitions:

* Dose Length Product (DLP) is a standardized parameter to measure computed tomography (CT) scanner radiation output to a patient and is a useful index to compare protocol-based outputs across different practices and scanners. Please see 1c.2. in the Measure Information Form for more description of DLP.

** Diagnostic reference levels (DRLs) are used as benchmarks for radiation protection and optimization of patient imaging. The intended use of DRLs is as a simple test for identifying situations where the levels of patient dose are unusually high and provide a means for facilities and clinicians to optimize dose to a lower level than a DRL.

Reference: Kanal KM, Butler PF, Sengupta D, Bhargavan-Chatfield M, Coombs LP, Morin RL. U.S. Diagnostic Reference Levels and Achievable Doses for 10 Adult CT Examinations. Radiology. 2017 Jul;284(1):120-133. doi: 10.1148/radiol.2017161911. Epub 2017 Feb 21. PMID: 28221093.

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)**

The concepts used to develop diagnostic reference levels (benchmarks for the composite measure and its components) used patient size as a justification for variation in dose indices.

As part of an ACR-published paper that developed the size-specific diagnostic reference levels underlying the construct of the measure, multivariable mixed regression analysis was conducted to determine whether dose indexes varied significantly by water-equivalent diameter and lateral thickness. Facility was included as a random effect, and fixed effects included facility characteristics, age, and sex. An analysis was performed for multiple comparisons among size bins for each body part to determine if the means of the dose indexes were significantly different from each other.

Size bins were constructed not by statistical significance but by using the distribution of the data—that is, the number of data points in each of the bins—and by keeping the clinical perspective and practical usefulness in mind. We considered collapsing the non–statistically significant bins into one but realized that the resulting bins would be confusing and lose their usability. All analyses were performed by using SAS software, version 9.3, of the SAS System for Windows (2015, SAS Institute, Chicago, Ill). (Kanal et. al, 2017)

The size categories were identified primarily using clinical concepts. However, the final size bins were constructed from a clinical perspective and for practical usefulness. Size bins were 2cm wide for head exams, and 4cm wide for abdomen-pelvis and chest, based on the number of data points in each bin.

Reference: Kanal KM, Butler PF, Sengupta D, Bhargavan-Chatfield M, Coombs LP, Morin RL. U.S. Diagnostic Reference Levels and Achievable Doses for 10 Adult CT Examinations. Radiology. 2017 Jul;284(1):120-133. doi: 10.1148/radiol.2017161911. Epub 2017 Feb 21. PMID: 28221093.

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

Regression analysis conducted as part of the underlying paper showed that size was a significant predictor of dose indices (after controlling for facility as random effect and facility characteristics, age, and sex as fixed effects). The final size bins were constructed from a clinical perspective and for practical usefulness. Size bins were 2cm wide for

head exams, and 4cm wide for abdomen-pelvis and chest, to allow for similar order of magnitude of number of data points in each bin.

Table 11. Comparison of DLP by patient size bins

CT Abdomen-pelvis With Contrast

Measures	N	Mean	StdErr	Median	P25	P75
<25	1071843	412.62	0.38	313.37	213.80	491.27
25 to <29	1707749	472.08	0.25	405.62	288.55	574.52
29 to <33	2278700	657.56	0.27	577.59	422.43	811.80
33 to <37	1806768	894.67	0.37	810.94	591.50	1085.23
37 to <41	1038415	1100.37	0.59	1003.79	739.28	1324.34
41+	1093874	1404.32	0.70	1263.19	949.29	1694.82

CT Chest Without Contrast

Measures	N	Mean	StdErr	Median	P25	P75
<25	209331	227.31	0.59	164.19	97.48	256.96
25 to <29	632346	239.61	0.28	193.30	123.11	284.67
29 to <33	1117103	300.59	0.24	253.95	147.09	388.80
33 to <37	984000	397.20	0.31	356.78	189.98	522.19
37 to <41	512215	519.07	0.55	470.72	268.58	672.75
41+	413959	644.99	0.77	561.93	358.35	790.04

