Measure Worksheet

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.
To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 3663e
Corresponding Measures:
Measure Title: Excessive Radiation Dose or Inadequate Image Quality for Diagnostic Computed Tomography (CT) in Adults (Facility Level)
Measure Steward: Alara Imaging
sp.02. Brief Description of Measure: This electronic clinical quality measure (eCQM) provides a standardized method for monitoring the performance of diagnostic CT to discourage unnecessarily high radiation doses, a risk factor for cancer, while preserving image quality. It is expressed as a percentage of eligible CT exams that are out-of-range based on having either excessive radiation dose or inadequate image quality, relative to evidence-based thresholds based on the clinical indication for the exam. All diagnostic CT exams of specified anatomic sites performed in inpatient and hospital outpatient care settings are eligible.
1b.01. Developer Rationale:
Diagnostic CT imaging occurs in more than a third of acute care hospitalizations (Vance 2013) and upwards of 90 million scans are performed annually in the U.S. (IMV 2020). The radiation doses used for these exams are frequently far higher than needed for diagnosis and vary up to 200-fold across facilities for patients imaged for the same clinical reason. (Smith-Bindman 2009, Smith-Bindman 2015, Smith-Bindman 2019, Miglioretti 2013, Demb 2017). Most of this variation reflects clinician preferences rather than appropriate differences based on patient and clinical indications (Smith-Bindman 2019). As described in section 1a.14, the inconsistency in how CT exams are performed represents a significant, unnecessary, and modifiable iatrogenic health risk, as there is extensive epidemiological and biological evidence that suggests exposure to radiation in the same range as that routinely delivered by CT increases a person's risk of developing cancer (Board of Radiation Effects 2006, Pearce 2012, Pierce 2000, Preston 2007, Brenner 2003, Hong 2019). It is estimated that 2% (36,000) of the 1.8 million cancers diagnosed annually in the U.S. are caused by CT exams (Berrington de Gonzalez 2009, NCI Cancer Statistics).
The measure focuses on reducing radiation dose in CT, an intermediate outcome important to cancer prevention. As radiation dose is known to be directly related and proportional to future cancer risk (Board of Radiation Effects 2006, Pearce 2012, Pierce 2000, Preston 2007, Brenner 2003, Hong 2019, Berrington de Gonzalez 2009), any reduction in radiation exposure would be expected to lead to a proportional reduction in cancers. Research suggests that when healthcare organizations and clinicians are provided with a summary of their CT radiation doses, their subsequent doses can be reduced without diminishing the diagnostic usefulness of these tests. Smith-Bindman et al. led a randomized controlled trial of two interventions to optimize CT radiation doses across 100 hospitals and imaging facilities and found that providing feedback to institutions along with education and opportunities for sharing best practices results in meaningful dose reductions. (Smith-Bindman 2020). Though results varied by anatomic region, following the intervention there was up to a 40% reduction in doses with a greater impact on the rate of high dose exams, meaning facilities with high

doses at the beginning of the trial were particularly likely to improve.

On the basis of the current estimated number of CT exams performed annually in the U.S. (IMV 2020), distribution in scan types and observed doses (Demb 2017, Smith-Bindman 2019), modelling of the cancer risk associated with CT at different ages of exposure (Berrington de Gonzalez 2009), and costs of cancer care (Dieguez 2017, Mariotto 2011), an estimated 18,643 cancers could be prevented annually in the U.S., 75% (13,982) of these among Medicare beneficiaries, resulting in $1.86 billion to $5.21 billion in annual cost savings to the Centers for Medicare & Medicaid Services.

References

1. Vance EA, Xie X, Henry A, Wernz C, Slonim AD. Computed tomography scan use variation: patient, hospital, and geographic factors. Am J Manag Care. 2013 Mar 1;19(3):e93-9. PMID: 23534948.
2. IMV 2019 CT Market Outlook Report, https://imvinfo.com/ct-departments-seek-workflow-improvements-to-address-increased-ct-utilization/.
3. Smith-Bindman R, Lipson J, Marcus R, et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. Arch Intern Med. 2009;169(22):2078-2086.
4. Smith-Bindman R, Moghadassi M, Wilson N, et al. Radiation Doses in Consecutive CT Examinations from Five University of California Medical Centers. Radiology. 2015;277(1):134-141.
5. Smith-Bindman R, Wang Y, Chu P, et al. International variation in radiation dose for computed tomography examinations: prospective cohort study. BMJ. 2019;364:k4931.
6. Miglioretti DL, Johnson E, Williams A, et al. The use of computed tomography in pediatrics and the associated radiation exposure and estimated cancer risk. JAMA Pediatr. 2013;167(8):700-707.
7. Demb J, Chu P, Nelson T, et al. Optimizing Radiation Doses for Computed Tomography Across Institutions: Dose Auditing and Best Practices. JAMA Intern Med. 2017;177(6):810-81
8. Board of Radiation Effects Research Division on Earth and Life Sciences National Research Council of the National Academies. Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2, Washington, D.C.: The National Academies Press; 2006.
9. Pearce MS, Salotti JA, Little MP, et al. Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. Lancet. 2012;380(9840):499-505.
10. Pierce DA, Preston DL. Radiation-related cancer risks at low doses among atomic bomb survivors. Radiation research. 2000;154(2):178-186.
11. Preston DL, Ron E, Tokuoka S, et al. Solid cancer incidence in atomic bomb survivors: 1958-1998. Radiation research. 2007;168(1):1-64.
12. Brenner DJ, Doll R, Goodhead DT, et al. Cancer risks attributable to low doses of ionizing radiation: assessing what we really know. Proc Natl Acad Sci U S A. 2003;100(24):13761-13766.
13. Hong JY, Han K, Jung JH, Kim JS. Association of Exposure to Diagnostic Low-Dose Ionizing Radiation with Risk of Cancer Among Youths in South Korea. JAMA Netw Open. 2019;2(9):e1910584.
14. Berrington de Gonzalez A, Mahesh M, Kim KP, et al. Projected cancer risks from computed tomographic scans performed in the United States in 2007. Arch Intern Med. 2009;169(22):2071-2077.
15. National Cancer Institute Cancer Statistics. Accessed May 25, 2021. https://www.cancer.gov/about-cancer/understanding/statistics
16. Smith-Bindman R, Chu P, Wang Y, et al. Comparison of the Effectiveness of Single-Component and Multicomponent Interventions for Reducing Radiation Doses in Patients Undergoing Computed Tomography: A Randomized Clinical Trial. JAMA Intern Med. 2020 May 1;180(5):666-675.
17. Dieguez G, Ferro C, Pyenson B. Milliman Research Report: A Multi-Year Look at the Cost Burden of Cancer Care. April 11, 2017. https://www.milliman.com/en/insight/2017/a-multi-year-look-at-the-cost-burden-of-cancer-care
18. Mariotto AB, Yabroff KR, Shao Y, Feuer EJ, Brown ML. Projections of the cost of cancer care in the United States: 2010-2020. J Natl Cancer Inst. 2011 Jan 19;103(2):117-28. Epub 2011 Jan 12. Erratum in: J Natl Cancer Inst. 2011 Apr 20;103(8):699. PMID: 21228314.

sp.12. Numerator Statement: Diagnostic CT exams that have a size-adjusted radiation dose value greater than the threshold specific to the CT category (reflecting the body region imaged and the radiation dose and image quality required for that exam given the reason for the exam), or a global noise value greater than a threshold specific to the CT Category.
sp.14. Denominator Statement: All diagnostic CT exams performed on adults (aged 18 years and older) during the measurement period of one year that have an assigned CT category, a size-adjusted radiation dose value, and a global noise value.
sp.16. Denominator Exclusions: Denominator exclusions are CT exams that simultaneously include multiple body regions outside of four commonly encountered multiple region groupings (specified as LOINC code 96914-7, CT Dose and Image Quality Category, Full Body). Denominator exclusions are also CT exams with missing patient age, missing size-adjusted radiation dose, or missing global noise. These are technical exclusions ("missing data") from the initial population. Technical exclusions will be flagged, corrected whenever possible, and tracked at the level of the accountable entity.

Measure Type: Outcome: Intermediate Clinical Outcome
sp.28. Data Source: Electronic Health Records
sp.07. Level of Analysis: 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:
sp.03. 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

- This is an intermediate-outcome measure electronic clinical quality measure (eCQM) utilizing electronic health data at the facility level that provides a standardized method for monitoring the performance of diagnostic Computed Tomography (CT) Scan radiation doses, a risk factor for cancer, while preserving image quality.

- The developer provided a logic model for this intermediate outcome measure which links physician choice of protocol, CT scan, with the intermediate outcome of patient exposure to radiation and the ultimate outcome of cancer.
- The developer cited two systematic reviews:
    - *Early life ionizing radiation exposure and cancer risks: systematic review and meta-analysis* published in Pediatric Radiology in January 2021:
        - The systemic review found that "CT exposure in childhood appears to be associated with increased risk of cancer (leukemia and brain tumors) while no significant association was observed with diagnostic radiographs."
        - The systematic review examined 21 observational studies, including 11 case-control studies and 10 cohort studies each with Newcastle-Ottawa Scale (NOS) scores ranging from seven to nine (with nine being the highest score possible).
        - This systematic review pertained to pediatric patients and not adult patients, which are the focus of this measure.
    - *Epidemiological Studies of Low-Dose Ionizing Radiation and Cancer: Summary Bias Assessment and Meta-Analysis* published in JNCI Monographs in July 2020 that included a combination of medical and non-medical exposures to radiation and the risk of cancer.
        - The review tested whether the median excess relative risk (ERR) per unit dose equals zero and assessed the impact of excluding positive studies with potential bias away from the null. In addition, there was a meta-analysis to quantify the ERR and assess consistency across studies for all solid cancers and leukemia.
        - The review of 26 studies concluded that these new epidemiological studies directly support excess cancer risks from low-dose ionizing radiation. Furthermore, the magnitude of the cancer risks from these low-dose radiation exposures was statistically compatible with the radiation dose-related cancer risks of the atomic bomb survivors.
- The developer also described the *Epidemiological study to quantify risks for paediatric computerized tomography and to optimise doses* (EPI-CT) study: a European pooled epidemiological study to quantify the risk of radiation-induced cancer from pediatric CT (Bernier, 2019). 4 contributing country-specific portions of the cohort are and show positive associations between CT and cancer incidence:
    - The British study reported a positive dose-response relationship between radiation dose and leukemia and CNS tumors in children and young adults.
    - The German study reported a significantly increased incidence of all cancer and lymphoma in exposed children compared with the general population.
    - The French and the German cohorts reported a dose-related increase for CNS tumors.
    - The Dutch study reported a dose-response relationship for CNS tumors.
- The developer also cited the ongoing Life Span Study (LSS) of atomic bomb survivors in Hiroshima and Nagasaki, Japan, which provides quantitative estimates of cancer risks associated with exposure to radiation and is a major source of human data used for risk assessment in establishing radiation safety standards.
    - The eligible cohort included 105,444 subjects who were alive and had no known history of cancer at the start of follow-up (1958-2009)

- ○ The developer states that these analyses demonstrate that solid cancer risks remain elevated more than 60 years after exposure and that approximately 10% of cancers in the cohort are due to the radiation.

Questions for the Committee:

- ○ *Does the Committee agree there is sufficient evidence presented by the developer that links this intermediate process outcome (i.e., radiation exposure) to an outcome (i.e., cancer)?*

Guidance from the Evidence Algorithm

Not a health outcome (Box 1)→ Systematic review and grading of the body of empirical evidence for the immediate-outcome measure is provided (Box 3) → → I Quality, quantity and consistency of the body of evidence from a systematic review provided (Box 4) → Quality (High), Quantity (Mod) and Consistency (Mod) → MODERATE

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

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

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

- The measure has been field tested across 7 health systems, representing 35,729 CT exams and 16 hospitals. The measure is reported at the level of the hospital (identified by CMS Certification number, CCN). Data were collected from an approximately four-week period at each testing site, spanning the years 2020-2021.
  - ○ The mean performance score was 31% with a standard deviation of 18% and a range of 2-100%

Disparities

- The developer examined differences based on age (-0.004 correlation) and sex and found minimal variation between male and female patients in the University of California, San Francisco (UCSF) Radiation Dose Registry.
- The developer states that studies have found that social factors including sex, race/ethnicity, and socioeconomic status are not predictive of radiation dose for CT exams., however patients living in poverty are at higher risk for comorbid conditions associated with exposure to multiple scans over time and increased cumulative exposure to ionizing radiation from diagnostic imaging.

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure for clinicians?
- Is there additional concerns about the presence of disparities in this measure?

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

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

- Evidence between radiation exposure and cancer is strong, but less clear the relationships with CT scans. Still, face validity for the measure based on the evidence is moderate. There is likely correlation at the facility level in terms of standardized practices.
- good evidence - moderate
- No concerns
- This is a new intermediate clinical outcome measure. It intends to improve the performance of diagnostic CT at facility level, by monitoring excessive radiation dose or inadequate imaging for adult patients. The developer provides the same logic model with evidence as in #3633e and 3662e. The rating on evidence is moderate.
- Link to cancer is evident.
- Solid, large scale evidence that links to eventual outcome/harm
- I have the same comment for all three related measures - I don't think it passes the evidence threshold - I vote "Low" because 2 systematic reviews cited, one is pediatric and not really applicable, the second one included mostly non medical exposure of radiation and only 4/26 studies were medical and of these 4 2 were pediatric again so I'm not sure there is sufficient evidence linking CT radiation exposure to cancers in adults. Who have potentially less early stage cells than kids and have less remaining lifetime to develop the cancers.
- Obvious high level of evidence linking radiation exposure to cancer. A point of discussion may be the evidence surrounding thresholds for "out-of-range" values to define numerator.
- same comment as for previous 2 related measures
- Substantial variation exists across facilities, suggesting a lot of opportunities for improvement.
- High opportunity for improvement
- No concerns
- The measure was tested across 7 health systems, representing 35,729 CT exams and 16 hospitals from 2020 to 2021. The mean performance score was 31% with a standard deviation of 18% and a range of 2-100%. The performance gap is rated as high. The only disparity that was identified is the patient population at economic disadvantage. These patients have a potential of higher risk of increased accumulative exposure to radiation scans due to comorbidity.
- Limited data on facility performance and disparities.
- There is clear variation that persists at facility and regional level
- The mean performance score was 31% with a standard deviation of 18% and a range of 2-100%
- no concerns

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.

- Submitted measure specification follows established technical specifications for eCQMs (QDM, HQMF, and CQL) as indicated Sub-criterion 2a1.
- Submitted measure specification is fully represented and is not hindered by any limitations in the established technical specifications for eCQMs.

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.

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.

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

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

Evaluators: Alex Sox-Harris, Samuel Simon, Zhenqiu Lin, Laurent Glance, Matt Austin, Terri Warholak, Jeffrey Geppert, Christie Teigland, Eugene Nuccio, Lacy Fabian, Marybeth Farquhar, Joseph Kunisch

Methods Panel Review (Combined)

Methods Panel Evaluation Summary:
This measure was reviewed by the Scientific Methods Panel. A summary of the measure and the Panel discussion is provided below.

Reliability

- Reliability testing at the Accountable Entity Level
  - The developer conducted a signal-to-noise analysis using ICC on electronic health records from 16 hospitals within 7 health systems and one vertically integrated organization from February 2020 to April 2021.
- The number of CT exams obtained during inpatient hospitalizations (n=15) in the one month of testing data ranged from 134-1,568 (mean 715); thus, the number of CT exams from inpatient settings per hospital is estimated to vary from 1,608-18,816 for a 12-month period.
- The estimated mean split-half ICC using 37,172 CT exams was 0.99. The number of exams per hospital in the one month of data used for testing ranged from 625 to 6,157 (mean=2,323); predicted reliability for 12 months exceeded 0.99 for every hospital.
- For the individual hospitals, the predicted reliability for 12 months of inpatient CT exams exceeded 0.99 for every hospital during the testing phase.

Validity

- Validity testing at Patient/Encounter Level
  - CT category – An ICD-10 based algorithm to assign the CT category was compared to chart review as the gold standard. The results, weighted by the distribution of CT categories in the UCSF International CT Dose Registry, were a sensitivity = 0.86 and specificity = 0.96 (n=978 CT exams). When tested across the 16 hospitals, the correct classification rate of the assignment of CT exams to CT category in field-testing was 92% on average and varied from 88-97% across the 16 hospitals.
  - Patient size – A previously validated algorithm that used cross-sectional imaging to generate patient size estimates was compared to how often this method generated clinically plausible and non-missing data. Size-adjusted radiation dose could be calculated and was within plausible range for 99% of CT exams and was missing for 0.4% of exams.

  - Radiation dose – Dose-length product is an element is generated by the CT machine for each examination and relies on published work. The developer tested how often this method generated clinically plausible and non-missing values for radiation dose in testing data.
  - Size-adjusted radiated dose - Using field testing data, the developer assessed whether it could calculate size-adjusted radiation dose within a plausible range and quantified missing data. Size-adjusted radiation dose could be calculated and was within plausible range for 99% of CT exams and was missing for 0.4% of exams.
  - Global noise – The developer tested whether global noise could be calculated within a plausible range and quantified missing data. Global noise was also correlated with physician dissatisfaction with image quality. Global noise could be calculated and was within a plausible range for 100% of CT exams in field-testing. Global noise was missing for 0.01% of examinations. The correlation between noise and physician dissatisfaction with image quality is 0.37 overall based on the image quality study (n=727 CT exams).
  - Thresholds for “out-of-range” values to define numerator – The developer used physician satisfaction with CT images as a basis for establishing the maximum radiation dose and minimum image quality thresholds for each CT category.
- Validity testing at the Accountable Entity Level
  - Gold standard comparison: The developer compared the eCQM against medical record review using field testing data collected from 8 health systems/vertically integrated organizations.
    - The "medical record review" was a human-reviewed indicator of whether the size-adjusted radiation dose or global noise of each sampled exam exceeds predetermined thresholds, thus constituting a “gold standard.”
    - In a sample of 7000 exams, the out-of-range results (measure score) from the medical record review and the eCQM computation were identical with no discrepancies between the two approaches
    - The developer stated the results indicate a correct and robust implementation of the measure logic.
  - Face validity: A 5-question poll was posed to a TEP which represented a diverse group of clinicians (N=10), patient advocates (N=2).
    - 100% (voted “very likely,” or “somewhat likely on a Likert scale) of the TEP agreed that radiation dose and imagine noise are relevant metrics of quality for CT imaging, size is an appropriate method for adjusting for radiation dose for a given indication, and performance on this measure of radiation dose and image quality, adjusted for size, stratified by indication, would be a representation of quality.
    - 94%-100% agreed that implementation of the measure in federal programs would lead to a reduction in average CT radiation dose while maintaining adequate CT image quality
- Missing data:
  - One SMP member expressed concerns about missing data only focusing on the "radiation dose" aspect of the measure. The missing data information provided in Table 2b-3 also made the SMP question where there could be issues with wider implementation of the measure.

Questions for the Committee regarding reliability:

- *Does the committee have concerns with the reliability of this measure?*
- *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:

- *Does the committee have concerns about the results or approach to the validity testing for this measure?*
- *The Scientific Methods Panel is satisfied with the validity testing for the measure. Does the Committee think there is a need to discuss and/or vote on validity?*

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

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

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

- No concerns; agree with SMP panel rating of high as it appears the the metrics can be accurately and consistently pulled from machines
- Review panel rated high
- No concerns
- It seems that the measure follows technical specifications for eCQMs and is not hindered by any limitations. I have no concerns.
- Reliability results are adequate.
- No concerns given simplicity
- Elements presented adequately.
- no concerns
- no concerns
- no
- No concerns
- Reviewed by SMP. Reliability testing was conducted at facility level. A signal-to-noise analysis was conducted. Electronic health records were from 16 hospitals within 7 health systems and one vertically integrated organization from February 2020 to April 2021. The number of CT exams from inpatient settings per hospital is estimated to vary from 1,608-18,816 for a 12-month period. For the individual hospitals, the predicted reliability for 12 months of inpatient CT exams exceeded 0.99for every hospital during the testing phase. The preliminary rating on reliability is high.
- no
- Strong, no concerns
- No concerns
- no concerns
- No, agree with a moderate/high ranking for validity based on evidence and frameworks.
- Review panel rate high
- No concerns
- Reviewed by SMP. Validity was also tested at both patient/encounter level and clinician group level. At patient/encounter level, the same 6 parameters were evaluated, CT category, patient size, radiation dose, size-adjusted radiated dose, and global noise. The CT category was signed using an ICD-10 algorithm and then compared to chart review as the gold standard. The analysis produced the same sensitivity of 0.86 and specificity of 0.96 as in the other two measures, #3633e and #3662e. When tested across the 16 hospitals, the correct classification rate was 92% on average and varied from 88-97% across the 16 hospitals. For testing at hospital level, the Gold standard comparison and face validity were evaluated. I have no concerns. The preliminary ratings for reliability and validity are both high.
- As an intermediate process measure it falls short of predicting patient harm from exposure because the overdose of ionizing radiation and age of the adult are not considered.
- Very strong, no concerns
- Not as presented
- no concerns
- Exclusions of multi-site CT scans described by developer seem appropriate. All adult patients are included. Risk adjustment for body size seem very well justified.
- none noted
- No concerns
- There does not appear to be any risk adjustment.

- na
- Well handled
- Are the ranges within generally agreed upon as standard of care?
- no concerns
- Missing data appears minimal (0.4%) and metrics seem consistent across CT scanners
- None noted
- No concerns
- SMP reviewed both reliability and validity. One SMP member expressed concerns about missing data only focusing on the "radiation dose" aspect of the measure.
- na
- Limited threat by these, esp. missingness and error
- N/a
- no concerns
- NA
- NA
- Not apply.
- na

Criterion 3. Feasibility

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

- Data for this measure generated or collected by and used by healthcare personnel during the provision of care
- ALL data elements are in defined fields in a combination of electronic sources
- The submission includes two measure specifications, a HQMF/QDM measure specification and a FHIR measure specification. Both measure specifications follow established technical specifications for eCQMs as indicated Sub-criterion 2a1.
- Submitted measure specifications are fully represented and are not hindered by any limitations in the established technical specifications for eCQMs.
- Using a simulated data set, the submission demonstrates that the evaluation of 100% of the measure logic can be automated.
- The Feasibility Scorecard indicated that the no data elements have issues with accuracy and 100% coverage in simulated data unit tests.
- There was concern from an SMP member that specification was heavily dependent on proprietary software developed by UCSF and Alara Imaging, Inc. to access and process primary data elements from the electronic systems to calculate the three variables required by the measure – CT category, size-adjusted radiation dose, and global noise. This software in turn requires access to raw imaging data. Although the developer states that this process has been tested in multiple settings, the SMP member was concerned that there was no evidence that a garden variety clinician could reliably replicate.

Questions for the Committee:

- Does this measure appear to be feasible as an eCQM?

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

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

- Very feasible, no concerns
- High
- No concerns
- The preliminary rating is again high. A SMP member raised concern that specification was heavily dependent on proprietary software developed by UCSF and Alara Imaging, Inc.
- no concerns
- Strong and seem extractable without big challenges to fidelity
- There was concern from an SMP member that specification was heavily dependent on proprietary software. Impact on clinicians without applicable software?
- no concerns

Criterion 4: Usability and Use

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

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

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

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

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

OR

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

Accountability program details

- The measure is not currently in use in any accountability programs.
- The developer states that this measure will be submitted for Centers for Medicare & Medicaid Services (CMS) Merit-based Incentive Payment System (MIPS). MIPS measures are publicly reported on Care Compare by 2026 because measures are not publicly reported for two years.
- The developer also states that this measure will be submitted to CMS' Measures Under Consideration list for 2022.

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

- The developer states that verbal feedback was provided by site participants on the video calls. Feedback from sites often reflected a recognition and understanding for why radiation doses were particularly high.

- Feedback received influenced the developer to the feedback for the measure to be more nuanced than the aggregate level to make the measure actionable.

Additional Feedback: N/A

Questions for the Committee:

- Can the performance results be used to further the goal of improving patient safety through reducing excessive radiation dosing?

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

- This eCQM is not currently used in any quality improvement program.

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

- One unexpected finding was the lack of consistency among facilities saving Radiation Dose Structured Reports (RDSR). The developer worked with sites to modify their systems to save the RDSR to capture 94% of dose reports.
  As the goal of this measure is the reduction of patient exposure to radiation, the developer noted a concern that radiation dose reduction might result in deteriorated image quality but did not find any evidence of poor image quality in the results. The developer stated that this potential issue will be monitored annually.

Potential harms

- There are no harms identified by the developer.

Additional Feedback: N/A

Questions for the Committee:

- How can the performance results be used to further the goal of safer care?
- Do the benefits of the measure outweigh any potential unintended consequences?

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

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

- Planned inclusion in an accountability program, not currently publicly reported. Plans for use in MIPS reporting and CMS
- Not currently publicly reported or used in accountability, plans for however
- No concerns
- The measure is currently not publicly reported in any accountability program. However, the developer indicates this measure will be submitted to the Centers for Medicare & Medicaid Services (CMS) Merit-based Incentive Payment System (MIPS). MIPS measures are publicly reported on Care Compare by

2026 because measures are not publicly reported for two years. The developer also states that this measure will be submitted to CMS' Measures Under Consideration list for 2022. Site participants provided feedbacks via video calls, which showed recognition and understanding high radiation doses. I think the performance results can be used to further the goal of improving patient safety through reducing excessive radiation dosing. The preliminary rating for use is pass.

- May improve patient safety.
- Not systematically used yet
- Not publicly reported and unclear intended use by measure developer outside of planned submission to CMS.
- no concerns
- Does not appear to be high risk of unintended consequences. Image quality concerns will be monitored per developer.
- Moderate
- No concerns
- The developer noted a concern that radiation dose reduction might result in deteriorated image quality but did not find any evidence of poor image quality in the results. The developer will monitor this potential issue annually. No potential harm was identified, though. The usability is preliminarily rated as moderate.
- The prevention of cancer depends on the magnitude of the overdose and the age of the patient. None the less, it seems that limiting overdoses matters.
- None expected, not in use yet
- Benefits > harms
- no concerns

Criterion 5: Related and Competing Measures

Related or competing measures

- Two measures were identified as related:
  - 2820: Pediatric Computed Tomography (CT) Radiation Dose (UCSF)
  - 3621: 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 (American College of Radiology)

Harmonization

3663e: Excessive Radiation Dose or Inadequate Image Quality for Diagnostic Computed Tomography (CT) in Adults (Facility Level)

- Population: All diagnostic CT exams performed on adults (aged 18 years and older) during the measurement period of one year that have an assigned CT category, a size-adjusted radiation dose value, and a global noise value.
- Outcome: Assesses radiation dose according to thresholds determined by the underlying clinical indication for imaging

3621: 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 (Facility; Clinician-Group Level)

- Population: Includes all patients regardless of age. Includes CT Abdomen-pelvis exams with contrast (single phase scans), CT Chest exams without contrast (single phase scans), and CT Head/Brain (single phase scans)
- Outcome: 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

2820: Pediatric Computed Tomography (CT) Radiation Dose

- Population: Diagnostic CT scans performed on children of the head, chest, abdomen/pelvis and chest/abdomen/pelvis in children.
- Outcome: Whether CT doses exceed published benchmarks

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

- related measures do not appear to be major competing.
- Pediatric and Composite measure
- No concerns
- Two measures were identified as related: (1) #2820: Pediatric Computed Tomography (CT) Radiation Dose (UCSF); (2) #3621: 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 (American College of Radiology). They are not competing or overlapping.
- Should be combined with #3662e and #3633e.
- Also competes w 3621. Not sure all needed

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 1/19/2022

- **Of the 1 NQF members who have submitted a support/non-support choice:**
  - **0 support the measure**
  - **1 do not support the measure**

Comments and Member Support/Non-Support Submitted as of: 2/2/2022

Comment 1 by: American Cancer Society

I am pleased to provide this comment in support of NQF quality measures 3633e, 3662e and 3663e. These comments reflect my personal opinion and not any other organization with which I may be affiliated. CT scans have assumed a primary role in the evaluation and diagnosis of many medical conditions, and are very commonly performed procedures. Less appreciated by the public and many professionals (including non-radiology physicians) is the variation in image quality and dose that has been recognized for many years by researchers who have evaluated these factors. As such, there can be substantial variation in CT scan dose and quality, even within the same institution. As a patient, this consideration has figured prominently in my own decisions as to whether or not to proceed with serial CT scans for follow-up of medical conditions. These measures have been carefully crafted to create an effective and validated method to monitor CT image and quality based on indications for the studies and in consideration of individual patient-related variables. As such, they provide a useful and meaningful way to offer our patients and the public the assurance that the scans they are receiving meet reasonable safety and professional standards--which is not routinely available otherwise. These quality measures will meaningfully improve the ability of physicians and heatlh systems alike to monitor the equipment utilized for these studies in a manner that minimizes interference with the tyipical workflow of a radiology center (or other center) where such studies are performed and will provide a significant and substantial increase in the quality of scans while reducing dose variability that can occur because of machine settings/performance or patient characteristics. Cumulative radiation dose should decline as a result of implementing these measures. At the very least, there will be assurance that the right dose is used for the right scan in the right patient. As a physician and patient advocate for many years, I offer my support for these measures for the reasons stated. And as someone who served as an advisor for this measure, I will add that I was impressed by the exceptional commitment of the developers and their colleagues to provide a meaningful, validated and effective quality measure as they created new processes to measure CT dose and quality, always with an eye towards making this measure acceptable to the professional and consumer communities. (Disclosures: As noted, I was an advisor during the development of this measure and received compensation for those services. I have also served on the NQF Cancer Committee without compensation. I have no other relevant conflicts.)

Comment 2 by: American College of Radiology

The American College of Radiology, representing more than 40,000 radiologists, radiation oncologists, medical physicists, and nuclear medicine physicians, appreciates the opportunity to submit comment on NQF #3633e, #3662e and #3663e: Excessive Radiation Dose or Inadequate Image Quality for Diagnostic Computed Tomography (CT) in Adults (Clinician Level, Clinician Group Level and Facility level, respectively). ***The ACR does not support the endorsement of NQF #3633e, #3662e, and #3663e.*** ***General Comments*** Protocol selection appropriate for a clinical indication is an important component of radiation dose management along with radiation dose optimization. Each component needs to be addressed as a separate quality action. The specific aspect(s) of performance to be improved is not intuitive due to the multiple components to the measures (size-adjusted dose, image quality, clinical indication). It is premature to measure performance on excessive radiation dose based on thresholds by clinical indication for an exam until the level of standardization and availability of national benchmarks is further along as discussed below. It is true that the most accurate way to address appropriate and safe use of multi-phase studies is to measure both the clinical indication of an exam and the radiation dose output (dose indices per exam) and look at the two separately or distinctly together. ***However, these measures conflate the appropriateness of protocol for the clinical indication and radiation dose optimization, disregarding applicability, from which a facility may not be able to determine if its performance could be improved by adjusting protocols or by focusing on appropriateness of the ordered exam. Therefore, improvement may be limited.[1]*** Dose optimization results in a quality action for facilities to adjust their protocols and is a responsibility of the team as a whole – physicists, technologists, and physicians who oversee the team at the facility. Protocol selection addresses the appropriateness of the exam for the clinical indication and other factors such as patient time on the scanner and optimal radiation dose. There are challenges with the implementation of an indications-based measure. Indications for exams do not have standardized language that could be used to track them. Most health and IT systems capture ICD-10 coding for reimbursement, but typically not enough standardized information to characterize the patient's condition. As a result, the clinical reason for performing an imaging exam is often extremely limited in the exam order. Electronic Health Records (EHRs) are notoriously incomplete with this type of information and interoperability issues exist with other software systems that might contain such information. ***A *validated* method for determining classification of studies using high-dose versus routine protocols appropriate to the indication must be incorporated into such a measure; these three measures include specifications which have not been validated.*** Please refer to the validity section below for more details. ***NQF #3633e, #3662e, and #3663e deviate from international standards, like diagnostic reference levels, and lack peer-reviewed, broadly**

accepted consensus on global noise. For these measures, global noise is defined solely by the measure developer. Endorsing this method may encourage facilities to accept a narrow view of image quality.* *The ACR requests the developer further clarify the global noise table used in calculating the numerator.* The benchmark source is not transparent, and its applicability is unclear. For example, Table sp-1, Size-adjusted radiation dose and global noise thresholds by CT category, has the same global noise threshold for several CT categories, such as head low dose, head routine dose, and head high dose. Is it intentional that the same global noise threshold should be applied to both low and high dose head CTs? If the image noise thresholds are the same, the size-adjusted radiation dose thresholds should be the same, unless the scan length is remarkably different between the 3 CT categories. Additionally, current CT scanners display dose values based on either a 16 cm or 32 cm phantom for a neck scan, which must be carefully accounted for in measure performance calculations. ***There is little to no acknowledgement of limitations.*** These measures have multiple limitations, including the lack of widespread acceptance and implementation, and the issues with the method of measuring global noise. The developer states their company can provide the service of quantifying the measure at a cost; this should also be included as a potential limitation. The measure developer does provide specifications for other entities to implement the measure, but the burden of implementation may be significant. Finally, the author cites publications from their group to justify the benchmarks, but they have not been vetted through a broader consensus process. ***The ACR strongly encourages the Patient Safety Standing Committee to re-vote on the scientific acceptability of these measures based on the following concerns.* *Validity/Feasibility*** These eCQMs require multiple variables that may be captured in software systems external to electronic health records (EHRs), such as dictation systems housing radiology reports or DICOM standard-based systems, such as CT device software. Data element validity testing should demonstrate that the testing sites were able to integrate and validate the variables used to construct the data elements used by the eCQM in addition to the usual validation of the eCQM's electronic output against the medical record review. ***We are uncertain that this validation has been completed. Therefore, this submission does not demonstrate the measure can be reproduced in a reliable and valid manner by practices or facilities across multiple settings.*** For example, for CT category (or other elements deriving/collecting data using custom natural language processing (NLP) tools), the developer used NLP for obtaining data such as reason for study or protocol name used in the calculation of this variable. The submission does not provide information on the NLP results' reliability and validity. Because ***this comparison of the NLP-derived data against a medical record review was only completed in a sample from one site (UCSF Health System), there is uncertainty whether the results are generalizable across EHRs or other databases**.* These measures rely on custom made NLP trained and validated on a small group of pilot sites; it is not clear whether this type of NLP would work outside these sites nor how sites would get access to use this custom NLP tool. Testing information does not demonstrate adequate validation of this critical data element. Additionally, ***sufficient evidence should demonstrate that the definitions/variables used are valid and do not rely on one study or use in a single system, such as what is provided to support the thresholds of "out of range" performance values.*** While the process to determine these thresholds is detailed, we do not believe that a Technical Expert Panel (TEP) conclusion in the absence of independent data validation is sufficient. *M**ultiple unstructured variables are required to construct the data elements for the numerator, denominator, and exclusions. Assessments of the feasibility of the integration of these unstructured data into the measure calculations would be useful to ensure that the underlying data can, in fact, be integrated if practices and facilities that choose not to use the edge device.*** For example, the level of effort required to integrate the Binning algorithm for the CT categories and ensure that the results are reproducible and valid remains unclear. The ACR is concerned with the selection bias for the accountable entity-level (measure score) validity. ***Assessing measure score face validity through the TEP that created these measures lessens the extent of credibility for these results**.* Although the TEP is knowledgeable and represents a variety of stakeholders, there is a vested interest in ensuring these measures are available for use. ***Most importantly, as one of the TEP members noted in the survey, the performance score from these measures does not clearly indicate what corrective action needs to be taken by the clinician, clinician**

group, and/or the facility to improve performance.* ***Usability*** While implementing these measures as specified may not impose a substantial burden on clinicians, ***it may necessitate substantial organizational effort to access and process the data elements required to calculate the measure score**.* The measure steward states that their software is available on a non-commercial basis to calculate this measure, and that other vendors may also develop their own software to implement the measure specifications using the information included in this submission. Will the measure steward review other vendors' software to ensure comparable calculation methods? Measure stewards frequently make specifications available "as is" without warranty, leaving it to the implementer to appropriately update any software or tools as measure specifications are changed. But the complexity of these measure specifications may warrant greater oversight. External vendor software will need to be maintained and updated to ensure the software's accuracy and reflect any changes in specifications and coding. ***For all the reasons stated above, the ACR does not support the endorsement of these three measures.*** We thank the NQF staff for their transparent endorsement process. **Reference: 1. '**Mahesh M. Benchmarking CT Radiation Doses Based on Clinical Indications: Is Subjective Image Quality Enough?Radiology. 2021 Nov 9:212624. doi: 10.1148/radiol.2021212624. Online ahead of print. PMID: 34751622

Comment 3 by: Angela Keyser,

What is AAPM: The American Association of Physicists in Medicine (AAPM) is the primary scientific and professional organization of physics in radiology and radiation oncology in the United States. The mission of AAPM is advancing medicine through excellence in the science, education and professional practice of medical physics; a broad-based scientific and professional discipline which encompasses physical principles with applications in biology and medicine. With 9717 members in 94 countries, AAPM supports the Medical Physics community with a focus on advancing patient care through education, improving safety and efficacy of radiation oncology and medical imaging procedures through research, education and the maintenance of professional standards. AAPM has a staff of 33 and an annual budget of $10.7M, and is located at 1631 Prince Street, Alexandria, VA 22314. **AAPM comments on the proposed measures:** AAPM does not support the endorsement of NQF #3633e, #3662e, and #3663e. This application proposes electronic clinical quality measures (eCQM) that monitor CT performance to discourage unnecessarily high radiation dose while maintaining adequate image quality. The proposed metrics require CT Category (i.e., the CT exam type), the size adjusted radiation dose [the patient's dose length product (DLP) adjusted by patient size], and the global noise (associated with the variance of the voxel values in CT images). The two reported measures are the percentage of eligible CT cases in a particular category deemed to be "out-of-range" compared to defined thresholds with respect to the size-adjusted radiation dose or the global noise in a set time period. While efforts to enhance consistency of CT practice are noble and include initiatives by AAPM and others worldwide, the proposal has significant limitations that impact its scientific and practical value and overall likelihood of clinical acceptance. These limitations include improper representation of image quality, improper estimation of radiation risk, and substantial oversimplified representation of implementation in practice, including not addressing the challenges of implementation. The authors indicate that their company (Alara Imaging, Inc.) can provide the service of quantifying the measures at a cost. A steward of measures requires an extensive track record for scientific and technical expertise and policy making that represents a broad consensus of the community. These important elements should be carefully reviewed within this application. One cited reference supports the proposed measure, however, this cited article has an accompanied editorial that highlights the limitations of the proposed approach [Mahesh M.Benchmarking CT Radiation Doses Based on Clinical Indications: Is Subjective Image Quality Enough? Radiology. 2021 Nov 9:212624. doi: 10.1148/radiol.2021212624. Online ahead of print. PMID: 34751622]. The editorial and stated limitations are not addressed in the proposal. The AAPM agrees that effort needs to be continually placed on ensuring diagnostic quality CT imaging, optimizing CT dose, and achieving consistency across facilities, considering differing technologies and practices. The non-profit entities of the AAPM, the American College of Radiology (ACR), and Image Wisely and Image Gently Alliances have

spent decades towards this goal and continue to do so through many initiatives. Among them, the non-profit ACR CT Dose Index Registry (DIR; https://www.acr.org/Practice-Management-Quality-Informatics/Registries/Dose-Index-Registry, established in 2011) has the significant stature of implementing a dose registry that enables facilities to compare dose indices nationally, to ensure the highest quality imaging with lowest possible dose. The ACR CT DIR implementation incorporates the expert, consensus opinions of the medical imaging community. ACR dose optimization measure recently endorsed by NQF provides a further valuable measure to manage imaging radiation dose (https://www.qualityforum.org/QPS/3621). The imaging community's valuable clinical benchmarks greatly benefit from consensus decisions based on sound scientific and technical review and discourse. The proposal herein should be carefully reviewed for any additional contributions or advantages it would provide to our existing robust consensus measures and resources, such as available with the ACR. After a detailed review of the measures by multiple expert members of the AAPM, we have concluded that the **AAPM does not support the endorsement of NQF #3633e, #3662e, and #3663e.** This position stems from eight major concerns about the proposed measures: 1) Unscientific characterization of CT scan risk: The proposal is based on estimation approaches that are not reflective of the consensus of the scientific community and do not acknowledge the uncertainties of the estimates. A NQF measure focused on radiation risk should uphold scientific objectivity, integrity, and responsibility not evident in the presentation and assessment of radiation risk in this proposal. 2) Inactionability of the measures to enable targeted change to improve practice: It is not evident how the proposed measures can be practically used to improve imaging practice and exactly how a facility can do to achieve compliance, given the wide varieties of factors and technologies involved. 3) Inadequate addressing of the complexity of CT categorization: The proposal does not address the magnitude of this challenge nor has suggested means to overcome it given that current standards are even lacking in uniform characterization of protocols. Inaccurate classification of data can lead to significant and misleading errors.
4) Inadequate assessment of noise: Noise in a CT image can be influenced by a variety of factors including justified differences in CT technologies including new reconstruction methods that dramatically alter noise. Further, noise does not have a singular value in a CT exam. A "global noise" ignores this diversity and can misrepresent the quality of an exam. 5) Inadequate assessment of image quality: Image quality is affected by a myriad of factors including resolution and contrast, as well as the intended purpose of the exam. A singular representation of image quality via global noise overly simplifies this space and can lead to gross misrepresentation of image quality and thus mis-service to patient care.
6) Flawed assumption on dose reduction vs dose optimization: The application focuses primarily on radiation dose reduction as oppose to right-sizing the dose for the best care of the patient. Individualization and optimization of care and safety should be the goal not minimization. This approach can lead to some patients getting under exposed, leading to missed diagnosis, while others may be over-dosed for their exact need and condition. 7) Inadequate accuracy in patient size estimation: Assessing a patient size is not a trivial task, stemming from significant variability in the differences in the habitus of different patients, coupled with the existential challenge that there is no single metric capturing the size of a patient of varying diameter at different cross-sectional locations. Algorithms are continuously evolving and no evidence is provided that the company can do this task with sufficient accuracy.
8) Limited expertise and track record of the company: The company is a new (2020) company with no experience of having previously performed a project of such wide scope, scientifically or technically. There is no scientific track record on CT technology, size estimation, or image quality assessment for the company to be considered a steward of measures on which there is a lack of expertise, publication, and scientific history. These concerns are detailed specially in our complete review submitted via email to patientsafety@qualityforum.org, along with selected specific observations on the proposal on January 19, 2022. The AAPM recognizes that this topic is complex, including scientific, technical and clinical components. We welcome the opportunity for greater in-depth discussion on meaningful measures of quality imaging practice.