CT Head Without Contrast

Measures	N	Mean	StdErr	Median	P25	P75
<14	4210321	849.18	0.28	781.89	568.85	991.17
14 to <16	4257564	988.31	0.28	893.80	725.67	1080.20

Measures	N	Mean	StdErr	Median	P25	P75
16 to <18	2557221	985.34	0.38	878.39	702.71	1077.66
18 to <20	1658580	956.70	0.42	879.88	710.00	1058.29
>20	979341	1163.53	1.07	919.43	716.63	1153.51

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.

The underlying variable, Dose Length Product, demonstrates variation by patient size categories, as shown in **Table 11** above. Statistical significance of the differences in mean DLP across size bins was conducted using ANOVA and the differences are found to all be statistically significant ($p<0.001$).

This warrants employing stratification by patient size in comparing performance in the measure components and overall composite measure.

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*)

Risk stratification for size was conducted by applying a size-specific exam-specific DRL to each of the identified size categories and assessing if each record met the benchmark specific to its size bin and description. In the absence of size-stratification, each record would be compared to an overall benchmark that spans all sizes.

We examined the effect of stratification by comparing performance (mean, std err, median) where 1) exams were stratified by size band and, 2) where exams were un-stratified by size. We performed this comparison for each component measure and for the overall composite measure.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b3.9

2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
2b3.9. Results of Risk Stratification Analysis:

Facility performance rates were compared for the stratified and un-stratified measures, by patient size. The un-stratified measure disregarded patient size and did not make any allowances for higher dose for large patients or require smaller doses for small patients.

Descriptive statistics and correlations by size are shown in **Table 12,** below. The stratified and un-stratified measures are highly correlated, in the middle size categories and less corelated in the smallest and largest size categories. **The un-stratified measure overestimates performance at smaller sizes and underestimates performance at larger sizes** relative to the stratified measure. For example, for abdomen/pelvis the un-stratified median score for size bin <25cm

is 98.80, compared to the stratified median score for size bin < 83.20. This indicates that for a smaller size of <25cm, when applying the unstratified benchmark, a facility may appear to meet performance even though the dose levels may be higher than required for a patient of that size; the unstratified measure is too lenient. In contrast, for the largest patients (>41cm), the median unstratified performance is 36.4% compared to 72.8% for the stratified measure, suggesting that the unstratified benchmark may be too strict.

Table 12. Descriptive statistics and correlations by size.

Abdomen-pelvis	Stratified measure				Unstratified measure				Correlation coefficient
Effective diameter	**N**	**Mean**	**StdErr**	**Median**	**N**	**Mean**	**StdErr**	**Median**	**Correlation coefficient**
<25	5,786	66.06	0.48	83.20	5,786	95.47	0.16	98.80	0.3656
25 to <29	6,382	64.72	0.44	78.00	6,382	91.54	0.22	98.80	0.5924
29 to <33	6,462	63.23	0.44	78.00	6,462	79.43	0.36	93.60	0.8460
33 to <37	6,383	64.80	0.43	78.00	6,383	60.13	0.44	67.60	0.9742
37 to <41	6,091	62.81	0.44	72.80	6,091	41.10	0.44	36.40	0.8060
41+	5,889	63.97	0.43	72.80	5,889	24.18	0.35	15.60	0.6049

Chest	Stratified measure				Unstratified measure				Correlation coefficient
Effective diameter	**N**	**Mean**	**StdErr**	**Median**	**N**	**Mean**	**StdErr**	**Median**	**Correlation coefficient**
<25	5,577	76.34	0.41	88.40	5,577	91.51	0.24	98.80	0.6151
25 to <29	6,370	74.32	0.39	88.40	6,370	90.74	0.24	98.80	0.6821
29 to <33	6,475	74.12	0.38	88.40	6,475	83.07	0.32	93.60	0.8879
33 to <37	6,373	77.19	0.36	88.40	6,373	70.53	0.39	83.20	0.9304
37 to <41	6,004	78.45	0.35	88.40	6,004	55.84	0.44	62.40	0.7233
41+	5,612	81.16	0.34	93.60	5,612	42.28	0.44	36.40	0.5330

Head, brain	Stratified measure				Unstratified measure				Correlation coefficient
Lat thickness	**N**	**Mean**	**StdErr**	**Median**	**N**	**Mean**	**StdErr**	**Median**	**Correlation coefficient**
<14	5,772	56.06	0.51	67.60	5,772	65.23	0.47	78.00	0.8984
14 to <16	6,371	60.60	0.46	72.80	6,371	66.41	0.44	83.20	0.9435
16 to <18	6,173	63.98	0.46	78.00	6,173	63.21	0.46	78.00	0.9948
18 to <20	4,710	65.20	0.54	83.20	4,710	60.36	0.56	72.80	0.9556
>20	4,427	61.09	0.57	72.80	4,427	49.02	0.59	52.00	0.8562

The risk stratification analysis is only performed at the level of the facility and not group. This is because groups are generally aggregations of facilities – a group supports one or more facilities. Any findings for patient size stratification applicable at the facility level formulation of the measure applies to group level as well.

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*)

The un-stratified measure overestimates performance at smaller sizes and underestimates performance at larger sizes relative to the stratified measure.

Additionally, the results quoted from the paper below show that size is an important risk factor to use for stratification:

Depending on the size of the patient relative to the size of the phantom used to report $CTDI_{vol}$, the actual dose to the patient may be considerably different (**4**,**5**).

However, radiation dose must increase with patient size (**13**) to maintain acceptable image quality.

4. McCollough CH, Leng S, Yu L, Cody DD, Boone JM, McNitt-Gray MF. CT dose index and patient dose: they are not the same thing. ***Radiology*** *2011;259(2):311–316. Link, Google Scholar*

5. Seibert JA, Boone JM, Wootton-Gorges SL, Lamba R. Dose is not always what it seems: where very misleading values can result from volume CT dose index and dose length product. ***J Am Coll Radiol*** *2014;11(3):233–237. Crossref, Medline, Google Scholar*

13. Shrimpton PC, Hiller MC, Meeson S, Golding SJ. Doses from computed tomography (CT) examinations in the UK – 2011 review. Public Health England website. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/349188/PHE_CRCE_013.pdf. Published 2014. Accessed November 4, 2016. Google Scholar

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*)

_____

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

Note: *Applies to the composite performance measure.*

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)*

To assess statistically significant differences in measure rates, the data described in sections above were used to calculate the mean, median, standard deviation, and interquartile range for the measure rates.

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*)

The table below shows the distribution of measure rates for all the component measures and the composite measure between 2017 and 2020. The mean rate for the composite measure was 78.87%, with a median rate of 84.57%, minimum rate of 11.01%, and maximum rate of 100%. The mean rate for CT abdomen-pelvis with contrast was 80.69%, with a median rate of 84.39%, minimum rate of 6.85%, and maximum rate of 100%. The mean rate for CT head/brain without contrast was 75.87%, with a median rate of 88.35%, minimum rate of .38%, and maximum rate of 100%. The mean rate for CT Chest without contrast/single phase scan was 83.96% with a median rate of 88.27%, minimum rate of 17.44%, and maximum rate of 100%.

Table 13. Variation and distribution of measure rates from 2017 to 2020.

Statistic	Composite	CT Abdomen-Pelvis	CT Head/Brain	CT Chest
Mean	78.87%	80.69%	75.87%	83.96%
Standard Deviation	18.17%	16.37%	28.18%	14.58%
Minimum	11.01%	6.85%	0.38%	17.44%
25th percentile	69.90%	72.39%	59.22%	74.91%
50th percentile (median)	84.57%	84.39%	88.35%	88.27%
75th percentile	92.77%	93.25%	98.38%	95.25%
Maximum	100%	100%	100%	100%
Interquartile Range	22.87%	20.86%	39.16%	20.34%
Student's t-test p-value	P<.0001	P<.0001	P<.0001	P<.0001

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?*)

The measure rates of the component measures and the composite measure show significant variation, with an interquartile range of 22.87%. There is a statically significant different in the measure rates between the top and bottom quartile of the testing (P<0.0001 at alpha = 0.05). This variation shows that there are statistically significant and clinically meaningful differences in performance. There is potential for improvement with this measure.