Respectfully submitted,

American Association of Physicists in Medicine (AAPM) January 19, 2022

Comment 4 by: Angela Keyser,

What is AAPM:

The American Association of Physicists in Medicine (AAPM) is the primary scientific and professional organization of physics in radiology and radiation oncology in the United States. The mission of AAPM is advancing medicine through excellence in the science, education and professional practice of medical physics; a broad-based scientific and professional discipline which encompasses physical principles with applications in biology and medicine. With 9717 members in 94 countries, AAPM supports the Medical Physics community with a focus on advancing patient care through education, improving safety and efficacy of radiation oncology and medical imaging procedures through research, education and the maintenance of professional standards. AAPM has a staff of 33 and an annual budget of $10.7M, and is located at 1631 Prince Street, Alexandria, VA 22314.

AAPM comments on the proposed measures:

AAPM does not support the endorsement of NQF #3633e, #3662e, and #3663e.

This application proposes electronic clinical quality measures (eCQM) that monitor CT performance to discourage unnecessarily high radiation dose while maintaining adequate image quality. The proposed metrics require CT Category (i.e., the CT exam type), the size adjusted radiation dose [the patient's dose length product (DLP) adjusted by patient size], and the global noise (associated with the variance of the voxel values in CT images). The two reported measures are the percentage of eligible CT cases in a particular category deemed to be "out-of-range" compared to defined thresholds with respect to the size-adjusted radiation dose or the global noise in a set time period.

While efforts to enhance consistency of CT practice are noble and include initiatives by AAPM and others worldwide, the proposal has significant limitations that impact its scientific and practical value and overall likelihood of clinical acceptance. These limitations include improper representation of image quality, improper estimation of radiation risk, and substantial oversimplified representation of implementation in practice, including not addressing the challenges of implementation. The authors indicate that their company (Alara Imaging, Inc.) can provide the service of quantifying the measures at a cost. A steward of measures requires an extensive track record for scientific and technical expertise and policy making that represents a broad consensus of the community. These important elements should be carefully reviewed within this application. One cited reference supports the proposed measure, however, this cited article has an accompanied editorial that highlights the limitations of the proposed approach [Mahesh M. Benchmarking CT Radiation Doses Based on Clinical Indications: Is Subjective Image Quality Enough? Radiology. 2021 Nov 9:212624. doi: 10.1148/radiol.2021212624. Online ahead of print. PMID: 34751622]. The editorial and stated limitations are not addressed in the proposal.

The AAPM agrees that effort needs to be continually placed on ensuring diagnostic quality CT imaging, optimizing CT dose, and achieving consistency across facilities, considering differing technologies and practices. The non-profit entities of the AAPM, the American College of Radiology (ACR), and Image Wisely and Image Gently Alliances have spent decades towards this goal and continue to do so through many initiatives. Among them, the non-profit ACR CT Dose Index Registry (DIR; https://www.acr.org/Practice-Management-Quality-Informatics/Registries/Dose-Index-Registry, established in 2011) has the significant stature of implementing a dose registry that enables facilities to compare dose indices nationally, to ensure the highest quality imaging with lowest possible dose. The

ACR CT DIR implementation incorporates the expert, consensus opinions of the medical imaging community. ACR dose optimization measure recently endorsed by NQF provides a further valuable measure to manage imaging radiation dose (https://www.qualityforum.org/QPS/3621). The imaging community's valuable clinical benchmarks greatly benefit from consensus decisions based on sound scientific and technical review and discourse. The proposal herein should be carefully reviewed for any additional contributions or advantages it would provide to our existing robust consensus measures and resources, such as available with the ACR.

After a detailed review of the measures by multiple expert members of the AAPM, we have concluded that the **AAPM does not support the endorsement of NQF #3633e, #3662e, and #3663e.** This position stems from eight major concerns about the proposed measures:

1) Unscientific characterization of CT scan risk: The proposal is based on estimation approaches that are not reflective of the consensus of the scientific community and do not acknowledge the uncertainties of the estimates. A NQF measure focused on radiation risk should uphold scientific objectivity, integrity, and responsibility not evident in the presentation and assessment of radiation risk in this proposal.

2) Inactionability of the measures to enable targeted change to improve practice: It is not evident how the proposed measures can be practically used to improve imaging practice and exactly how a facility can do to achieve compliance, given the wide varieties of factors and technologies involved.

3) Inadequate addressing of the complexity of CT categorization: The proposal does not address the magnitude of this challenge nor has suggested means to overcome it given that current standards are even lacking in uniform characterization of protocols. Inaccurate classification of data can lead to significant and misleading errors.

4) Inadequate assessment of noise: Noise in a CT image can be influenced by a variety of factors including justified differences in CT technologies including new reconstruction methods that dramatically alter noise. Further, noise does not have a singular value in a CT exam. A "global noise" ignores this diversity and can misrepresent the quality of an exam.

5) Inadequate assessment of image quality: Image quality is affected by a myriad of factors including resolution and contrast, as well as the intended purpose of the exam. A singular representation of image quality via global noise overly simplifies this space and can lead to gross misrepresentation of image quality and thus mis-service to patient care.

6) Flawed assumption on dose reduction vs dose optimization: The application focuses primarily on radiation dose reduction as oppose to right-sizing the dose for the best care of the patient. Individualization and optimization of care and safety should be the goal not minimization. This approach can lead to some patients getting under exposed, leading to missed diagnosis, while others may be over-dosed for their exact need and condition.

7) Inadequate accuracy in patient size estimation: Assessing a patient size is not a trivial task, stemming from significant variability in the differences in the habitus of different patients, coupled with the existential challenge that there is no single metric capturing the size of a patient of varying diameter at different cross-sectional locations. Algorithms are continuously evolving and no evidence is provided that the company can do this task with sufficient accuracy.

8) Limited expertise and track record of the company: The company is a new (2020) company with no experience of having previously performed a project of such wide scope, scientifically or technically. There is no scientific track record on CT technology, size estimation, or image quality assessment for the company to be considered a steward of measures on which there is a lack of expertise, publication, and scientific history.

These concerns are detailed specially in our complete review submitted via email to patientsafety@qualityforum.org, along with selected specific observations on the proposal on January 19, 2022.

The AAPM recognizes that this topic is complex, including scientific, technical and clinical components. We welcome the opportunity for greater in-depth discussion on meaningful measures of quality imaging practice.

Respectfully submitted,
American Association of Physicists in Medicine (AAPM)
January 19, 2022

Comment 5 by: Dawn Ritzwoller

I am a college student and Environmental Biology (E-bio) major, and I am pediatric cancer survivor. I am writing today in support of this radiation dose quality measure. Beginning ten years ago, and both during and after I finished treatement, I recieved multiple CTs (to multiple parts of my body) as part of my diagnostic and follow-up care. Not once during this period, did any of my doctors or other, discuss with me the downstream risk of all of the radiation exposure I experienced. It was only years after my treatment ended, and now via classes I have take for my E-bio major, that I am beginning to understand the risk associated with radiation exposure. What is also now clear to me is the importance that providers usie the most appropriate (low) dose for the specific diagnostic or follow-up exam. I know that image quality is important for diagnosis, but patients (like me) need the confidence that their doctors and hospitals are using the best and lowest dose possibe for the exam that they order. Thank you!

Comment 6 by: Debra Ritzwoller

I am writing in support of this important measure. I am a cancer health services researcher *and* a mother of a pediatric cancer survivor. It is well documented in the literature that there has been a significant secular increase in CT use within and across most patient populations. While CT use, and therefore radiation exposure has increased over time, I know that personally and professionally that excessive radiation dose remains a significant quality issue, and it is one that is often not adequately addressed by researchers and healthcare providers/delivery systems. This quality metric is necessary now, in order to provide the incentives and the resources needed to generate the metrics and the benchmarks that may actually influence practice that may in turn translate into a meaningful reductions in the radiation dose that patients are exposed to. This metric is designed to address the clinical indication associated with the respective exam, rather than just the type of advanced imaging that is performed. The measure is also constructed to ensure that the dose benchmarking does not adversely impact the quality of the metric. Given the noted harms of CT based radiation exposure (e.g USPSTF Lung Cancer Screening "B" recommendation), this measure is addresses a timely and needed quality metric.

Comment 7 by: Ehsan Samei

Duke University, Ravin Advanced Imaging Laboratories (Ravin Labs) and Clinical Imaging Physics Group (CIPG), Durham, NC 27710 The Ravin Labs is a 50-member leading translation imaging research laboratory in the country with over 30 years of history. The lab conducts rigorous NIH-funded research with an additional mandate to practice its science through CIPG, an imaging physics group of 15 experts dedicated to quality and safety in the practice of radiology. The group, highly integrated into the clinical domain, has devised and put to practice imaging dose and image quality monitoring systems at the level of individual patients within the Duke University Health System with additional pilot installations at MD Anderson Cancer Center and Stanford University. The group has published extensively on its technology and findings (upward of 500 papers), with over 30 referred publications on dose and quality monitoring alone. The effort has led to significant reduction of patient radiation dose at our facilities and right-sizing it per individual needs of patients. **We do not support the proposed measures.** The rationale is detailed below. **Overall:** While we applaud the effort to introduce new quality measures in the practice of medical imaging, the proposed electronic clinical quality measures (eCQM) are misleading and overly simplistic leading to significant unintended consequences. The limitations stem from the fact that the proposed risk measures are based on CT scanner output and not the actual dose burden to individual patients at the organ level, the quality measure is based on noise alone ignoring the multi-faceted reality of diagnostic quality, and lack of methods that standardize protocols across vast diversity of examinations. There is significant ambiguity in the exact method used for noise and size estimation with no track record or peer review of otherwise black-box methods. This approach will likely produce measures that can be orders of magnitude off from their actual values, and therefore lack clinical relevance and fidelity. Measures can lead to misleading and erroneous conclusions while also potentiallyjeopardizing the use and development of better approaches, as inaccurate low-bar measures can prevent accurate ones in the future. But most importantly, the measure can lead to unintended consequences and even harm the patient. For example, an imaging team can take an action that is not in the best interest of a patient, like applying too little dose for some patients such that disease would be missed, a “wasted dose” with no medical benefit and health and cost consequence of a miss. Conversely others might get more radiation than needed as the measures do not account for individual patient needs and tasks. Improving consistency in imaging practice is a laudable goal that needs a proper solution anchored to scientific understanding of radiation risk, image quality need of patients, diversity of practices, and the CT technology. The proposal is lacking on all these four fronts. A solution to inconsistency in images can only be brought forth through a broad consensus of the scientific and practicing communities (including ACR, AAPM, Image Gently, and Image Wisely), CT manufacturers (represented by MITA), standard methods of data categorizations and measures (supported by the medical community), and evidence-based radiation risk and image quality measures at the level of indication and organ where they are actually relevant to the individual patient. A for-profit company with no track record or transparency of its methods cannot be considered a steward of such a space. Below we further detail 12 concerns regarding the proposed measures:

1. **Inadequate attention to image quality:**The measures are heavily dose related, emphasizing this over measures of quality. Dose and minimizing it is important but equally important is image quality as an inadequate image quality would be a dis-service to the patient regardless of the dose. This is explicitly stated in the International Commission of Radiological Protection (ICRP) in Publication n. 135.

2. **Inaccurate assessment of radiation risk:**The measure of size-adjusted radiation risk, adjusting the CT scanner outputs with ‘patient size’ to perform risk estimation is not a standard method nor endorsed by any scientific or professional body. The method is in fact explicitly <u>discouraged</u> by the AAPM Task Group 204. Patient risk can only be assessed with the knowledge of organ doses that is not even mentioned in the application let alone pursued. The proposed method CANNOT be used as surrogate for future cancer risk.

3. **Incomplete/Inaccurate representation of image quality:**The measures include image noise. Yet, noise is just one component of image quality. For example, the noise of an image can be fine but image quality totally inadequate. And conversely noise can be too high but image quality totally adequate. To assess image quality properly, one should include the actual task at hand (eg, detecting a pancreatic cancer vs bowel obstruction vs kidney stone) as well as other equally important facets of quality, like noise texture, resolution, and contrast. These factors have not been even mentioned let alone tackled in this application. Focusing on noise as a singular metric

of quality can lead to major mis-representation of the needs of a quality and safe imaging practice.

4. **Neglecting the impact of image rendition:**Critical and relevant to clinical practice, the measure of noise proposed does not take into consideration how differing reconstruction algorithms and parameters affect noise (up to 200%). Without considering this influence, a measure of noise as proposed is irrelevant and misleading.

5. **Subjectivity:**The measures are anchored to subjective perception by radiologists as how they "like" the images. There is in fact no evidence provided that the measures can lead to an improvement in diagnostic accuracy. In fact, it might lead to a degradation.

6. **Lack of integrating dose and quality:**There is no indication as to how image quality is linked to radiation dose and at what level; or instance, how they propose to manage multiple reconstructions of the same exposure event.

7. **Not addressing the multiplicity of exam components**: A CT exam often includes multiple phases (series) each of which has a noise and radiation dose of its own. Averaging noise across series is meaningless. The measures do not recognize or account for this multiplicity and diversity.

8. **Under-recognizing the diversity of exams:**The measures do not address the notable diversity of exam nomenclature across institutions and practices. This is a significant component of any dose or quality monitoring system. Without a standard for CT protocols, which cannot be devised by a for-profit company without consensus of manufacturers and users, the data can be mislabeled and mishandled leading to major errors in the results and subsequent negative effect on mis-dosing and mis-diagnosing patients.

9. **Inaccurate assessment of patient size:**The measure of size proposed is calibrated to earlier work and publication from our group at Duke University for academic purposes. That early method they have embraced has had major errors (upward of 300% is certain applications) that have been corrected in subsequent versions that have not been shared. Without essential newer refinements to assure fidelity, the company cannot be a responsive steward of the measure that it has had no expertise to advance or maintain.

10. **Inaccurate assessment of noise:**The measure of noise proposed references earlier work and publication from our group at Duke University. That early method exhibited errors, corrected in subsequent versions that have not been shared. Without essential newer refinements, the company cannot be a responsive steward of the measure that it has had no expertise to advance or maintain.

11. **Lack of guidance toward compliance:** To us it is difficult to defend (1) measuring imaging practices based on ambiguous and questionably-relevant metrics promoted to represent the actual safety or quality of CT practice, and (2) not offering any guidance as to how a practitioner responsible for "outlier" examinations can bring their practice to the proposed definition of compliance. Together, these can easily create signification confusion and potential disruption in the imaging practice

12. **Lack of support from manufacturers:**Having worked in dose and image quality monitoring for over a decade, academic centers of excellence, including ourselves, have a close connection with major CT manufacturers including MITA, Medical Imaging Technology Alliance, which comprises all CT manufactures. Our discussions regarding this measure lead us to believe that there will be little support from scanner manufacturers for a non-transparent and unpredictable product that lacks maturity from a private for-profit entity. There are substantial differences in image processing, detector efficiency, and such across scanners that will have significant bearing

on the CT image. The proposed measure does not account for such important nuances, leading to erroneous results.

Comment 8 by: Krishna Nallamshetty

I would like to submit a comment regarding this measure. As a practicing radiologist for greater than 15 years, we have seen tremendous growth in medical imaging that requires radiation, specifically computed tomography (CT). The public awareness of the potential long-term effects of ionizing radiation has become mainstream and as a result, a primary objective of the Americal College of Radiology and other governing bodies. The objective focuses on reducing radiation exposure as much as possible without compromising the diagnostic information that is obtained This measure evaluates radiation dose for every patient who undergoes CT *based on the clinical indication for imaging* rather than solely on the type of examination that is performed. It ensures patients receive the most appropriate CT acquisition protocol and level of radiation for their individual condition. The measure also assesses image noise, safeguarding image quality against potential effects of dose reduction, and is the first quality measure to do so.

The measure would have a large, positive impact on patients and protect them from unnecessary over-exposure of radiation without compromising the diagnostic value of medical imaging. It would be the first time a measure addresses both radiation and image quality.

Comment 9 by: Mary White

I am writing in support of this CT radiation dose safety measure. As a cancer epidemiologist, I recognize that excessive exposure to medical radiation increases cancer risk. And I understand that this measure will be valuable for protecting patients from unnecessarily high levels of radiation from CT imaging. The measure is designed to evaluate radiation dose for every patient based on the clinical indication for imaging. The measure also assesses image noise, ensuring adequate image quality despite the reduction in radiation dose. This measure fills an important quality void and has the potential to substantially reduce the contribution of CT scans to the incidence of cancer in the population.

Comment 10 by: Matthew Nielsen

I am writing in support of this important measure. The utilization of CT imaging int he United States has dramatically increased over recent decades, providing numerous benefits to patients and clinicians in the management of countless medical conditions. There has also been increasing recognition of the potential for unintended harms due to potentially avoidable variation in radiation in radiation dose for many patients. Evidence from research and quality improvement efforts demonstrates the potential to mitigate these harms witha feedback loop and benchmarking to radiologists and staff. This measure provides needed resources to disseminate these early successes, preserving the benefit of advanced imaging with CT while providing a means for healthcare facilities and clinicians to improve the safety of the studies they provide patients. The design of this measure importantly takes into account the indication for the study as the framework for dose benchmarking, with balancing measures of image quality to assure that efforts to reduce dose do not come at the expense of diagnostic quality. Given the increased recognition from patients and providers of the potential harms of imaging-associated radiation, this measure fills a timely and important gap in the current measurement portfolio.

Comment 11 by: Suz Schrandt

As a patient advocate with significant experience navigating the healthcare system--including repeated exposures to a variety of diagnostic imaging studies--I submit these comments in endorsement of this measure. The measure takes into account different contexts and parameters for a given patient and his or her unique benefit/risk profile. At a more foundational level, the measure calls into focus the significant variation in practices in CT imaging that can expose patients to unnecessary and/or unsafe levels of radiation, a risk many patients are not even aware of. The wide-spread use of this measure could standardize imaging practices and should the measure be adopted, I strongly encourage a robust dissemination plan to inform patients and families of its existence. Our ability to access safe and effective care should not be left to change; measures such as this help to close key gaps in our system.

Comment 12 by: The Leapfrog Group

Founded in 2000 by large employers and other purchasers, The Leapfrog Group is a national nonprofit organization driving a movement for giant leaps forward in the quality and safety of American health care. The flagship Leapfrog Hospital Survey collects and transparently reports hospital performance, empowering purchasers to find the highest-value care and giving consumers the lifesaving information they need to make informed decisions. For the past several year's Leapfrog has been collecting and publicly reporting hospital performance on an NQF-endorsed Pediatric CT Radiation Dose (NQF 2820) measure. The new Excessive Radiation Dose or Inadequate Image Quality for Diagnostic Computed Tomography (CT) in Adults (Clinician Group Level) fills a critical gap in evaluating radiation dose for adult patients who undergoe CT. Additionally, because the measure is based on the clinical indication for imaging – rather than on the type of examination the radiologist chose to perform – it can help ensure patients receive the right type of CT and amount of radiation for their individual condition, which is a primary concern of Leapfrog and our purchaser and employer membership. The measure also assesses image noise, safeguarding image quality against potential effects of dose reduction, and is the first quality measure to do so. Leapfrog strongly supports this measure.

Scientific Acceptability: Preliminary Analysis Form

Measure Number: *3663e*

Measure Title: *Excessive Radiation Dose or Inadequate Image Quality for Diagnostic Computed Tomography (CT) in Adults (Facility Level)*

Measure is:

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

RELIABILITY: SPECIFICATIONS

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

   Submission document: Items sp.01-sp.30

   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.**

   For example: 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?

   Reviewer 3: Overall this is a well specified measure. The main concern i have is with the technical exclusion associated with missing data; in particular, missing radiation dose. In table 2b-3, we can see that

H6 had 1,761 missing radiation dose, H14 had 408 missing radiation dose. Excluding so many CT exams could potentially have a major impact on the measure score validity.

Reviewer 5: none.