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

Note: *Applies to all component measures, unless already endorsed or are being submitted for individual endorsement.*

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*)

2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b5.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted?*)

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

Note: *Applies to the overall composite measure.*

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*)

The ACR NRDR DIR participants use software that directly transmits all CT scanner data in Digital Imaging and Communications in Medicine (DICOM) format to the registry. ACR encourages all registry users to review submitted data at least every two months to ensure data are flowing accurately and completely to the DIR. Automated e-mail reminders are sent for data quality checks. Users are able to view several reports to ensure data are not missing, such as a data summary report and a scanner report, to compare the volume of exams received with the volume of exams sent. Sites can identify which scanner may not be transmitting the correct volume of exams.

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*)

There are no missing data for this measure. The data are generated and transmitted automatically from each scanner. By eliminating manual data entry, the registry reduces errors and resource burden for each group/facility.

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*)

No missing data was found through testing, nor would missing data be expected to occur in the future. The automation of all data collection for each facility ensures accurate and unbiased results.

2c. EMPIRICAL ANALYSIS TO SUPPORT COMPOSITE CONSTRUCTION APPROACH

***Note:** If empirical analyses do not provide adequate results—or are not conducted—justification must be provided and accepted in order to meet the must-pass criterion of Scientific Acceptability of Measure Properties. Each of the following questions has instructions if there is no empirical analysis.*

2d1. Empirical analysis demonstrating that the component measures fit the quality construct, add value to the overall composite, and achieve the object of parsimony to the extent possible.

2d1.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used; **if no empirical analysis**, provide justification)*

The ACR performed pairwise correlation between the performance rates of the component measures to ensure that the composite is not redundant.

We assessed proportion of total practice covered by the three complementary component measures to demonstrate the clinical validity of combining the measures into a composite.

2d1.2. What were the statistical results obtained from the analysis of the components? (e.g., *correlations, contribution of each component to the composite score, etc.*; ***if no empirical analysis****, identify the components that were considered and the pros and cons of each*)

The results of the analysis are in the table below.

Table 14. Component pairwise correlation analysis.

Measures	*Abdomen-pelvis TO Chest*	*Abdomen-pelvis TO Head-Brain*	*Chest TO Head-Brain*
2017	*-0.09*	*-0.10*	*0.17*
2018	*-0.17*	*0.02*	*-0.03*
2019	*-0.15*	*-0.04*	*0.04*
2020	*0.10*	*0.04*	*0.12*

The selected component measures collectively represent a substantial proportion of the work performed by most groups, but the distribution between the components may vary. We will submit analyses to demonstrate this with our final submission.

2d1.3. What is your interpretation of the results in terms of demonstrating that the components included in the composite are consistent with the described quality construct and add value to the overall composite? (i.*e., what do the results mean in terms of supporting inclusion of the components;* ***if no empirical analysis****, provide rationale for the components that were selected)*

The correlation between the performance rates of the component measures is very low. This indicated that the composite measure is not redundant.

The measures are identical in construct, so they do reflect the same underlying quality. The data show that together, the exams covered reflect NN% of a group's practice. Therefore, looking at the 3 components together in a composite allows for a parsimonious assessment of the overall group performance on this measure, and the component measures provide guidance of where the group's performance gaps may lie.

2d2. Empirical analysis demonstrating that the aggregations and weighting rules are consistent with the quality construct and achieve the objective of simplicity to the extent possible

2d2.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used;* ***if no empirical analysis****, provide justification)*

We compared a simple average composite with a weighted average composite to assess the difference in measured performance.

2d2.2. What were the statistical results obtained from the analysis of the aggregation and weighting rules? (e.g., *results of sensitivity analysis of effect of different aggregations and/or weighting rules;* ***if no empirical analysis****, identify the aggregation and weighting rules that were considered and the pros and cons of each*)

Table 14. Comparison of simple average vs. weighted average

Measures	*% Measure Met (Mean)*	*Standard Deviation*	*Min*	*25th percentile*	*Median*	*75th percentile*	*Max*
Weighted average (current measure)	79.62	18.24	13.57	70.88	85.74	93.10	100
Simple average	80.33	15.28	18.48	70.69	84.08	92.47	100

2d2.3. What is your interpretation of the results in terms of demonstrating the aggregation and weighting rules are consistent with the described quality construct? (i.*e., what do the results mean in terms of supporting the selected rules for aggregation and weighting;* ***if no empirical analysis****, provide rationale for the selected rules for aggregation and weighting*)

The weighting mechanisms generate very similar results. The weighted average method is more reflective of a group's practice. It is not especially difficult to calculate. Therefore, we believe the weighted average is the best approach for these component measure.

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

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

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 a combination of electronic sources

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: DIR_NQF_Feasibility_Scorecard.xlsx

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.

The initial setup for submitting data requires the site to have staff resources for installing data collection software. It is a small amount of time to set up the CT equipment to transmit the dose information and to map the site exam names to standardized DIR names for comparison. Occasionally, if done incorrectly, this can require a site to review the set-up and standardized formatting.

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)*.

Minimal participation fee to participate in the DIR, which is based on facility size, number of facilities and number of radiologists in each practice. The fee is typically about $500-$1000 per year. The primary purpose of participating sites in DIR is quality improvement, but an additional benefit of this specific measure is the accountability purpose.

NRDR and Participation Fees: https://nrdrsupport.acr.org/support/solutions/articles/11000029012-registration-and-participation-fees

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 Merit-based Incentive Payment System qpp.cms.gov Quality Improvement (Internal to the specific organization) ACR Dose Index Registry https://www.acr.org/Practice-Management-Quality-Informatics/Registries/Dose-Index-Registry

*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

The CMS Merit-based Incentive Payment System (MIPS) is a quality improvement accountability program. They reward high-value, high-quality Medicare clinicians with payment increases and reduce payments to those

clinicians who aren't meeting performance standards. Over 10,000 physicians and approximately 2.4 million patients are included in the program for this measure. A variety of geographic areas in the United States are measured. Measurement is performed at the individual and group level.
The ACR Dose Index Registry (DIR) allows facilities to compare their CT dose indices to regional and national values. Facilities receive quarterly feedback reports comparing their results to aggregate results by body part and exam type. Participation offers participants additional ways to fulfill reporting requirements for the Merit-based Incentive Payment System (MIPS) and also allows credit for Maintenance of Certification (MOC) Part IV requirements of the American Board of Radiology (ABR). Over 2000 facilities and over 200 groups submit data to the DIR, with over 2 million patients included. Measurement is performed at the facility and group level.

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?*)
This measure is currently used in an accountability program.

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.*)

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.

The measure specifications are updated annually and included in the CMS Quality Payment Program for MIPS. This measure is reported via the ACR National Data Radiology Database (NRDR) Qualified Clinical Data Registry (QCDR) with measure ID ACRad34. Detailed specifications are publicly available on the ACR website.

Assistance with interpretation for this measure is provided through the ACR help desk and through the CMS help desk. Users can submit their questions and receive a response from ACR staff within 72 hours.

Performance results are provided in two ways. The first is through the ACR NRDR DIR, where users upload their data to the registry and can compare their performance against registry benchmarks in real time. Users must have an account with the registry to view results and are able to view their performance online. The second is through CMS' MIPS Feedback Reports, which are issued annually. These feedback reports are based on performance benchmarks, which are calculated in deciles. These reports are not specific nor necessarily indicative of a group's performance. These reports are available online through the user's CMS account.