Reviewer 6: No concerns

Reviewer 7: The specification is heavily dependent on proprietary software developed by UCSF and Alara Imaging, Inc. to access and process primary data elements from the electronic systems to calculate the three variables required by the measure – CT category, size-adjusted radiation dose, and global noise. This software in turn requires access to raw imaging data. Although the developer states that this process has been tested in multiple settings, that is not evidence that a garden variety clinician could reliability replicate.

Reviewer 8: Would like to know more about the software and integrated edge device that seems to be required and/or the approach to "export from HER and radiology electronic clinical data systems via "custom reports"--what the cost or no cost alternatives might be to use this proprietary measure.

Reviewer 9: The determination of numerator ("failed value based on table of specifications by body part and size-adjusted radiation dose and global noise") is very complex. Hopefully, the developer evaluated the reliability of the "failed" determination, especially if there are higher incidents of "failed" for some body parts. In a later section the developer reports that five body regions (head, chest, cardiac, abdomen, and combined head & neck) have "low, routine, and high" radiation dose categories that were not included in the data table included in the materials. This introduces additional complications to determining failure. Time period for data collection seems inconsistent "One calendar year, although shorter periods can be used for high-volume entities." Operational definition of "high-volume" was not presented. Denominator exclusions (typically multiple areas scanned) may be problematic if these types of scans are the most common an the source of problems with too low or high dosages.

Reviewer 11: No concerns

Reviewer 12: No concerns

RELIABILITY: TESTING

Type of measure:

☐ **Process** ☒ **Process: Appropriate Use** ☐ **Structure** ☐ **Efficiency** ☐ **Cost/Resource Use**
☐ **Outcome** ☐ **Outcome: PRO-PM** ☒ **Outcome: Intermediate Clinical Outcome** ☐ **Composite**

Data Source:

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

Level of Analysis:

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

Reviewer 7: Raw images

Submission document: Questions 2a.01-09

3. **Reliability testing level**

For example: for some types of measures, if patient/encounter level validity is demonstrated, additional reliability testing is not required. Please review table above.

☒ **Accountable-Entity Level** ☒ **Patient/Encounter Level** ☐ **Neither**

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

   NOTE: "level of analysis" reflects which entity is being assessed or held accountable by the measure.

   For example: If a measure is specified for a clinician level of analysis, but facility-level testing is provided, then testing does NOT match level of analysis. Or, if two levels of analysis are specified (e.g., clinician and facility) but testing is conducted for only one, then testing does NOT match level of analysis. Or, if claims data are selected as a data source, but testing data doesn't include claims data, then testing does NOT match data source.

   Also, check "NO" if only descriptive statistics are provided or submitter only describes process for data management/cleaning/computer programming.

   ☒ **Yes** ☐ **No**

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

   According to current guidance patient/encounter level validity testing can be used for patient/encounter level reliability testing. Answer ONLY if you responded "Neither" on question #3 and/or "No" to question #4. Note that for some types of measures, additional reliability testing is not required IF patient/encounter level validity is demonstrated.

   ☐ **Yes** ☒ **No**

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

   Submission document: Question 2a.10

   For example: Is the method(s) appropriate? If not, please explain (and offer potential alternatives if possible). Does the testing conform to NQF criteria and guidance? Was testing was conducted with the data source and level of analysis indicated for this measure? Address each level of testing provided, and each analysis under each method.

   Reviwer 1: acceptable

   Reviewer 3: The developer calculated the split-half reliability for reliability testing. The method as described seems appropriate.

   Reviewer 4: Split-sample reliability testing: ICC 0.99

   Reviewer 5: Appropriate method. Calculated the ICC on a random split-half sample.

   Reviewer 6: Split half ICC was used

   Reviewer 8: Measure score reliability was estimated at the Facility level using the intraclass correlation coefficient (ICC), using randomly split samples for each accountable entity with 1,000 repetitions, applying a one-way random effects model, assuming that both entity effects and residual effects are random, independent, and normally distributed with mean 0. The Spearman-Brown prophecy formula was applied to adjust reliability from one-month test samples to the anticipated 12-month sample (i.e., (12*r)/(1 + (11*r)). These ICC(1) estimates (bounded between 0 and 1) were then logit-transformed and used to model the linear relationship between entity volume and logit reliability. By ranking predicted reliabilities across the complete range of potential volumes, the volume threshold that would correspond to ICC(1)=0.9 for an accountable entity was estimated.

   Reviewer 9: Intraclass correlations coefficient (ICC) was used. Description of the actual calculation methodology was vague ("we estimated the measure score reliability…"). The logit-transformed process was cryptic.

   Reviewer 10: ICC

Reviewer 11: Testing was conducted with the data source at the level of analysis indicated for this new measure and is appropriate.

Reviewer 12: estimated measure score reliability at the accountable entity 0.99 level using the intraclass correlation coefficient (ICC), a reliability coefficient that conceptually represents the true (between-entity) variance in a measure divided by the sum of true variance and error (within-entity) variance. Used randomly split samples for each accountable entity with 1,000 repetitions, applying a one-way random effects model, assuming that both entity effects and residual effects are random, independent, and normally distributed with mean 0. This approach corresponds to Case 1 or the ICC

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

   Submission document: Question 2a.11

   For example: Is the test sample adequate to generalize for widespread implementation? Is there high or moderate confidence that the measure results and/or the data used in the measure are reliable? Address each level of testing provided, and each analysis under each method.

   Reviewer 1: very high

   Reviewer 3: Consistent with the high volume of both inpatient and outpatient CT exams at hospital level, the reliability testing results were excellent. However, at the bottom of Table 2a-3, the developer mentioned "based on the method described above, a minimum of 28 CT exams are required to achieve 90% reliability." The developer should provide additional information to describe how this number was arrived at.

   Reviewer 4: Reliability is acceptable

   Reviewer 5: Predicted reliability for 12 months exceeded 0.99 for every hospital. According to the scale developed by Koo and Li, an ICC estimate greater than 0.90 may be interpreted as excellent reliability. (Koo 2016)

   Reviewer 6: The mean split half ICC was 0.99. This is good reliability.

   Reviewer 8: The estimated mean split-half ICC using 37,172CT exams collected from 16 Facilities was 0.99 (after Spearman-Brown adjustment to a 12-month data collection period). The number of exams per hospital for the one month time used in testing ranged from 625 to 6,157; predicted reliability for 12 months was 0.99 for every facility.

   Reviewer 9: If the calculation methodology is correct, then the reported reliability values are impressive (>0.9).

   Reviewer 10: 0.99

   Reviewer 11: Testing methods are appropriate.

   Reviewer 12: The estimated mean split-half ICC using 37,172 CT exams collected from 16 hospitals was 0.99 (after Spearman-Brown adjustment to a 12-month data collection period). The number of exams per hospital in the one month of data used for testing ranged from 625 to 6,157 (mean=2,323); predicted reliability for 12 months exceeded 0.99 for every hospital.

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: Question 2a.10-12

   For example: Appropriate signal-to-noise analysis; random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.

   ☒ **Yes**

   ☐ **No**

   ☐ **Not applicable**

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

   Submission document: Question 2a.10-12

   For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements

   Answer NO if: only assessed percent agreement; did not assess separately for all critical data elements (or at minimum, for numerator, denominator, exclusions)

   ☐ **Yes**

   ☒ **No**

   ☒ **Not applicable** (patient/encounter level 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** accountable-entity level testing has been conducted)

   ☒ **Moderate** (NOTE: Moderate is the highest eligible rating if accountable-entity 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.**

   Reviewer 3: The resulting ICC is high and as expected for a measure like this, binary process measure with large denominator.

   Reviewer 4: Split-sample reliability testing: ICC 0.99

   Reviewer 5: Used appropriate method for testing measure score reliability. Results demonstrated excellent reliability.

   Reviewer 9: Given the lack of specificity in the description, the reported results may or may not be correct. The rating is a "benefit of the doubt" value.

   Reviewer 10: Appropriate test and results.

   Reviewer 11: Testing results.

   Reviewer 12: No concerns, Very sound statistical analysis

VALIDITY: TESTING

12. **Validity testing level (check all that apply):**

   ☒ **Accountable-Entity Level** ☒ **Patient or Encounter-Level** ☒ **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*:* Questions 2b.01-02.

   For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.

   Answer NO if: only assessed percent agreement; did not assess separately for all critical data elements (or at minimum, for numerator, denominator, exclusions)

   ☒ **Yes**

   ☐ **No**

☐ **Not applicable** (patient/encounter level testing was not performed)

14. **Method of establishing validity at the *accountable-entity level:***

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

    Submission document: Questions 2b.01-02

    ☒ **Face validity**

    ☒ **Empirical validity testing at the accountable-entity level**

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

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

    Submission document: Question 2b.02

    For example: Correlation of the accountable-entity level on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score

    ☒ **Yes**

    ☐ **No**

    ☒ **Not applicable** (accountable-entity level testing was not performed)

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

    Submission document: Question 2b.02

    For example:

    - *If face validity the only testing conducted: Was it accomplished through a systematic and transparent process, by identified experts, explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality, and the degree of consensus and any areas of disagreement provided/discussed?*
    - *If a maintenance measure, but no empirical testing conducted, was justification provided?*
    - *If construct validation conducted, was the hypothesized relationship (including strength and direction) described and does it seem reasonable?*

    Reviewer 1: robust
    Reviewer 3: Data element validity testing is by comparing eCQM computation to medical record review. Measure score face validity is ascertained via TEP survey.
    Reviewer 4: Established face validity using TEP – very high level of agreement with questions on face validity. Data element validity – accuracy of measure algorithm to assign CT category had 95% accuracy. Should have used Kappa analysis or sensitivity/specificity instead. The eCQM computed identical results for a sample of 8,000 CT exams, compared to medical record review.
    Reviewer 5: Data-element: For the different data elements, relied on prior literature or comparisons to a gold standard. Score-level: Relied on face validity.
    Reviewer 6: No concerns
    Reviewer 8: CT category: The measure uses an algorithm to assign each CT exam to one of 18 CT categories based on the diagnosis associated with the exam order (codified in ICD-10-CM codes) and procedure performed (codified in CPT® codes). Developers used criterion validity to compare agreement between the CT category assigned using this method versus a gold standard method based on expert review of the complete medical record.

Patient size: Methods for measuring patient diameter on CT images have been previously validated including measuring patient size on axial and coronal images. Developer relied on published work and tested how often this method generated clinically plausible and non-missing values for size in testing data.

Radiation Dose: The measure uses a standardized data element, generated by virtually (>99%) all CT machines, that is well validated and used broadly to reflect the radiation dose delivered to the patient. The proposed measure adjusted DLP for patient size to ensure that differences in patient mix would not result in differences in measure scores across reporting entities. Developers relied on this published work and tested how often this method generated clinically plausible and non-missing values for radiation dose in testing data.

Size-Adjusted Radiation Dose: When out-of-range rates are unadjusted for patient size, observed failure rates are strongly associated with size, with almost all failures occurring in larger patients. When failure rates are adjusted for size, there is no association. Using field testing data, developers assessed whether we could calculate size-adjusted radiation dose within a plausible range and quantified missing data.

Global noise: Adapted previously validated approaches. Developer assessed whether they could calculate global noise within a plausible range and quantified missing data using field-testing data.

They also calculated the correlation between global noise and physician dissatisfaction with image quality using data from the Image Quality Study and explored the rate of physician dissatisfaction in CT exams that exceeded global noise thresholds.

Thresholds for "out-of-range" values to define numerator: Radiologists' satisfaction with CT images was used as a basis for establishing the maximum radiation dose and minimum image quality thresholds for each CT category.

Empirical validity testing: validated the eCQM output (encounter-level validity) against medical record review using field testing data collected from electronic clinical data systems from 8 health systems/vertically integrated organizations.

Accountable entity-level (measure score) validity was tested using systematic assessment of face validity of measure score as an indicator of quality through a 6-question poll to the Technical Expert Panel (TEP) assembled for the creation of this measure. The TEP represents a diverse group of clinicians (N=10), patient advocates (N=2), and leaders of medical specialty societies, payers, and healthcare safety and accrediting organizations. TEP members were identified by reaching out to key stakeholder organizations and advocates and identifying researchers who had contributed to the relevant literature.

Reviewer 9: Face validity method produced a very high level of agreement that the measure and its components were valid.

Reviewer 10: Criterion validity, comparisons with literature, comparisons with field testing data, correlation, face validity, medical record review

Reviewer 11: Face validity via expert panel and abstraction from EMR used. No inter-rater correlation (kappa score). Used NLP but validated using abstraction.

Reviewer 12: Empirical and face validity performed

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

Submission document: Questions 2b.03-04

For example: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient validity so that conclusions about quality can be made? Do you agree that the score from this measure as specified is an indicator of quality?

Reviewer 1: no concerns

Reviewer 3: TEP survey produced complete support on the face validity of this measure. eCQM computation led to identical results compared with medical record review for a sample of 7,000 CT exams.

Reviewer 4: Established face validity using TEP – very high level of agreement with questions on face validity. Data element validity – accuracy of measure algorithm to assign CT category had 95% accuracy. Should have used Kappa analysis or sensitivity/specificity instead. The eCQM computed identical results for a sample of 8,000 CT exams, compared to medical record review.

Reviewer 5: Data-element: Mean correct classification of exam was 92% Score-level: 100% of members agreed that measure was a valid measure of quality

Reviewer 6: No concerns

Reviewer 7: The results were generally consistent with high validity

Reviewer 8: CT category: Results, weighted by the distribution of CT categories in the UCSF International CT Dose Registry, were: sensitivity = 0.86 and specificity = 0.96 (n=978 CT exams). When tested across the 16 facilities, the correct classification rate of the assignment of CT exams to CT category in field-testing was 92% on average and varied from 88-97% across groups.

Size-Adjusted Radiation Dose: In field testing data, size-adjusted radiation dose could be calculated and was within plausible range for 99% of CT exams and was missing for 0.4% of exams.

Global Noise: Global noise could be calculated and was within a plausible range for 100% of CT exams in field-testing. Global noise was missing for 0.01% of examinations. The correlation between noise and physician dissatisfaction with image quality is 0.37 overall based on the image quality study (n=727 CT exams).

Based on the field-testing data, there were few exams which exceeded the global noise thresholds. There were 4 CT categories with exams in which global noise exceeded the allowable threshold. For other CT categories, exams were not observed above the threshold.

Empirical Validity Testing: The results of the medical record review were compared with the results of the eCQM computation by selecting a sample of exams (N=7,000) representative of exams generated by the 16 hospitals across the 7 health systems. The out-of-range results (measure score) from the medical record review and the eCQM computation were identical with no discrepancies between the two approaches, indicating a correct and robust implementation of the measure logic.

Face validity results were very strong with items having 100% agreement.

Reviewer 9: Face validity method produced a very high level of agreement that the measure and its components were valid."

Reviewer 10: Agreement from TEP, low correlation .37, 92% data element validity

Reviewer 11: Sufficient validity.

Reviewer 12: Yes, no concerns

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

Submission document: Questions 2b.15-18.

For example: Are there exclusions? If so, are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? Are any patients or patient groups inappropriately excluded from the measure? If patient preference (e.g., informed decision-making) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent? If you have concerns based on a clinical rationale, please note here as well as in question #29.

Reviewr 3: As indicated earlier, missing data technical exclusion should be looked at carefully. If one facility has too many missing radiation dose, for example, it may be more appropriate to not report on this facility instead of relying on non-missing data.

Reviewer 5: Data-element: Mean correct classification of exam was 92% Score-level: 100% of members agreed that measure was a valid measure of quality

Reivewer 6: No concerns

Reviewer 11: No concerns.

Reviewer 12: No concerns, exclusions appropriate

19. **Risk Adjustment**

Submission Document: Questions 2b.19-32

Applies to all outcome, cost, and resource use measures. Please answer all checkbox questions (19a -19d), then elaborate on your answers in your response to 19e.

19a. **Risk-adjustment method**

☒ None ☒ Statistical model ☒ Stratification

☒ Other method assessing risk factors (please specify)

Reviewer 3: Risk models were used to derive size correction coefficients.

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**

For example: If measure is risk adjusted:

- *If the developer asserts there is no conceptual basis for adjusting this measure for social risk factors, do you agree with the rationale?*
- *How well do social risk factor variables that were available and analyzed align with the conceptual description provided?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)?*
- *If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision?*
- *Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)?*
- *Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*

If measure is NOT risk-adjusted:

- *Is a justification for not risk adjusting provided (conceptual and/or empirical)?*
- *Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

Reviewer 1: The measure is not risk adjusted in the usual sense. Patient size is used in the calculation of the outcome, not risk adjusted afterwards.

Reviewer 3: I would defer this to TEP as this is not a typical risk adjustment issue.

Reviewer 4: Risk adjustment model is not intended as a predictive model, but only to adjust for need to use higher radiation does to adequately image larger structures and patients. Unclear to me why the Rsquared value for the model should not be used to assess model performance. Nor is it clear to me why they did not assess model performance using entire data set which included all CT body regions and patient weights.

Reviewer 5: Adjusted for patient size. R-squared close to zero for almost all CT categories.

Reviewer 6: No concerns

Reviewer 7: Although the approach is described as "risk adjustment" it is really the definition of the outcome variable that happens to vary based on a patient characteristics. The results would be uninterpretable without it.

Reviewer 11: Agree with risk-adjustment approach.

Reviewer 12: Justification for not including social risk factor. Only risk factor is size

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

Submission document: Questions 2b.05-07

For cost/resource use measures, does this measure identify meaningful differences about cost and resource use between the measured entities?

Reviewer 3: This range of measure scores is wide. This measure can identify meaningful differences among facilities.

Reviewer 5: None. Across the 16 hospitals, found 24-57% of exams to be out-of-range (+/-5% of national mean).

Reviewer 6: No concerns

Reviewer 7: Although there is variability in performance whether these results are clinically meaningful to the patient is not directly addressed

Reviewer 9: Meaningful differences description was confusing. Simplify the presentation and emphasize the # of clinicians who meet the minimum # of produced CTs, then the % who fail low, high by grouped # of CTs, etc.

Reviewer 10: None

Reviewer 11: No concerns.

Reviewer 12: None

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

Submission document: Questions 2b.11-14.

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****. 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.*

Note if not applicable. Note if applicable but not addressed. If multiple sets of specification (e.g., due to different data sources or methods of data collection): Do analyses indicate they produce comparable results?

Reviewer 3: no concern

Reviewer 5: Not applicable.

Reviewer 10: None

Reviewer 12: None

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

Submission document: Questions 2b.08-10.

For example: Are there any sources of missing data not considered? Is it clear how missing data are handled? Is missing data more of a problem for some providers or patients than others? Does the extent of missing data impact the validity of the measure?

Reviewer 3: I would recommend a missing data threshold on key data elements for reporting.

Reviewer 5: None. 92% of exams had full data. Missing data appears to be within the control of the accountable entity.

Reviewer 6: No concerns

Reviewer 7: There was significant missing data even among study hospital that had all the advantages of mentoring by the study team. The "real world" level of missing data is likely to be much higher.

Reviewer 9: The data seem dependent upon installing software package. If we endorse the measure, are we imposing the cost of this software package on all entities that produce CT scans? Additionally, the developer raised the issue of the cost of CT hardware for poorer communities.

Reviewer 10: None

Reviewer 11: No concerns. Approve of the developer's approach to missing data.

Reviewer 12: No concerns, missing data was small and exclusion of cases were appropriate

For cost/resource use measures ONLY:

If not cost/resource use measure, please skip to question 25.

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

Consider these specific aspects of the measure specifications: attribution, cost categories, target population.

☐ **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):**

Attribution: Does the accountable entity have reasonable control over the costs/resources measured? Is this approach aspirational (intending to drive change) or was it developed based on current state?

Costing Approach: Do the cost categories selected align with the measure intent, target population and care settings? Is the approach for assigning dollars to resources

Carve Outs: Has the developer addressed how carve outs in the data source are handled (or should be handled for other users)? For example, if pharmacy data is carved out (missing) from the data set, can a measure that focuses on cost of care for asthmatics still be valid?

Truncation (approach to outliers): What is the threshold for outliers (i.e., extremely high cost or low cost cases) and how are they handled?

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 accountable-entity level testing has been conducted)

☒ **Moderate** (NOTE: Moderate is the highest eligible rating if accountable-entity 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 accountable-entity level and the patient/encounter 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.**

Reviewer 3: Data element validity results are excellent. Face validity survey results are also strong.
Reviewer 4: Based on TEP.
Reviewer 5: 100% of expert panel thought measure did a good job of capturing quality. Measure developer took reasonable steps to validate the validity of data elements.
Reviewer 7: There are several statements in the submission which seem to contradict clinician level validity: technical decisions on how to perform CT are made at the facility level rather than at the individual patient level. Because decisions are made at the level of patient groups, rather than individual patients, the logic model does not include varying technical parameters for individual patients Given that this measure is an eCQM, no patient-reported data were collected. Therefore, social risk factors were not available and not analyzed (this sentence just doesn't make sense)
Reviewer 9: The rating is based on the strong Face Validity results and the fact that this is a new measure.
Reviewer 10: Appeared to use what was available when gold standards weren't feasible; multiple methods
Reviewer 11: Based on the methods used and the validation results.

For composite measures ONLY

If not composite, please skip this section.

Submission documents: Questions 2c.01-08

Examples of analyses:

1) If components are correlated - analyses based on shared variance (e.g., factor analysis, Cronbach's alpha, item-total correlation, mean inter-item correlation).

2) If components are not correlated - analyses demonstrating the contribution of each component to the composite score (e.g., change in a reliability statistic such as ICC, with and without the component measure; change in validity analyses with and without the component measure; magnitude of regression coefficient in multiple regression with composite score as dependent variable, or clinical justification (e.g., correlation of the individual component measures to a common outcome measure).

3) Ideally, sensitivity analyses of the effect of various considered aggregation and weighting rules and the rationale for the selected rules; at a minimum, a discussion of the pros and cons of the considered approaches and rationale for the selected rules.

4) Overall frequency of missing data and distribution across providers. Ideally, sensitivity analysis of the effect of various rules for handling missing data and the rationale for the selected rules; at a minimum, a discussion of the pros and cons of the considered approaches and rationale for the selected rules.

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?**

For example: Do the component measures fit the quality construct and add value? Are the objectives of parsimony and simplicity achieved while supporting the quality construct? Do analyses demonstrate the aggregation and weighting rules fit the quality construct and rationale?

☐ **High**

☐ **Moderate**

☐ **Low**

☐ **Insufficient**

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

ADDITIONAL RECOMMENDATIONS

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

Scientific Acceptability

Evaluating Scientific Acceptability: Instructions

Scientific Acceptability: 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 subcriteria for both reliability and validity to pass this criterion.

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating. Relevant measure documents are at the bottom of the SharePoint site.
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form **if your measure is a composite**.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***examples in italics*** to help you answer them.
- ***Please refer to the 2021*** Measure Evaluation Criteria and Guidance document ***(pages 17-27) and the 2-page*** Key Points document ***when evaluating your measures***.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff as soon as possible (methodspanel@qualityforum.org). Is it possible that we can obtain the needed information, but only if requested in a timely manner.
- ***Remember*** that testing at either the patient/encounter level **OR** the accountable-entity level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures. Please review the box below to guide your rating.
- If a measure you are evaluating includes multiple measures (e.g., the Hopsital CAHPS measure submsission acutally includes 11 performance measures), all included measures must be rated. You may decide that one rating applies to all included measures, or you may need to provide separate ratings (e.g., if results are substantially better for one measure than for another).

Measure type	**Requirements for reliability testing**	**Requirements for validity testing**
Instrument-based measures	BOTH patient/encounter and accountable-entity level testing	BOTH patient/encounter and accountable-entity level testing
Composite measures	Accountable-entity, score-level testing of the composite measure score and testing of the individual components. Testing of the components is not sufficient.	At initial endorsement only, empirical or face validity testing of the components OR face validity of the composite is acceptable. At maintenance, accountable-entity level testing of the composite measure score is desired.

Measure type	Requirements for reliability testing	Requirements for validity testing
eCQMs	All eCQMs must be tested using the eCQM specifications. These must use the latest industry accepted eCQM technical specifications: health quality measure format (HQMF), Quality Data Model (QDM), Clinical Quality Language (CQL), and value sets vetted through the National Library of Medicine's Value Set Authority Center (VSAC). Reliance on data from structured data fields is expected; otherwise, unstructured data must be empirically shown to be both reliable and valid (and this must be tested empirically). Thus, testing for elements that are not included in structured data fields should be tested at the patient/encounter/data element level.	All eCQMs must be tested using the eCQM specifications. These must use the latest industry accepted eCQM technical specifications: health quality measure format (HQMF), Quality Data Model (QDM), Clinical Quality Language (CQL), and value sets vetted through the National Library of Medicine's Value Set Authority Center (VSAC). Reliance on data from structured data fields is expected; otherwise, unstructured data must be empirically shown to be both reliable and valid (and this must be tested empirically). Thus, testing for elements that are not included in structured data fields should be tested at the patient/encounter/data element level. Empirical testing is expected, and as of Summer 2019, data element validation will be required unless justification is provided/accepted. Face validity alone will not be sufficient. Use of a simulated data set (e.g., BONNIE) is no longer accepted for testing validity of data elements.
Cost and Resource Use Cost and Resource Use Measure Evaluation Criteria	EITHER patient/encounter level or accountable-entity level testing	Validity is considered in the context of measure intent and threats to validity based on these cost measure-specific components: • Attribution approach • Cost categories • Approach to outliers • Impact of Carve Outs EITHER patient/encounter level or accountable-entity level testing; face validity not accepted for maintenance measures unless justification provided/accepted.
All others	EITHER patient/encounter level or accountable-entity level testing	EITHER patient/encounter level or accountable-entity level testing; face validity not accepted for maintenance measures unless justification provided/accepted; if patient/encounter level validity is demonstrated, additional reliability testing is not required

Developer Submission

1. 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

Please separate added or updated information from the most recent measure evaluation within each question response in the Importance to Measure and Report: Evidence section. For example:

2021 Submission:
Updated evidence information here.

2018 Submission:
Evidence from the previous submission here.

Evidence

1a.01. Provide a logic model.

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.

[Response Begins]

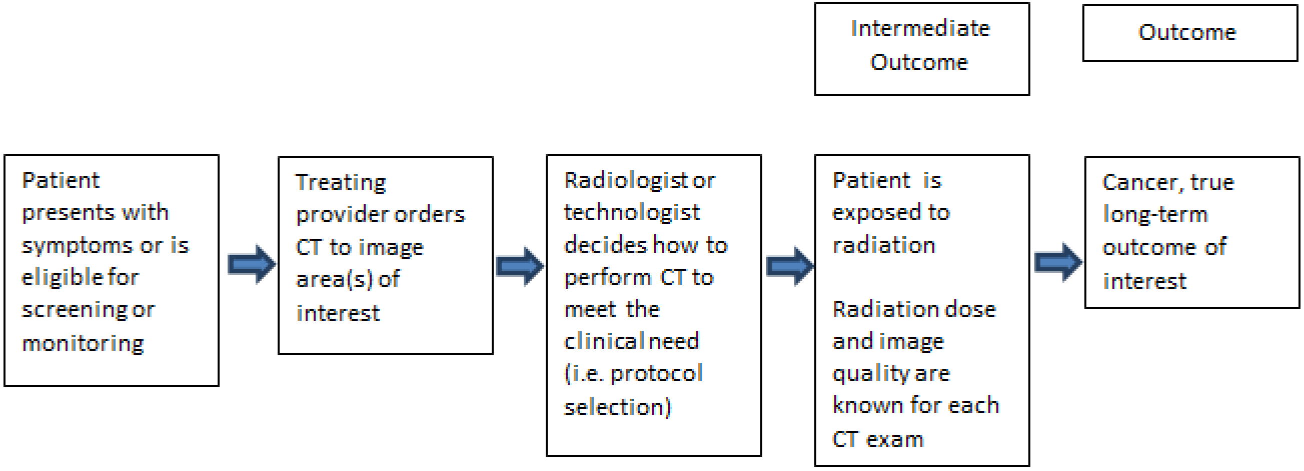

Intermediate Outcome
Outcome
Patient presents with symptoms or is eligible for screening or monitoring
Treating provider orders CT to image area(s) of interest
Radiologist or technologist decides how to perform CT to meet the clinical need (i.e. protocol selection)
Patient is exposed to radiation
Radiation dose and image quality are known for each CT exam
Cancer, true long-term outcome of interest

Figure 1a-1. Logic model demonstrating the steps and relationships between imaging based on clinical indication, the intermediate outcome (radiation dose), and the ultimate outcome of interest (cancer).

There is substantial variation in the radiation doses used for CT exams (Kanal 2017, Smith-Bindman 2009) which is primarily due to differences in how radiologists choose to perform them – in other words, their choice of a specific imaging protocol (for example, a single or multiple phase CT) and the specific technical parameters used such as scan length, milliampere-seconds, and kilovoltage peak. (Smith-Bindman 2019) More than patient or CT machine characteristics, this subjective protocol selection is the single greatest predictor of radiation dose. (Smith-Bindman 2019) However, there are no benchmarks currently available to guide practice from this point of evaluating patients with particular symptoms. In practice, patients are often assigned to a protocol that uses a higher radiation dose than the underlying indication warrants. The proposed measure directly assesses size-adjusted radiation dose and image quality used in CT exams *based on the clinical indication for imaging, shown as the first step in the process*. In this framework, the measure assesses both the earlier step of protocol selection and the later step of radiation dose (and image quality) given the protocol selected.

There is also substantial evidence (discussed later in this section) that radiation doses used for CT are carcinogenic, and that the risk of cancer is directly proportional to the doses used. Therefore, risks would be directly reduced by reducing doses. However, it is not feasible to identify the incidence of cancer associated with the physician's imaging decisions and resultant patient doses because of the potentially long lag between exposure and cancer onset. As highlighted in this application, cancer risks continue to be elevated for over 50 years after exposure. However, the cancer risk will be directly related to the radiation dose used, which is known at the time of the exam. Thus, the radiation dose for each CT exam is an intermediate outcome that can be used as a surrogate for (future) cancer risk.

[Response Ends]

1a.02. Select the type of source for 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.

[Response Begins]

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

[Response Ends]

If the evidence is not based on a systematic review, skip to the end of the section and do not complete the repeatable question group below. If you wish to include more than one systematic review, add additional tables by clicking "Add" after the final question in the group.

Evidence - Systematic Reviews Table (Repeatable)

Group 1 - Evidence - Systematic Reviews Table

1a.03. Provide the title, author, date, citation (including page number) and URL for the systematic review.

[Response Begins]

Early life ionizing radiation exposure and cancer risks: systematic review and meta-analysis.

Abalo KD, Rage E, Leuraud K, Richardson DB, Le Pointe HD, Laurier D, Bernier MO.

Pediatr Radiol. 2021 Jan;51(1):45-56. doi: 10.1007/s00247-020-04803-0.

https://link.springer.com/article/10.1007/s00247-020-04803-0

[Response Ends]

1a.04. Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the systematic review.

[Response Begins]

“CT exposure in childhood appears to be associated with increased risk of cancer (leukemia and brain tumors) while no significant association was observed with diagnostic radiographs.” Although the benefits of diagnostic radiation examinations may outweigh the risks associated with the doses delivered by these procedures (benefits were not evaluated in the studied patients), the results of this analysis justify continued efforts to optimize doses to patients.
[Response Ends]

1a.05. Provide the grade assigned to the evidence associated with the recommendation, and include the definition of the grade.

[Response Begins]
Newcastle-Ottawa Scale (NOS) for studies of radiation exposure in children = 7 to 9

The NOS assesses the quality of non-randomized studies, using 8 items grouped into 3 domains (I.e., selection, comparability/confounding, and outcome/exposure assessment), with 9 being the best possible score. NOS scores of 6 to 9 equate with “good quality” in the Agency for Healthcare Research and Quality (AHRQ) standards for observational studies. Good quality is the highest possible rating on the AHRQ scale.
[Response Ends]

1a.06. Provide all other grades and definitions from the evidence grading system.

[Response Begins]
The DerSimonian and Laird random-effect model was used to estimate the overall effect size to account for within- and between-study heterogeneities. The authors reported moderate heterogeneity (I^2 = 60%, p=0.03) among 6 studies of the risk of leukemia following childhood CT exposures, but no substantial alteration of the aggregate excess relative risk (ERR) with exclusion of individual studies from the meta-analysis (with one exception, where exclusion of a Dutch study led to a higher pooled ERR). There was small heterogeneity (I^2 = 32%) among 5 studies reporting on the risk of brain tumors following childhood CT exposures.

Publication and selection bias were assessed and tested using the Egger test. Some evidence of publication bias was reported (p=0.03) in the leukemia analysis, suggesting that studies of small size with negative results were less often published, but this seemed “not to be a major limitation of our analysis as demonstrated by statistical tests.” There was no evidence of publication or selection bias in the brain cancer analysis (p=0.16).
[Response Ends]

1a.07. Provide the grade assigned to the recommendation, with definition of the grade.

[Response Begins]
N/A - there is no direct recommendation
[Response Ends]

1a.08. Provide all other grades and definitions from the recommendation grading system.

[Response Begins]
N/A
[Response Ends]

1a.09. Detail the quantity (how many studies) and quality (the type of studies) of the evidence.

[Response Begins]
21 observational studies, including 11 case-control studies and 10 cohort studies, were included in the systematic review. All studies were assessed to be of good quality, with NOS scores ranging from 7 to 9. (Additional included studies looked at prenatal exposure, but the findings discussed below relate only to childhood exposure).
[Response Ends]

1a.10. Provide the estimates of benefit, and consistency across studies.

[Response Begins]
This review assesses only the risk associated with radiation exposure from medical imaging, not the benefit.
[Response Ends]

1a.11. Indicate what, if any, harms were identified in the study.

[Response Begins]
The authors report pooled excessive relative risk (ERR) per unit (Gray, Gy) of exposure for leukemia and brain tumors. ERR is the most commonly reported measure in this domain. Overall, the pooled analysis included over 11 million subjects including 437 cases of leukemia and 478 brain tumor cases. The authors observed a significant increased risk for leukemia (ERR_{pooled}=26.9 Gy^{-1}, 95% CI: 2.7–57.1), which represents an increase of 2.69% per mGy of dose over the background risk of leukemia. The pooled ERR for brain tumors was also significantly increased (ERR_{pooled}=9.1 Gy^{-1}, 95% CI: 5.2–13.1), which represents an increase of 0.91% per mGy of dose over the background risk of brain tumors. In other words, for a CT exam delivering 10 mGy to the red bone marrow, the risk of leukemia increases by about 27% over the background risk, holding all other factors constant. In 2017, this was the average bone marrow exposure from one CT in a child, and just slightly above the average bone marrow dose for an abdomen CT in an adult. For a CT exam delivering 10 mGy to the brain, the risk of brain tumor increases by about 9% over the background risk, holding all other factors constant.
[Response Ends]

1a.12. Identify any new studies conducted since the systematic review, and indicate whether the new studies change the conclusions from the systematic review.

[Response Begins]
N/A – the systematic review is from 2021.
[Response Ends]

Group 2 - Evidence - Systematic Reviews Table

1a.03. Provide the title, author, date, citation (including page number) and URL for the systematic review.

[Response Begins]
Epidemiological Studies of Low-Dose Ionizing Radiation and Cancer: Summary Bias Assessment and Meta-Analysis.

Michael Hauptmann, Robert D. Daniels, Elisabeth Cardis, Harry M. Cullings, Gerald Kendall, Dominique Laurier, Martha S. Linet, Mark P. Little, Jay H. Lubin, Dale L. Preston, David B. Richardson, Daniel O. Stram, Isabelle Thierry-Chef, Mary K. Schubauer-Berigan, Ethel S. Gilbert, Amy Berrington de Gonzalez

J Natl Cancer Inst Monogr (2020) 2020(56): lgaa010

https://academic.oup.com/jncimono/article/2020/56/188/5869934vv
[Response Ends]

1a.04. Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the systematic review.

[Response Begins]
This systematic review and meta-analysis concludes that “new epidemiological studies directly support excess cancer risks from low-dose ionizing radiation,” in the radiation dose range used in CT imaging. “Furthermore, the magnitude of the cancer risks from these low-dose radiation exposures was statistically compatible with the radiation dose-related cancer risks of the atomic bomb survivors.”
[Response Ends]

1a.05. Provide the grade assigned to the evidence associated with the recommendation, and include the definition of the grade.

[Response Begins]
No specific grading system was used, but included studies were assessed for bias in the following ways:

1. To identify bias in dose estimates, the authors "assessed the strengths and weaknesses of dosimetry systems with respect to the directness, complexity, and completeness of the dosimetry, the dosimetric uncertainty, and the validity of dose estimates."

2. In assessing the evidence for confounding and selection bias, they "summarized methods to control confounding and assessed the likelihood of uncontrolled confounding as well as its direction."

3. They "reviewed the possible impact of differential outcome ascertainment across radiation dose levels, and considered loss to follow-up, under- or over ascertainment of cancer outcomes, misclassification of outcomes, and changing classifications over time."

4. They then "performed a summary of the assessments of different biases for each study and considered both the direction of the observed effect and the direction of the bias."

Of 26 eligible studies, 3 had known or suspected bias in dose estimates that could bias the risk estimate away from the null, and 1 study was likely biased toward the null. Various sources of confounding and selection bias were identified, but the authors could not "draw a definitive conclusion on the impact of bias adjustment with the available data." Four studies "may have had cancer ascertainment possibly differential by radiation exposure"; three of these were likely biased away from the null, and one was likely biased toward the null.
[Response Ends]

1a.06. Provide all other grades and definitions from the evidence grading system.

[Response Begins]
In performing the meta-analysis of excess relative risk (ERR), they tested for homogeneity and variance due to heterogeneity (by computing Cochran's Q and the I^2 statistic, respectively.) Heterogeneity was very low for all analyses after excluding one study that contributed significant heterogeneity.
[Response Ends]

1a.07. Provide the grade assigned to the recommendation, with definition of the grade.

[Response Begins]
N/A – there is no direct recommendation
[Response Ends]

1a.08. Provide all other grades and definitions from the recommendation grading system.

[Response Begins]
N/A
[Response Ends]

1a.09. Detail the quantity (how many studies) and quality (the type of studies) of the evidence.

[Response Begins]
There were 26 eligible human studies on low-dose radiation exposure and cancer risk. Of 22 studies on solid cancer risk, 4 positive studies with potential positive bias were excluded. Of 25 studies on leukemia risk, 5 positive studies with

potential positive bias were excluded. Following these exclusions, the authors were able to exclude bias as the cause of the positive associations between low-dose ionizing radiation and elevated cancer risk.
[Response Ends]

1a.10. Provide the estimates of benefit, and consistency across studies.

[Response Begins]
The study assesses the risk associated with radiation exposure from medical imaging, not the benefit.
[Response Ends]

1a.11. Indicate what, if any, harms were identified in the study.

[Response Begins]
For solid cancers, after excluding 4 positive studies with potential positive bias, 12 of 18 studies reported positive excess relative risks (ERR) per unit of dose. For leukemia, 17 of 20 studies were positive. For both meta-analyses, the authors rejected the null hypothesis that the median ERR per unit of radiation dose equals zero. For adulthood exposure, the meta-ERR at 100 mGy was 0.029 (95% CI = 0.011 to 0.047) for solid cancers and 0.16 (95% CI = 0.07 to 0.25) for leukemia. For childhood exposure, the meta-ERR at 100 mGy for leukemia was 2.84 (95% CI = 0.37 to 5.32). The authors concluded that the majority of studies reported positive risk estimates and that these data directly support excess cancer risks from low-dose ionizing radiation.
[Response Ends]

1a.12. Identify any new studies conducted since the systematic review, and indicate whether the new studies change the conclusions from the systematic review.

[Response Begins]
This systematic review was published in 2020; the developers are not aware of any newer studies that have changed the conclusion from this systematic review.
[Response Ends]

1a.13. If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, describe the evidence on which you are basing the performance measure.

[Response Begins]
In addition to the systematic reviews described in 1a.03-1a.12 above, further epidemiological evidence derived from literature review is provided in 1a.14 below.
[Response Ends]

1a.14. Briefly synthesize the evidence that supports the measure.

[Response Begins]
There is extensive epidemiological and biological evidence that suggests exposure to radiation in the same range as that routinely delivered by CT (10-100 milli-Sieverts, mSv) increases a person's risk of developing cancer (Board of Radiation Effects 2006, Pearce 2012, Pierce 2000, Preston 2007, Brenner 2003, Hong 2019). **It was estimated in 2009 that 2% of cancers diagnosed annually are the result of CT**; in 2019 that would amount to 36,000 cancers diagnosed each year due to the use of CT. (Berrington de Gonzalez 2009, NCI Cancer Statistics).

The relationship between exposure to radiation and cancer has been shown across a large epidemiological literature, including numerous case control studies, cohort studies including the follow up of individuals exposed to radiation from the atomic bombs, and in recent years, cohort studies showing a direct association between CT imaging and cancer risk. For example, Pearce showed that **among 178,604 children exposed to CT radiation between 1985-2002 and followed through 2008, bone marrow and brain organ doses in the range of 30-50 mGy tripled the risk of leukemia and brain cancer within 10 years.** (Pearce 2012) Far from uncommon, these absorbed radiation doses are frequently delivered by

CT imaging. (Miglioretti 2013, Stewart 2021) In the longest follow-up study of survivors of the Hiroshima and Nagasaki atomic bombings (where the median dose to survivors was 40 mSv, in the same range as a single CT exam), the survivors remain at significantly elevated risk for every cancer type through all years of follow up. (Sadakane 2019, Brenner 2020, Sakata 2019, Sugiyama 2020) Overall, more than 10% of cancers in this population are attributed to the radiation exposure.

There have been several systematic reviews, summarized above, assessing the relationship between diagnostic medical radiation exposure and cancer. Abalo et al. (2021) performed a literature search of five electronic databases covering publications from 2000 to 2019 on the relationship between medical radiation exposure in children up to age 21 and cancer. Pooled excess relative risk (ERR) was reported, representing the excess of leukemia and brain tumor risk per unit (Gray, Gy) of organ dose – this metric reflects the proportional increase in risk over the background rate of cancer (in the absence of exposure), per unit of dose. The authors observed a significantly increased risk for leukemia (ERR_{pooled}=26.9 Gy−1, 95% CI: 2.7–57.1), which represents an increase of 2.69% per mGy of dose over the background risk of leukemia. The pooled ERR for brain tumors was also significantly increased (ERR_{pooled}=9.1 Gy−1, 95% CI: 5.2–13.1), which represents an increase of 0.91% per mGy of dose over the background risk of brain tumors.

Dr. Amy Berrington De Gonzalez, Chief of Radiation Epidemiology at the National Cancer Institute, was the senior author of a second systematic review and meta-analysis of studies evaluating the association between radiation exposure and cancer. (Hauptmann 2020) The authors identified 26 studies which: 1) reported a mean dose of less than 100 mGy (corresponding to exposures used in medical imaging); 2) individualized dose estimates, risk estimates, and confidence intervals (CI) for the dose-response relationship; and 3) were published between 2006-2017. They systematically assessed the potential for bias from each primary study and performed a meta-analysis to quantify the ERR and to assess consistency across studies for all solid cancers and leukemia. For adulthood exposure, the meta-ERR at 100 mGy was 0.029 (95% CI: 0.011 to 0.047) for solid cancers and 0.16 (95% CI: 0.07 to 0.25) for leukemia. For childhood exposure, the meta-ERR at 100 mGy for leukemia was 2.84 (95% CI: 0.37 to 5.32). The authors concluded that **the majority of studies reported positive risk estimates and that these data directly support excess cancer risks from low-dose ionizing radiation.** Furthermore, the magnitude of the cancer risks from these low-dose radiation exposures was statistically compatible with the radiation dose-related cancer risks of atomic bomb survivors.

A number of cohort studies are being conducted as part of the EPI-CT study: a European pooled epidemiological study to quantify the risk of radiation-induced cancer from pediatric CT (Bernier, 2019). The full results are forthcoming, but 4 contributing country-specific portions of the cohort have been published and show positive associations between CT and cancer incidence (Table 1a-1):

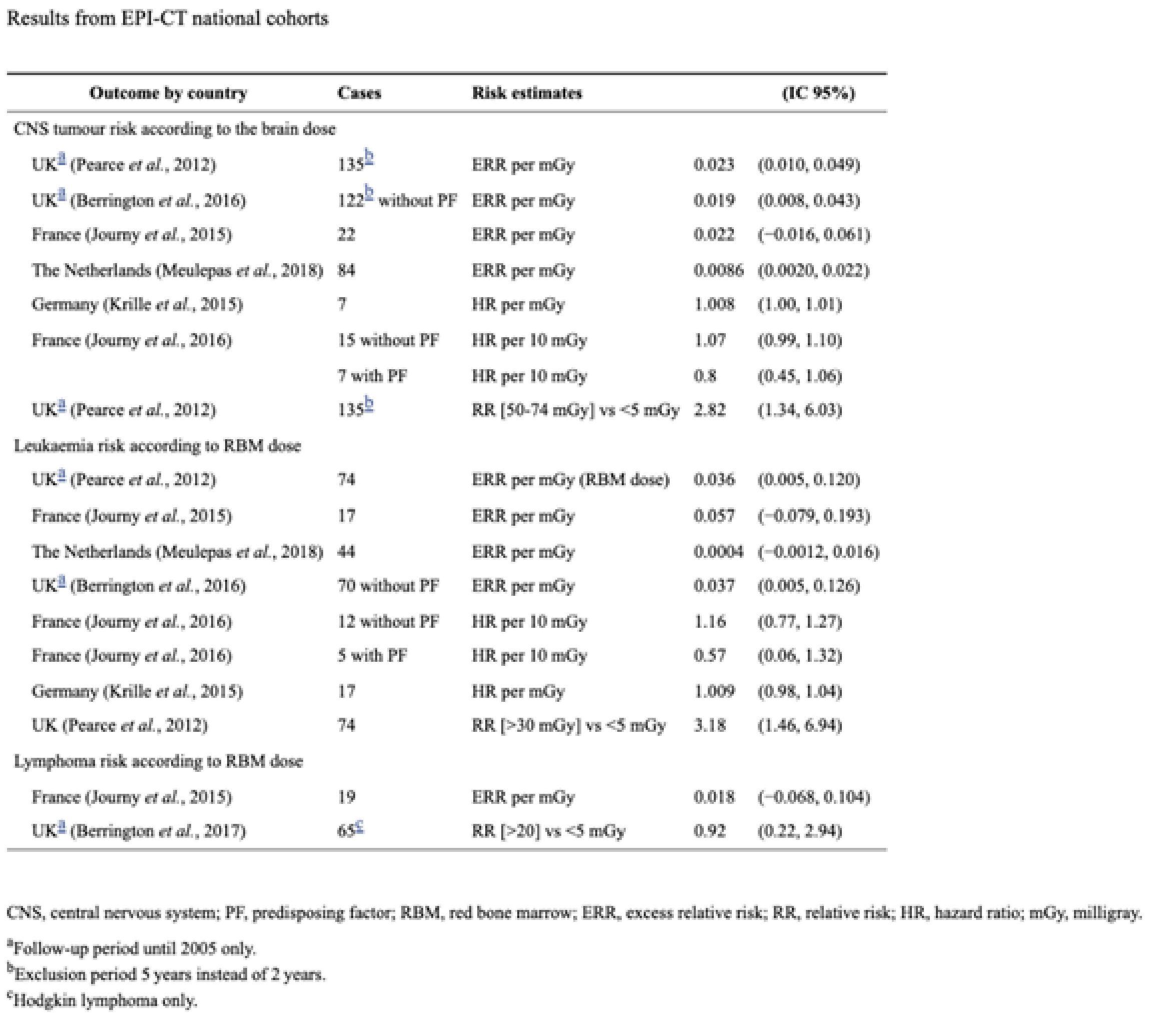
(1) **The British study reported a positive dose-response relationship between radiation dose and leukemia and CNS tumors in children and young adults.** (Pearce 2012, Berrington 2016)

(2) **The German study reported a significantly increased incidence of all cancer and lymphoma in exposed children** compared with the general population. (Krille 2015)

(3) **The French and the German cohorts reported a dose-related increase for CNS tumors**. (Journy 2015, Journy 2016, Krille 2015)

(4) **The Dutch study reported a dose-response relationship for CNS tumors.** (Meulepas 2016, Meulepas 2019)

Table 1a-1. Results from EPI CT National Cohort (Bernier 2019).



Results from EPI-CT national cohorts

Outcome by country	Cases	Risk estimates		(IC 95%)
CNS tumour risk according to the brain dose				
UK[a] (Pearce *et al.*, 2012)	135[b]	ERR per mGy	0.023	(0.010, 0.049)
UK[a] (Berrington *et al.*, 2016)	122[b] without PF	ERR per mGy	0.019	(0.008, 0.043)
France (Journy *et al.*, 2015)	22	ERR per mGy	0.022	(−0.016, 0.061)
The Netherlands (Meulepas *et al.*, 2018)	84	ERR per mGy	0.0086	(0.0020, 0.022)
Germany (Krille *et al.*, 2015)	7	HR per mGy	1.008	(1.00, 1.01)
France (Journy *et al.*, 2016)	15 without PF	HR per 10 mGy	1.07	(0.99, 1.10)
	7 with PF	HR per 10 mGy	0.8	(0.45, 1.06)
UK[a] (Pearce *et al.*, 2012)	135[b]	RR [50-74 mGy] vs <5 mGy	2.82	(1.34, 6.03)
Leukaemia risk according to RBM dose				
UK[a] (Pearce *et al.*, 2012)	74	ERR per mGy (RBM dose)	0.036	(0.005, 0.120)
France (Journy *et al.*, 2015)	17	ERR per mGy	0.057	(−0.079, 0.193)
The Netherlands (Meulepas *et al.*, 2018)	44	ERR per mGy	0.0004	(−0.0012, 0.016)
UK[a] (Berrington *et al.*, 2016)	70 without PF	ERR per mGy	0.037	(0.005, 0.126)
France (Journy *et al.*, 2016)	12 without PF	HR per 10 mGy	1.16	(0.77, 1.27)
France (Journy *et al.*, 2016)	5 with PF	HR per 10 mGy	0.57	(0.06, 1.32)
Germany (Krille *et al.*, 2015)	17	HR per mGy	1.009	(0.98, 1.04)
UK (Pearce *et al.*, 2012)	74	RR [>30 mGy] vs <5 mGy	3.18	(1.46, 6.94)
Lymphoma risk according to RBM dose				
France (Journy *et al.*, 2015)	19	ERR per mGy	0.018	(−0.068, 0.104)
UK[a] (Berrington *et al.*, 2017)	65[c]	RR [>20] vs <5 mGy	0.92	(0.22, 2.94)

CNS, central nervous system; PF, predisposing factor; RBM, red bone marrow; ERR, excess relative risk; RR, relative risk; HR, hazard ratio; mGy, milligray.

[a] Follow-up period until 2005 only.

[b] Exclusion period 5 years instead of 2 years.

[c] Hodgkin lymphoma only.

Lastly, the ongoing Life Span Study (LSS) of atomic bomb survivors in Hiroshima and Nagasaki, Japan, provides quantitative estimates of cancer risks associated with exposure to radiation and is a major source of human data used for risk assessment in establishing radiation safety standards. Although this is not a systematic review, it is the gold standard, epidemiological study of radiation in the same dose range as encountered with CT. The most recent publications describe solid cancer incidence in the LSS cohort through 2009. (Brenner 2020, Grant 2017, Sadakane 2019, Sakata 2019, Sugiyama 2020) The eligible cohort included 105,444 subjects who were alive and had no known history of cancer at the start of follow-up. The follow-up period was 1958-2009, providing 3,079,484 person-years of follow-up. Cases were identified by linkage with population-based Hiroshima and Nagasaki Cancer Registries. Poisson regression methods were used to elucidate the nature of the radiation-associated risks per Gy of weighted absorbed organ doses using both excess relative risk (ERR) and excess absolute risk (EAR) models adjusted for smoking and other covariates**. These analyses demonstrate that solid cancer risks remain elevated more than 60 years after exposure and that approximately 10% of cancers in the cohort are due to the radiation.** Studies by type of tumor confirm the strong association between radiation exposure and particular cancer types such as CNS tumors (Braganza, 2012 and Brenner, 2020), upper gastrointestinal tract tumors (Sakata, 2019) and liver and pancreas tumors (Sadakane, 2019) and colon tumors (Sugiyama, 2020)

There is also increasing understanding of the mechanisms involved in carcinogenesis. In a prospective evaluation of 67 adults undergoing cardiac CT, patients underwent extensive blood work just prior to and following the exam to look for cellular processes implicated in carcinogenesis. (Nguyen, 2015) Immunohistochemistry and full gene sequencing were performed, and diverse markers of DNA damage, repair, and cell death were evaluated. The average exposure from a single CT exam was 30 mSv (similar to the Hiroshima and Nagasaki exposures), and there was a three-fold increase in markers of DNA damage and cell death. These changes were seen at doses of 7 mSv and greater, and these changes persisted for at least a month.

Despite the known risks of CT, its use has grown substantially over the last few decades (Harvey L Neiman 2017), with 91.4 million CT exams performed in the United States in 2019 (IMV 2020), including 428 exams per 1000 patients aged 65 years and older (Smith-Bindman 2019). The radiation doses used for CT exams are frequently far higher than needed for diagnosis and have been shown to vary up to 200-fold across facilities for patients imaged for the same clinical reason. (Smith-Bindman 2009, Smith-Bindman 2015, Smith-Bindman 2019, Miglioretti 2013, Demb 2017). For example, the American College of Radiology reported that CT exams to assess kidney stones had an average dose of 10 mSv, while the optimum dose is 2-4 mSv. (Lukasiewicz, 2014) In a prospective randomized trial of different imaging strategies for patients with suspected kidney stones, 5% of patients received an appropriate dose of 4 mSv or less. (Smith-Bindman, 2014)

Evidence of the association between medical imaging and cancer risk has been reviewed by many professional societies and government, quality, and oversight organizations, which have all identified CT radiation dose reduction as a safety imperative and issued guidelines asking radiologists to track, optimize, and lower CT radiation doses. These organizations include: the American College of Radiology (Kanal 2017); the Radiology Society of North America (Hricak 2010); The Society of Interventional Radiology (Stecker 2009); The Society of Cardiovascular CT (Halliburton 2011); Cardiovascular Imaging Societies (Writing Committee 2018); Image Wisely (a joint initiative of the American College of Radiology, Radiological Society of North America, American Society of Radiological Technologists, and American Association of Physicists in Medicine); and the FDA (US Food and Drug Administration 2019).

[Response Ends]

1a.15. Detail the process used to identify the evidence.

[Response Begins]
The evidence was obtained through comprehensive searches or PubMed, Embase, and Web of Science from inception to August 2021. Each search consisted of Medical Imaging, Cancer and Epidemiology concept blocks with additional search terms including Computed Tomography and CT. References of all publications were searched to identify additional publications. Additionally, there are a small number of investigators who lead studies in this area (such as Dr. Amy Berrington De Gonzales, Chief of Radiation Epidemiology at the NCI and Dr. Alina Brenner at the Radiation Effects Research Foundation) whose names were added to searches.
[Response Ends]

1a.16. Provide the citation(s) for the evidence.

[Response Begins]

1. Abalo KD, Rage E, Leuraud K, Richardson DB, Le Pointe HD, Laurier D, Bernier MO. Early life ionizing radiation exposure and cancer risks: systematic review and meta-analysis. Pediatr Radiol 2021;51(1):45-56. doi: 10.1007/s00247-020-04803-0
2. Bernier MO, Baysson H, Pearce MS, Moissonnier M, Cardis E, Hauptmann M, Struelens L, Dabin J, Johansen C, Journy N, Laurier D, Blettner M, Le Cornet L, Pokora R, Gradowska P, Meulepas JM, Kjaerheim K, Istad T, Olerud H, Sovik A, Bosch de Basea M, Thierry-Chef I, Kaijser M, Nordenskjold A, Berrington de Gonzalez A, Harbron RW, Kesminiene A. Cohort Profile: the EPI-CT study: a European pooled epidemiological study to quantify the risk of radiation-induced cancer from paediatric CT. Int J Epidemiol 2019;48(2):379-381g. doi: 10.1093/ije/dyy231
3. Berrington de Gonzalez A, Mahesh M, Kim KP, et al. Projected cancer risks from computed tomographic scans performed in the United States in 2007. Arch Intern Med. 2009;169(22):2071-2077.
4. Berrington de Gonzalez A, Salotti JA, McHugh K, Little MP, Harbron RW, Lee C, Ntowe E, Braganza MZ, Parker L, Rajaraman P, Stiller C, Stewart DR, Craft AW, Pearce MS. Relationship between paediatric CT scans and subsequent risk of leukaemia and brain tumours: assessment of the impact of underlying conditions. Br J Cancer 2016;114(4):388-394. doi: 10.1038/bjc.2015.415
5. Berrington de Gonzalez A, Daniels RD, Cardis E, et al. Epidemiological Studies of Low-Dose Ionizing Radiation and Cancer: Rationale and Framework for the Monograph and Overview of Eligible Studies. J Natl Cancer Inst Monogr. 2020;2020(56):97-113.

6. Board of Radiation Effects Research Division on Earth and Life Sciences National Research Council of the National Academies. Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2, Washington, D.C.: The National Academies Press; 2006.
7. Braganza MZ, Kitahara CM, Berrington de Gonzalez A, Inskip PD, Johnson KJ, Rajaraman P. Ionizing radiation and the risk of brain and central nervous system tumors: a systematic review. Neuro Oncol 2012;14(11):1316-1324. doi: 10.1093/neuonc/nos208
8. Brenner AV, Sugiyama H, Preston DL, Sakata R, French B, Sadakane A, Cahoon EK, Utada M, Mabuchi K, Ozasa K. Radiation risk of central nervous system tumors in the Life Span Study of atomic bomb survivors, 1958-2009. Eur J Epidemiol 2020;35(6):591-600. doi: 10.1007/s10654-019-00599-y
9. Brenner DJ, Doll R, Goodhead DT, et al. Cancer risks attributable to low doses of ionizing radiation: assessing what we really know. Proc Natl Acad Sci U S A. 2003;100(24):13761-13766.
10. Demb J, Chu P, Nelson T, et al. Optimizing Radiation Doses for Computed Tomography Across Institutions: Dose Auditing and Best Practices. JAMA Intern Med. 2017;177(6):810-81
11. Grant EJ, Brenner A, Sugiyama H, Sakata R, Sadakane A, Utada M, Cahoon EK, Milder CM, Soda M, Cullings HM, Preston DL, Mabuchi K, Ozasa K. Solid Cancer Incidence among the Life Span Study of Atomic Bomb Survivors: 1958-2009. Radiation research 2017;187(5):513-537. doi: 10.1667/RR14492.1
12. Halliburton SS, Abbara S, Chen MY, et al. SCCT guidelines on radiation dose and dose-optimization strategies in cardiovascular CT. J Cardiovasc Comput Tomogr. 2011;5(4):198-224.
13. Harvey L Neiman Health Policy Institute. Harvey L Neiman Health Policy Institute. Medicare Part B Total Computed Tomography Procedures. 2017; https://www.neimanhpi.org/data_series/medicare-part-b-total-computed-tomography-procedures/#/graph/2017/2017/true, November 12, 2019.
14. Hauptmann M, Daniels RD, Cardis E, et al. Epidemiological Studies of Low-Dose Ionizing Radiation and Cancer: Summary Bias Assessment and Meta-Analysis. J Natl Cancer Inst Monogr. 2020;2020(56):188-200.
15. Hong JY, Han K, Jung JH, Kim JS. Association of Exposure to Diagnostic Low-Dose Ionizing Radiation with Risk of Cancer Among Youths in South Korea. JAMA Netw Open. 2019;2(9):e1910584.
16. Hricak H, Brenner DJ, Adelstein SJ, et al. Managing Radiation Use in Medical Imaging: A Multifaceted Challenge. Radiology. 2010.
17. Image Wisely. https://www.imagewisely.org/.
18. IMV 2019 CT Market Outlook Report, https://imvinfo.com/ct-departments-seek-workflow-improvements-to-address-increased-ct-utilization/.
19. Journy N, Rehel JL, Ducou Le Pointe H, Lee C, Brisse H, Chateil JF, Caer-Lorho S, Laurier D, Bernier MO. Are the studies on cancer risk from CT scans biased by indication? Elements of answer from a large-scale cohort study in France. Br J Cancer 2015;112(1):185-193. doi: 10.1038/bjc.2014.526
20. Journy N, Roue T, Cardis E, et al. Childhood CT scans and cancer risk: impact of predisposing factors for cancer on the risk estimates. J Radiol Prot. 2016;36(1):N1-7.
21. 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;284(1):120-133.
22. Krille L, Dreger S, Schindel R, et al. Risk of cancer incidence before the age of 15 years after exposure to ionising radiation from computed tomography: results from a German cohort study. Radiation and environmental biophysics. 2015;54(1):1-12.
23. Lukasiewicz A, Bhargavan-Chatfield M, Coombs L, Ghita M, Weinreb J, Gunabushanam G, Moore CL. Radiation Dose Index of Renal Colic Protocol CT Studies in the United States: A Report from the American College of Radiology National Radiology Data Registry. Radiology 2014;271(2):445-451. doi: 10.1148/radiol.14131601
24. Meulepas JM, Ronckers CM, Smets A, Nievelstein RAJ, Gradowska P, Lee C, Jahnen A, van Straten M, de Wit MY, Zonnenberg B, Klein WM, Merks JH, Visser O, van Leeuwen FE, Hauptmann M. Radiation Exposure From Pediatric CT Scans and Subsequent Cancer Risk in the Netherlands. J Natl Cancer Inst 2019;111(3):256-263. doi: 10.1093/jnci/djy104
25. Meulepas JM, Ronckers CM, Merks J, Weijerman ME, Lubin JH, Hauptmann M. Confounding of the association between radiation exposure from CT scans and risk of leukemia and brain tumors by cancer susceptibility syndromes. J Radiol Prot 2016;36(4):953-974. doi: 10.1088/0952-4746/36/4/953
26. Miglioretti DL, Johnson E, Williams A, et al. The use of computed tomography in pediatrics and the associated radiation exposure and estimated cancer risk. JAMA Pediatr. 2013;167(8):700-707.
27. National Cancer Institute Cancer Statistics. Accessed May 25, 2021. https://www.cancer.gov/about-cancer/understanding/statistics

28. Nguyen PK, Lee WH, Li YF, Hong WX, Hu S, Chan C, Liang G, Nguyen I, Ong SG, Churko J, Wang J, Altman RB, Fleischmann D, Wu JC. Assessment of the Radiation Effects of Cardiac CT Angiography Using Protein and Genetic Biomarkers. JACC Cardiovasc Imaging 2015;8(8):873-884. doi: 10.1016/j.jcmg.2015.04.016
29. Pearce MS, Salotti JA, Little MP, et al. Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. Lancet. 2012;380(9840):499-505.
30. Pierce DA, Preston DL. Radiation-related cancer risks at low doses among atomic bomb survivors. Radiation research. 2000;154(2):178-186.
31. Preston DL, Ron E, Tokuoka S, et al. Solid cancer incidence in atomic bomb survivors: 1958-1998. Radiation research. 2007;168(1):1-64.
32. Sadakane A, French B, Brenner AV, Preston DL, Sugiyama H, Grant EJ, Sakata R, Utada M, Cahoon EK, Mabuchi K, Ozasa K. Radiation and Risk of Liver, Biliary Tract, and Pancreatic Cancers among Atomic Bomb Survivors in Hiroshima and Nagasaki: 1958-2009. Radiation research 2019;192(3):299-310. doi: 10.1667/RR15341.1
33. Sakata R, Preston DL, Brenner AV, Sugiyama H, Grant EJ, Rajaraman P, Sadakane A, Utada M, French B, Cahoon EK, Mabuchi K, Ozasa K. Radiation-Related Risk of Cancers of the Upper Digestive Tract among Japanese Atomic Bomb Survivors. Radiation research 2019;192(3):331-344. doi: 10.1667/RR15386.1
34. Smith-Bindman R, Lipson J, Marcus R, et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. Arch Intern Med. 2009;169(22):2078-2086.
35. Smith-Bindman R, Aubin C, Bailitz J, Bengiamin RN, Camargo CA, Jr., Corbo J, Dean AJ, Goldstein RB, Griffey RT, Jay GD, Kang TL, Kriesel DR, Ma OJ, Mallin M, Manson W, Melnikow J, Miglioretti DL, Miller SK, Mills LD, Miner JR, Moghadassi M, Noble VE, Press GM, Stoller ML, Valencia VE, Wang J, Wang RC, Cummings SR. Ultrasonography versus Computed Tomography for Suspected Nephrolithiasis. The New England Journal of Medicine. 2014;371(12):1100-1110. doi: 10.1056/NEJMoa1404446.
36. Smith-Bindman R, Kwan ML, Marlow EC, et al. Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016. JAMA. 2019;322(9):843-856.
37. Smith-Bindman R, Moghadassi M, Wilson N, et al. Radiation Doses in Consecutive CT Examinations from Five University of California Medical Centers. Radiology. 2015;277(1):134-141.
38. Smith-Bindman R, Wang Y, Chu P, et al. International variation in radiation dose for computed tomography examinations: prospective cohort study. BMJ. 2019;364:k4931.
39. Stecker MS, Balter S, Towbin RB, et al. Guidelines for patient radiation dose management. J Vasc Interv Radiol. 2009;20(7 Suppl):S263-273.
40. Stewart C, Smith-Bindman R. It Is Time to Inform Patients of Medical Imaging Risks. JAMA Netw Open. 2021 Oct 1;4(10):e2129681. doi: 10.1001/jamanetworkopen.2021.29681.
41. Sugiyama H, Misumi M, Brenner A, Grant EJ, Sakata R, Sadakane A, Utada M, Preston DL, Mabuchi K, Ozasa K. Radiation risk of incident colorectal cancer by anatomical site among atomic bomb survivors: 1958-2009. Int J Cancer 2020;146(3):635-645. doi: 10.1002/ijc.32275
42. U.S. Food and Drug Administration. FDA White Paper: Initiative to Reduce Unnecessary Radiation Exposure from Medical Imaging. 2019. https://www.fda.gov/radiation-emitting-products/initiative-reduce-unnecessary-radiation-exposure-medical-imaging/white-paper-initiative-reduce-unnecessary-radiation-exposure-medical-imaging.
43. Writing Committee M, Hirshfeld JW, Jr., Ferrari VA, et al. 2018 ACC/HRS/NASCI/SCAI/SCCT Expert Consensus Document on Optimal Use of Ionizing Radiation in Cardiovascular Imaging-Best Practices for Safety and Effectiveness, Part 2: Radiological Equipment Operation, Dose-Sparing Methodologies, Patient and Medical Personnel Protection. J Am Coll Cardiol. 2018.

[Response Ends]

Performance Gap

1b.01. Briefly explain the rationale for this measure.

Explain how the measure will improve the quality of care, and list the benefits or improvements in quality envisioned by use of this measure.

[Response Begins]

Diagnostic CT imaging occurs in more than a third of acute care hospitalizations (Vance 2013) and upwards of 90 million scans are performed annually in the U.S. (IMV 2020). The radiation doses used for these exams are frequently far higher than needed for diagnosis and vary up to 200-fold across facilities for patients imaged for the same clinical reason. (Smith-Bindman 2009, Smith-Bindman 2015, Smith-Bindman 2019, Miglioretti 2013, Demb 2017). Most of this variation reflects clinician preferences rather than appropriate differences based on patient and clinical indications (Smith-Bindman 2019). As described in section 1a.14, the inconsistency in how CT exams are performed represents a significant, unnecessary, and modifiable iatrogenic health risk, as there is extensive epidemiological and biological evidence that suggests exposure to radiation in the same range as that routinely delivered by CT increases a person's risk of developing cancer (Board of Radiation Effects 2006, Pearce 2012, Pierce 2000, Preston 2007, Brenner 2003, Hong 2019). It is estimated that 2% (36,000) of the 1.8 million cancers diagnosed annually in the U.S. are caused by CT exams (Berrington de Gonzalez 2009, NCI Cancer Statistics).

The measure focuses on reducing radiation dose in CT, an intermediate outcome important to cancer prevention. As radiation dose is known to be directly related and proportional to future cancer risk (Board of Radiation Effects 2006, Pearce 2012, Pierce 2000, Preston 2007, Brenner 2003, Hong 2019, Berrington de Gonzalez 2009), any reduction in radiation exposure would be expected to lead to a proportional reduction in cancers. Research suggests that when healthcare organizations and clinicians are provided with a summary of their CT radiation doses, their subsequent doses can be reduced without diminishing the diagnostic usefulness of these tests. Smith-Bindman et al. led a randomized controlled trial of two interventions to optimize CT radiation doses across 100 hospitals and imaging facilities and found that providing feedback to institutions along with education and opportunities for sharing best practices results in meaningful dose reductions. (Smith-Bindman 2020). Though results varied by anatomic region, following the intervention there was up to a 40% reduction in doses with a greater impact on the rate of high dose exams, meaning facilities with high doses at the beginning of the trial were particularly likely to improve.

On the basis of the current estimated number of CT exams performed annually in the U.S. (IMV 2020), distribution in scan types and observed doses (Demb 2017, Smith-Bindman 2019), modelling of the cancer risk associated with CT at different ages of exposure (Berrington de Gonzalez 2009), and costs of cancer care (Dieguez 2017, Mariotto 2011), an estimated 18,643 cancers could be prevented annually in the U.S., 75% (13,982) of these among Medicare beneficiaries, resulting in $1.86 billion to $5.21 billion in annual cost savings to the Centers for Medicare & Medicaid Services.

References

1. Vance EA, Xie X, Henry A, Wernz C, Slonim AD. Computed tomography scan use variation: patient, hospital, and geographic factors. Am J Manag Care. 2013 Mar 1;19(3):e93-9. PMID: 23534948.
2. IMV 2019 CT Market Outlook Report, https://imvinfo.com/ct-departments-seek-workflow-improvements-to-address-increased-ct-utilization/.
3. Smith-Bindman R, Lipson J, Marcus R, et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. Arch Intern Med. 2009;169(22):2078-2086.
4. Smith-Bindman R, Moghadassi M, Wilson N, et al. Radiation Doses in Consecutive CT Examinations from Five University of California Medical Centers. Radiology. 2015;277(1):134-141.
5. Smith-Bindman R, Wang Y, Chu P, et al. International variation in radiation dose for computed tomography examinations: prospective cohort study. BMJ. 2019;364:k4931.
6. Miglioretti DL, Johnson E, Williams A, et al. The use of computed tomography in pediatrics and the associated radiation exposure and estimated cancer risk. JAMA Pediatr. 2013;167(8):700-707.
7. Demb J, Chu P, Nelson T, et al. Optimizing Radiation Doses for Computed Tomography Across Institutions: Dose Auditing and Best Practices. JAMA Intern Med. 2017;177(6):810-81
8. Board of Radiation Effects Research Division on Earth and Life Sciences National Research Council of the National Academies. Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2, Washington, D.C.: The National Academies Press; 2006.
9. Pearce MS, Salotti JA, Little MP, et al. Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. Lancet. 2012;380(9840):499-505.
10. Pierce DA, Preston DL. Radiation-related cancer risks at low doses among atomic bomb survivors. Radiation research. 2000;154(2):178-186.
11. Preston DL, Ron E, Tokuoka S, et al. Solid cancer incidence in atomic bomb survivors: 1958-1998. Radiation research. 2007;168(1):1-64.
12. Brenner DJ, Doll R, Goodhead DT, et al. Cancer risks attributable to low doses of ionizing radiation: assessing what we really know. Proc Natl Acad Sci U S A. 2003;100(24):13761-13766.

13. Hong JY, Han K, Jung JH, Kim JS. Association of Exposure to Diagnostic Low-Dose Ionizing Radiation with Risk of Cancer Among Youths in South Korea. JAMA Netw Open. 2019;2(9):e1910584.
14. Berrington de Gonzalez A, Mahesh M, Kim KP, et al. Projected cancer risks from computed tomographic scans performed in the United States in 2007. Arch Intern Med. 2009;169(22):2071-2077.
15. National Cancer Institute Cancer Statistics. Accessed May 25, 2021. https://www.cancer.gov/about-cancer/understanding/statistics
16. Smith-Bindman R, Chu P, Wang Y, et al. Comparison of the Effectiveness of Single-Component and Multicomponent Interventions for Reducing Radiation Doses in Patients Undergoing Computed Tomography: A Randomized Clinical Trial. JAMA Intern Med. 2020 May 1;180(5):666-675.
17. Dieguez G, Ferro C, Pyenson B. Milliman Research Report: A Multi-Year Look at the Cost Burden of Cancer Care. April 11, 2017. https://www.milliman.com/en/insight/2017/a-multi-year-look-at-the-cost-burden-of-cancer-care
18. Mariotto AB, Yabroff KR, Shao Y, Feuer EJ, Brown ML. Projections of the cost of cancer care in the United States: 2010-2020. J Natl Cancer Inst. 2011 Jan 19;103(2):117-28. Epub 2011 Jan 12. Erratum in: J Natl Cancer Inst. 2011 Apr 20;103(8):699. PMID: 21228314.

[Response Ends]

1b.02. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.

Include mean, std dev, min, max, interquartile range, and 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 (4b) under Usability and Use.

[Response Begins]
The measure has been field tested across 7 health systems, representing 35,729 CT exams and 16 hospitals. The measure is reported at the level of the hospital (identified by CMS Certification number, CCN). They represent diverse practices with regard to community vs. academic, urban vs. nonurban care settings, and geographic location (Alabama, California, Michigan, New York). Data were collected from an approximately four-week period at each testing site, spanning the years 2020-2021.

Performance data at the hospital level is as follows:
Mean measure (out-of-range) score: 31%, standard deviation: 7%
Range: minimum = 20%, maximum = 43%
Interquartile range: 9% (27%-36%)
Measure scores by percentile:

- 10th = 21%
- 20th = 23%
- 30th = 27%
- 40th = 28%
- 50th = 30%
- 60th = 31%
- 70th = 36%
- 80th = 37%
- 90th = 40%

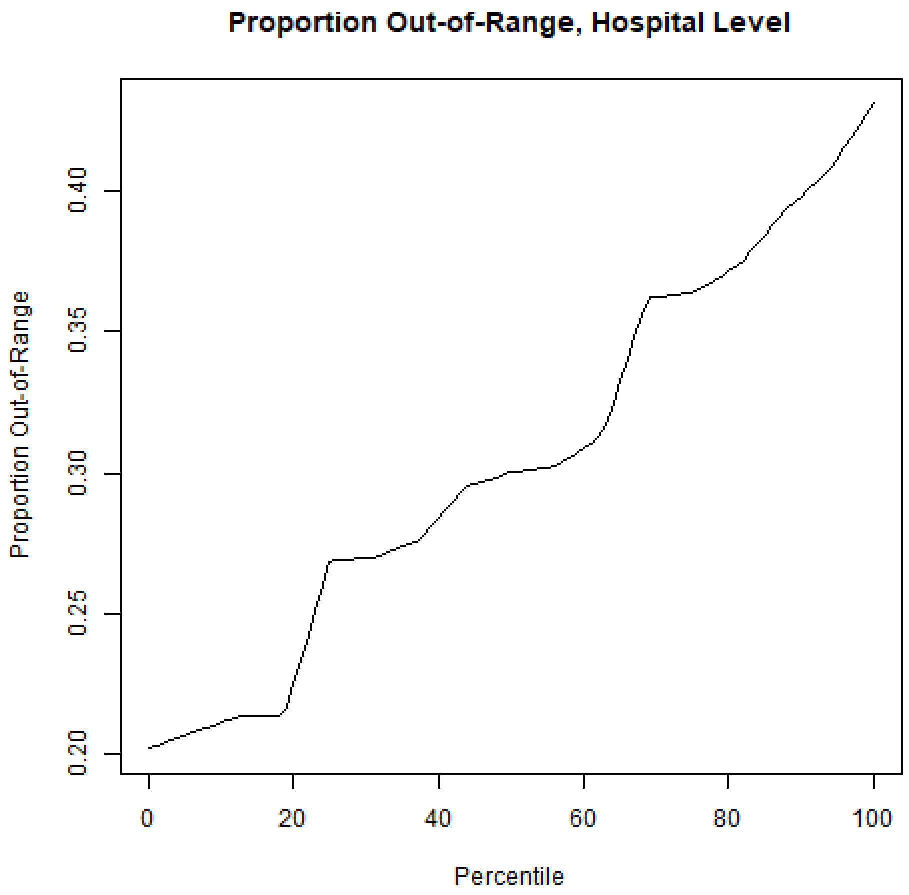


Figure 1b-1. Out-of-range scores by percentile. Hospitals with the lowest (best) out-of-range scores are on the bottom left.

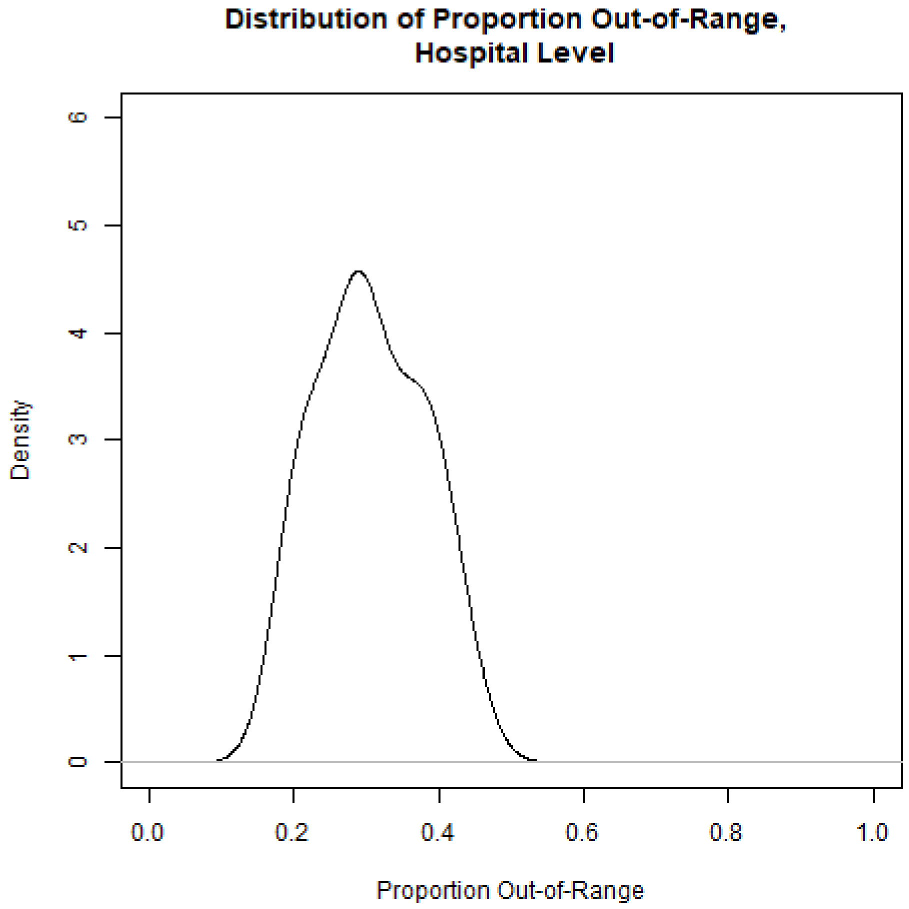


Figure 1b-2. Distribution in proportion out-of-range scores by hospital. This probability distribution is presented as an estimated density function, which is defined as a smooth function such that the probability of an outcome lying between any two given points on the x-axis is equal to the area under the curve of those two points (i.e. the area under the entire curve equals 1).

GLOBAL NOISE

Virtually all out-of-range scores are driven by excessive radiation doses, rather than global noise, which ranged from 0-0.6% across hospitals. This finding suggests image quality as reflected by global noise is not currently a large problem, and that there is considerable opportunity to optimize radiation doses without impacting quality. However, it is important to include the global noise in the measure as a balancing component to ensure that incentivizing the reduction of size-adjusted radiation doses does not compromise image quality.

PERFORMANCE IN THE UCSF INTERNATIONAL CT DOSE REGISTRY

When we applied the proposed measure to data assembled in the UCSF International CT Dose Registry – a repository of CT data containing over 6.5 million exams from 161 hospitals and imaging facilities – overall 33% of CT exams were out-of-range based on radiation dose exceeding thresholds. Overall, 135 facilities (84%) had out-of-range scores over 10%. Global noise cannot be assessed in the registry, but given the out-of-range values for global noise were <1% in field-testing data, we would expect it to also be low in the Registry. It is not possible to identify clinician groups in the UCSF registry, only facility-level performance.

Performance data at the facility level is as follows:

Mean measure (out-of-range) score: 30%, standard deviation: 18%

Range: minimum = 2%, maximum = 100%

Interquartile range: 27% (16%-43%)

Scores by percentile:

- 10th = 7%
- 20th = 11%
- 30th = 17%
- 40th = 22%
- 50th = 27%
- 60th = 31%
- 70th = 39%
- 80th = 46%
- 90th = 53%

[Response Ends]

1b.03. If no or limited performance data on the measure as specified is reported above, 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. Include citations.

[Response Begins]
Previous studies support the same performance gaps observed in our field-testing. The radiation doses used for CT exams are frequently far higher than needed for diagnosis and have been shown to vary up to 200-fold across facilities for patients imaged for the same clinical reason. (Demb 2017, Hricak 2010, Miglioretti 2013, Raff 2009, Smith-Bindman 2009, Smith-Bindman 2015, Smith-Bindman 2019,Tack 2014). For example, in a study of 151 organizations across seven countries, even after adjusting for patient characteristics, abdominal CT exams had a four-fold range in mean effective radiation dose and a 17-fold range in the proportion of high dose exams. (Smith-Bindman 2019)

There is also evidence that radiation doses can be reduced meaningfully without compromising the diagnostic usefulness of CT. In general, a direct relationship exists between radiation dose and image quality. As the dose increases, the image quality increases until a threshold is reached at which point no further benefit in image quality occurs. There is a concern that reducing radiation dose will compromise image quality, undermining the clinical value of CT exams. However, several studies suggest that radiation doses may be lowered 50-90% without impacting image quality or diagnostic accuracy because there is such a wide range in quality that is acceptable and that does not impact accuracy. (Catalano 2007, Smith-Bindman 2020, Konda 2016, Huppertz 2015, den Harder 2018, Rob 2017). A randomized trial of audit feedback combined with an educational intervention across 100 imaging facilities achieved 23-58% reductions in the proportion of high-dose exams (Smith-Bindman 2020), without any reduction in physician satisfaction with image quality.

References

1. Catalano C, Francone M, Ascarelli A, et al. Optimizing radiation dose and image quality. Eur Radiol. 2007 Dec;17 Suppl 6:F26-32.
2. Demb J, Chu P, Nelson T, et al. Optimizing Radiation Doses for Computed Tomography Across Institutions: Dose Auditing and Best Practices. JAMA Intern Med. 2017;177(6):810-81.
3. Den Harder AM, Willemink MJ, van Doormaal PJ, et al. Radiation dose reduction for CT assessment of urolithiasis using iterative reconstruction: A prospective intra-individual study. Eur Radiol. 2018;28(1):143-150.
4. Hricak H, Brenner DJ, Adelstein SJ, et al. Managing Radiation Use in Medical Imaging: A Multifaceted Challenge. Radiology. 2010.
5. Huppertz A, Lembcke A, Sariali el H, et al. Low Dose Computed Tomography for 3D Planning of Total Hip Arthroplasty: Evaluation of Radiation Exposure and Image Quality. J Comput Assist Tomogr. 2015;39(5):649-656.
6. Konda SR, Goch AM, Leucht P, et al. The use of ultra-low-dose CT scans for the evaluation of limb fractures: is the reduced effective dose using CT in orthopaedic injury (REDUCTION) protocol effective? Bone Joint J. 2016;98-B
7. Lukasiewicz A, Bhargavan-Chatfield M, Coombs L, et al. Radiation Dose Index of Renal Colic Protocol CT Studies in the United States: A Report from the American College of Radiology National Radiology Data Registry. Radiology. 2014;271(2):445-451.

8. Miglioretti DL, Johnson E, Williams A, et al. The use of computed tomography in pediatrics and the associated radiation exposure and estimated cancer risk. JAMA Pediatr. 2013;167(8):700-707.
9. Raff GL, Chinnaiyan KM, Share DA, et al. Radiation dose from cardiac computed tomography before and after implementation of radiation dose-reduction techniques. JAMA : the journal of the American Medical Association. 2009;301(2
10. Rob S, Bryant T, Wilson I, Somani BK. Ultra-low-dose, low-dose, and standard-dose CT of the kidney, ureters, and bladder: is there a difference? Results from a systematic review of the literature. Clin Radiol. 2017;72(1):11-15.
11. Smith-Bindman R, Chu P, Wang Y, et al. Comparison of the Effectiveness of Single-Component and Multicomponent Interventions for Reducing Radiation Doses in Patients Undergoing Computed Tomography: A Randomized Clinical Trial.
12. Smith-Bindman R, Lipson J, Marcus R, et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. Arch Intern Med. 2009;169(22):2078-2086.
13. Smith-Bindman R, Moghadassi M, Wilson N, et al. Radiation Doses in Consecutive CT Examinations from Five University of California Medical Centers. Radiology. 2015;277(1):134-141.
14. Smith-Bindman R, Wang Y, Chu P, et al. International variation in radiation dose for computed tomography examinations: prospective cohort study. BMJ. 2019;364:k4931.
15. Tack D, Jahnen A, Kohler S, et al. Multidetector CT radiation dose optimization in adults: short- and long-term effects of a clinical audit. Eur Radiol. 2014;24(1):169-175.

[Response Ends]

1b.04. 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.

Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included. Include mean, std dev, min, max, interquartile range, and scores by decile. 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 (4b) under Usability and Use.

[Response Begins]
Age and sex were explored in the general population, represented by all data in the UCSF International CT Dose Registry. No meaningful differences in radiation dose were identified either based on patient age or sex, after adjustment for patient size. The correlation between size-adjusted radiation dose and patient age is -0.004, with minimal variation between CT categories. The prevalence of out-of-range size-adjusted dose averaged 34% for female patients and 35% for male patients, with minimal variation between CT categories. A similarly comprehensive dataset was not available to assess the relationship between image noise and patient age or sex in the general population, though testing data shows that noise contributes minimally to the body of exams determined as "out-of-range" in our measure.

Despite this lack of disparity in the overall population, and despite no clinical justification for dosing differences by age or sex, individual clinicians, clinician groups, or hospitals may still express disparities between age and sex groups due to localized practice, and the proposed measure may have a role in reducing disparities.

Age and sex were explored in the testing data. Notable differences in radiation dose and noise out-of-range prevalence based on patient age and sex were identified in some hospitals.

Table 1b-1. Proportion out-of-range by age and sex by hospital.

Hospital (H)	Sex Female	Sex Male	Age 18-20	Age 21-30	Age 31-40	Age 41-50	Age 51-60	Age 61-70	Age 71-80	Age 80-89
H1	0.27	0.37	0.00	0.31	0.44	0.29	0.26	0.26	0.36	0.29
H2	0.20	0.24	0.20	0.28	0.29	0.21	0.17	0.24	0.24	0.13
H3	0.30	0.33	0.25	0.36	0.20	0.31	0.32	0.28	0.37	0.33

Hospital (H)	Sex Female	Sex Male	Age 18-20	Age 21-30	Age 31-40	Age 41-50	Age 51-60	Age 61-70	Age 71-80	Age 80-89
H4	0.34	0.45	0.11	0.26	0.53	0.42	0.40	0.40	0.36	0.42
H5	0.36	0.46	0.38	0.32	0.37	0.42	0.43	0.43	0.40	0.41
H6	0.19	0.23	0.09	0.19	0.19	0.20	0.22	0.22	0.24	0.20
H7	0.26	0.30	0.17	0.25	0.20	0.28	0.25	0.37	0.27	0.32
H8	0.33	0.39	0.36	0.44	0.39	0.40	0.35	0.34	0.34	0.29
H9	0.29	0.26	0.19	0.28	0.31	0.26	0.27	0.28	0.25	0.31
H10	0.35	0.39	0.30	0.25	0.26	0.33	0.37	0.41	0.37	0.43
H11	0.43	0.44	0.40	0.32	0.35	0.45	0.47	0.45	0.41	0.42
H12	0.31	0.29	0.18	0.27	0.30	0.34	0.27	0.30	0.32	0.30
H13	0.17	0.25	0.11	0.17	0.22	0.17	0.22	0.22	0.22	0.24
H14	0.24	0.30	0.09	0.27	0.24	0.25	0.27	0.29	0.27	0.28
H15	0.33	0.42	0.45	0.38	0.36	0.36	0.37	0.36	0.40	0.40
H16	0.28	0.32	0.40	0.31	0.32	0.33	0.31	0.28	0.27	0.30

Other social factors were not analyzed in field testing, because this information was not available to the developers and there was no *a priori* reason to believe that social factors such as insurance status, socioeconomic status, and/or functional status/disability would affect CT radiation dose. Therefore, disparities data by other population groups are not available.

[Response Ends]

1b.05. If no or limited data on disparities from the measure as specified is reported above, 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 above.

[Response Begins]
To the extent they have been studied, social factors including sex, race/ethnicity, and socioeconomic status are not predictive of radiation dose for CT exams. (Strauchler 2012, Freeman 2012, Hou 2014, Messenger 2015). However, as described in the studies led by Strauchler and Freeman, patients living in poverty are at higher risk for comorbid conditions associated with exposure to multiple scans over time and increased cumulative exposure to ionizing radiation from diagnostic imaging. Thus, it is particularly important to ensure that the doses used for CT in these individuals are not excessive, because vulnerable patients are at greatest risk of chronic disease and more likely to be exposed to many irradiating exams.

References

1. Freeman K, Strauchler D, Miller TS. Impact of socioeconomic status on ionizing radiation exposure from medical imaging in children. J Am Coll Radiol. 2012 Nov;9(11):799-807. doi: 10.1016/j.jacr.2012.06.005. PMID: 23122347.
2. Hou, J.K., Malaty, H.M. & Thirumurthi, S. Radiation Exposure from Diagnostic Imaging Studies Among Patients with Inflammatory Bowel Disease in a Safety-Net Health-Care System. Dig Dis Sci 59, 546–553 (2014). https://doi.org/10.1007/s10620-013-2852-1
3. Messenger B, Li D, Nasir K, Carr JJ, Blankstein R, Budoff MJ. Coronary calcium scans and radiation exposure in the multi-ethnic study of atherosclerosis. Int J Cardiovasc Imaging. 2016 Mar;32(3):525-9. doi: 10.1007/s10554-015-0799-3. Epub 2015 Oct 29. PMID: 26515964.
4. Strauchler D, Freeman K, Miller TS. The impact of socioeconomic status and comorbid medical conditions on ionizing radiation exposure from diagnostic medical imaging in adults. J Am Coll Radiol. 2012 Jan;9(1):58-63. doi: 10.1016/j.jacr.2011.07.009. PMID: 22221637.

[Response Ends]

2. 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.

sp.01. Provide the measure title.

Measure titles should be concise yet convey who and what is being measured (see What Good Looks Like).

[Response Begins]
Excessive Radiation Dose or Inadequate Image Quality for Diagnostic Computed Tomography (CT) in Adults (Facility Level)
[Response Ends]

sp.02. Provide a brief description of the measure.

Including type of score, measure focus, target population, timeframe, (e.g., Percentage of adult patients aged 18-75 years receiving one or more HbA1c tests per year).

[Response Begins]
This electronic clinical quality measure (eCQM) provides a standardized method for monitoring the performance of diagnostic CT to discourage unnecessarily high radiation doses, a risk factor for cancer, while preserving image quality. It is expressed as a percentage of eligible CT exams that are out-of-range based on having either excessive radiation dose or inadequate image quality, relative to evidence-based thresholds based on the clinical indication for the exam. All diagnostic CT exams of specified anatomic sites performed in inpatient and hospital outpatient care settings are eligible.
[Response Ends]

sp.04. Check all the clinical condition/topic areas that apply to your measure, below.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.

Please do not select:

- *Surgery: General*

[Response Begins]
Other (specify)
Diagnostic radiology
[Response Ends]

sp.05. Check all the non-condition specific measure domain areas that apply to your measure, below.

[Response Begins]
Safety
[Response Ends]

sp.06. Select one or more target population categories.

Select only those target populations which can be stratified in the reporting of the measure's result.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.

Please do not select:

- *Populations at Risk: Populations at Risk*

[Response Begins]
Adults (Age >= 18)
[Response Ends]

sp.07. Select the levels of analysis that apply to your measure.

Check ONLY the levels of analysis for which the measure is SPECIFIED and TESTED.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.

Please do not select:

- *Clinician: Clinician*
- *Population: Population*

[Response Begins]
Facility
[Response Ends]

sp.08. Indicate the care settings that apply to your measure.

Check ONLY the settings for which the measure is SPECIFIED and TESTED.
[Response Begins]
Inpatient/Hospital
Outpatient Services
[Response Ends]

sp.09. 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. If no URL is available, indicate "none available".

[Response Begins]
https://www.alaracare.com/qualitymeasures

Please note, we have developed and tested the eCQM in both a Quality Data Model (QDM) format, to allow immediate implementation, and a FHIR format to align with CMS's strategy for increasing interoperability. The human readable outputs for both QDM and FHIR formats are attached to this application and available at the website above.
[Response Ends]

sp.10. Indicate whether Health Quality Measure Format (HQMF) specifications are attached.

Attach the zipped output from the eCQM 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).
[Response Begins]
HQMF specifications are attached.
[Response Ends]

Attachment: Human_readable_1074_QDM_Facility.pdf

Attachment: Human_readable_1075_FHIR_Facility.pdf
Attachment: CMS1074-v0-0-012-QDM-5-6.zip
Attachment: CMS1075FHIR-v0-0-012-FHIR-4-0-1.zip

sp.11. Attach the data dictionary, code table, or value sets (and risk model codes and coefficients when applicable). Excel formats (.xlsx or .csv) are preferred.

Attach an excel or csv file; if this poses an issue, contact staff. Provide descriptors for any codes. Use one file with multiple worksheets, if needed.

[Response Begins]

Available in attached Excel or csv file

[Response Ends]

Attachment: Binning algorithm CPT ICD List_2021.08.02 v18.xlsx
Attachment: LOINC_code_table.xlsx

sp.12. State the numerator.

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.

[Response Begins]

Diagnostic CT exams that have a size-adjusted radiation dose value greater than the threshold specific to the CT category (reflecting the body region imaged and the radiation dose and image quality required for that exam given the reason for the exam), or a global noise value greater than a threshold specific to the CT Category.

[Response Ends]

sp.13. Provide details needed to calculate the numerator.

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 sp.11.

[Response Begins]

The numerator represents the total number of out-of-range (i.e. failed) exams.

Through this application, these LOINC variable names will be shortened for brevity, as follows:

Calculated CT Size-Adjusted Dose = size-adjusted radiation dose

Calculated CT Global Noise = global noise

CT Dose and Image Quality Category = CT category

Definitions

Size-adjusted radiation dose reflects the total radiation dose delivered during a CT, risk-adjusted for patient size. The total radiation dose is recorded for each CT exam using the standardized metric of dose length product (ACR–AAPM–SPR: Practice parameter, European Commission, Radiation Protection No. 185, ICRP Publication 135, Kanal 2017, Smith-Bindman 2019). The patient size is defined as the effective diameter of the anatomic area scanned in millimeters, computed on the mid-slice of the scan. Where axial images are available showing the entire anatomic area, the patient size is computed as the average effective patient diameter on the axial image (Cheng 2013). If axial images showing the entire anatomic area are unavailable, the effective diameter is computed on the coronal localizer image (Christianson 2012). The dose length product is adjusted for patient size using log-transformed linear regression models. The size-adjusted radiation dose value is compared with thresholds that vary by the CT category.

Global noise reflects the image quality of the CT exam. Noise is the most widely used measure of CT image quality. (Catalano 2007, Christianson 2012, Malkus 2017, Schindera 2009, Smith 2008, Szczykutowicz 2017, Szczykutowicz 2021, Willemink 2014) Noise represents differences in the appearance of homogenous areas of tissue that is not a result of inherent tissue composition, but rather of the quality due to imaging technique. In general, image noise in CT reflects the number of x-ray photons hitting the detector, and this will be influenced by the x-ray tube voltage and tube current, as well as patient factors such as the patient's body habitus, the body region being evaluated, and other scanning parameters such as the slice thickness. Different clinical questions require different values of noise, yet in general, the greater the noise, the worse the image quality and the poorer the diagnostic accuracy, although this is not a simple linear relationship. Diagnostic accuracy may be acceptable for a large range of noise values, but unacceptable only at a high value. Noise can be quantified in CT images by positioning standard elliptical regions of interest in a known density structure (e.g. water, air, soft tissue) and measuring the standard deviation of the measured values in Hounsfield units. (Catalano 2007). Noise as defined in this measure is calculated on every CT image within a scan (a single irradiating event), and the global noise value for each scan is the mean value across all images. For CT exams that have multiple scans (for example a scan without contrast, followed by a scan with contrast, followed by a delayed scan), the exam is assigned the "best" global noise value across all scans, i.e. the highest quality scan. The global noise value for each scan is also standardized to a 3 mm slice thickness. (Alshipli 2017) The global noise value is compared with thresholds that vary by the CT category.

Details needed to calculate the numerator

To calculate the numerator, the size-adjusted radiation dose and global noise for each CT exam are compared against the following evidence-based thresholds specific to the CT Category (Table sp-1). If a CT exam has a size-adjusted radiation dose and/or global noise value exceeding these thresholds, the exam is considered out-of-range (i.e. "failed") and is counted in the numerator.

Table sp-1. Size-adjusted radiation dose and global noise thresholds by CT category.

CT Category	**Size-Adjusted Radiation Dose THRESHOLD (Dose length product, mGy-cm)**	**Global Noise THRESHOLD (Hounsfield units)**
Abdomen and Pelvis Low Dose	598	64
Abdomen and Pelvis Routine Dose	644	29
Abdomen and Pelvis High Dose	1260	29
Cardiac Low Dose	93	55
Cardiac Routine Dose	576	32
Chest Low Dose	377	55
Chest Routine Dose	377	49
Cardiac High Dose or Chest High Dose	1282	49
Head Low Dose	582	115
Head Routine Dose	1025	115
Head High Dose	1832	115
Extremity	320	73
Neck or Cervical Spine	1260	25
Thoracic or Lumbar Spine	1260	25
Simultaneous Chest and Abdomen and Pelvis	1637	29
Simultaneous Thoracic and Lumbar Spine	2520	25
Simultaneous Head and Neck Routine Dose	2285	25
Simultaneous Head and Neck High Dose	3092	25

References

ACR–AAPM–SPR: Practice parameter for diagnostic reference levels and achievable doses in medical x-ray imaging. Revised 2018.Alshipli M and Kabir NA, Effect of slice thickness on image noise and diagnostic content of single-source-dual energy computed tomography 2017 *J. Phys.: Conf. Ser.* 851 012005

Catalano C, Francone M, Ascarelli A, Mangia M, Iacucci I, Passariello R. Optimizing radiation dose and image quality. Eur Radiol. 2007;17 Suppl 6:F26-32.

Cheng PM. Automated estimation of abdominal effective diameter for body size normalization of CT dose. J Digit Imaging. 2013 Jun;26(3):406-11.

Christianson O, Li X, Frush D, Samei E. Automated size-specific CT dose monitoring program: assessing variability in CT dose. Med Phys. 2012 Nov;39(11):7131-9.

Christianson O, Winslow J, Frush DP, Samei E. Automated Technique to Measure Noise in Clinical CT Examinations. AJR Am J Roentgenol. 2015 Jul;205(1):W93-9.

European Commission, Radiation Protection No. 185, European guidelines on diagnostic reference levels for paediatric imaging. 2018.

ICRP, 2017. Diagnostic reference levels in medical imaging. ICRP Publication 135. Ann. ICRP 46(1).

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;284(1):120-133.

Malkus A, Szczykutowicz TP. A method to extract image noise level from patient images in CT. Med Phys. 2017 Jun;44(6):2173-2184. doi: 10.1002/mp.12240. Epub 2017 Apr 25.

Schindera ST, Nelson RC, Yoshizumi T, et al. Effect of automatic tube current modulation on radiation dose and image quality for low tube voltage multidetector row CT angiography: phantom study. Acad Radiol. 2009;16(8):997-1002.

Smith AB, Dillon WP, Lau BC, et al. Radiation dose reduction strategy for CT protocols: successful implementation in neuroradiology section. Radiology. 2008;247(2):499-506.

Smith-Bindman R, Wang Y, Chu P, et al. International variation in radiation dose for computed tomography examinations: prospective cohort study. BMJ. 2019;364:k4931.

Szczykutowicz TP, DuPlissis A, Pickhardt PJ. Variation in CT Number and Image Noise Uniformity According to Patient Positioning in MDCT. AJR Am J Roentgenol. 2017 May;208(5):1064-1072. doi: 10.2214/AJR.16.17215. Epub 2017 Mar 7.

Szczykutowicz TP, Nett B, Cherkezyan L, et al. Protocol Optimization Considerations for Implementing Deep Learning CT Reconstruction. AJR American journal of roentgenology. 2021;216(6):1668-1677.

Willemink MJ, Takx RA, de Jong PA, et al. Computed tomography radiation dose reduction: effect of different iterative reconstruction algorithms on image quality. J Comput Assist Tomogr. 2014;38(6):815-823.

[Response Ends]

sp.14. State the denominator.

Brief, narrative description of the target population being measured.

[Response Begins]
All diagnostic CT exams performed on adults (aged 18 years and older) during the measurement period of one year that have an assigned CT category, a size-adjusted radiation dose value, and a global noise value.
[Response Ends]

sp.15. Provide details needed to calculate the denominator.

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 sp.11.

[Response Begins]
Target population

The target population includes all diagnostic CT exams of specified anatomic sites performed on adults during the measurement period.

On a practical level, to be included, the exam must have an assigned CT category and must have a size-adjusted radiation dose value and a global noise value (meaning the relevant CT data must be available to allow calculation of patient size and image quality.)

CT exams performed in conjunction with nuclear medicine (such as SPECT and PET-CT), biopsies, procedures related to an intervention, assessments of bone mineral density, where the body region is not specified, or where no primary images were obtained, are not included as they are not diagnostic CT.

Definitions

<u>CT Dose and Image Quality Category (short term: “CT category”)</u>: reflects the type of exam performed based on the body

region and the clinical indication for the exam. Each CT category has a specific set of radiation dose and global noise thresholds. The categories are:

1. Abdomen and Pelvis Low Dose
2. Abdomen and Pelvis Routine Dose
3. Abdomen and Pelvis High Dose
4. Cardiac Low Dose
5. Cardiac Routine Dose
6. Chest Low Dose
7. Chest Routine Dose
8. Cardiac High Dose or Chest High Dose
9. Head Low Dose
10. Head Routine Dose
11. Head High Dose
12. Extremity
13. Neck or Cervical Spine
14. Thoracic or Lumbar Spine
15. Simultaneous Chest and Abdomen and Pelvis
16. Simultaneous Thoracic and Lumbar Spine
17. Simultaneous Head and Neck Routine Dose
18. Simultaneous Head and Neck High Dose

Time period for data collection

One calendar year, although shorter periods can be used for high-volume entities

Codes

LOINC codes representing the data elements required for this measure are published in the Value Set Authority Center (VSAC). They are attached in section sp.11. The data elements themselves and data sources are described in section sp.29.

[Response Ends]

sp.16. Describe the denominator exclusions.

Brief narrative description of exclusions from the target population.

[Response Begins]
Denominator exclusions are CT exams that simultaneously include multiple body regions outside of four commonly encountered multiple region groupings (specified as LOINC code 96914-7, CT Dose and Image Quality Category, Full Body). Denominator exclusions are also CT exams with missing patient age, missing size-adjusted radiation dose