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.

Feedback is provided to all DIR participants reporting this quality measure daily. Feedback is based on registry benchmarks. ACR educational webinars are conducted bimonthly to explain measure requirements and interpretation of performance results.

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.

Feedback is obtained through email, the ACR help desk, the CMS quality help desk, and CMS contractor QMMS.

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

Feedback on this measure is positive. Facilities are able to evaluate when their CT exam protocols should be reviewed and/or updated to optimize radiation dose exposure to patients.

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

No other feedback has been provided from entities other than individuals that could report the measure.

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.

This feedback is considered during the annual measure specification update process with CMS. The ACR Metrics Committee reviews feedback for measure changes.

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.

Performance has remained steady in the 79-80% for this measure. There hasn't been a significant performance improvement, which demonstrates that there is still a gap in care for optimizing radiation dose to patients. Improving performance in this measure would demonstrate that a facility is adjusting radiation dose protocols.

4b2. Unintended Consequences

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

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

We are not aware of any unintended consequences related to this measurement.

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

All benefits from this measure are intended.

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.

Yes

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

2820 : Pediatric Computed Tomography (CT) Radiation Dose

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

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

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

Are the measure specifications harmonized to the extent possible?

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

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.)

Our measure, NQF #3621, evaluates the whole population and is not limited to pediatric patients as for NQF #2820. In NQF #3621 performance for facilities and groups is calculated comparing dose indices to published benchmarks.

NQF #2820, "provides a simple framework for how facilities can assess their dose, compare their doses to published benchmarks (Smith-Bindman, Radiology, 2015) and identify opportunities to improve if their doses are higher than the benchmarks". Measure users thus are self-calculating results against one of three published benchmarks themselves using one of three benchmarks published benchmarks for both levels of measurement (group and facility).

NQF #3621 uses data published in the ACR 2017 study, U.S. Diagnostic Reference Levels and Achievable Doses for 10 Adult CT Examinations, identifying DRLs and Achievable Doses (ADs) for the 10 most common CT adult examinations performed in the United States. It represents the first time that national adult DRLs and ADs have been developed as a function of patient size, a milestone in optimizing radiation dose to patients. NQF #3621 has eight years of performance data for each measure component, as well as four years of data for the composite. Using electronic data sources, NQF #3621 has high feasibility and low collection burden, which minimizes missing data bias. NQF #3621 provides greater consistency and level of comparison across facilities

and groups, providing more validity and reliability for use in quality improvement and specifically for accountability programs.
Reference: Kanal KM, Butler PF, Sengupta D, Bhargavan-Chatfield M, Coombs LP, Morin RL. U.S. Diagnostic Reference Levels and Achievable Doses for 10 Adult CT Examinations. Radiology. 2017 Jul;284(1):120-133. doi: 10.1148/radiol.2017161911. Epub 2017 Feb 21. PMID: 28221093.

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 College of Radiology

Co.2 Point of Contact: Karen, Campos, kcampos@acr.org, 800-227-5463-5848

Co.3 Measure Developer if different from Measure Steward: American College of Radiology

Co.4 Point of Contact: Karen, Campos, kcampos@acr.org, 800---

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.

- William Sensakovic, PhD
- Steven Don, MD
- Loretta Johnson, PhD
- Clinton Jokerst, MD
- Aaron Jones, PhD
- Phillip Koo, MD
- Tony Seibert, MD, FACR
- Keith Strauss, MS, FACR
- Kalpana Kanal, PhD, FACR
- Mythreyi Chatfield, PhD

- Dustin Gress
- Penny Butler
- Judy Burleson, MHSA
-

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2019

Ad.3 Month and Year of most recent revision: 09, 2020

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

Ad.5 When is the next scheduled review/update for this measure? 09, 2021

Ad.6 Copyright statement: n/a

Ad.7 Disclaimers:

Ad.8 Additional Information/Comments: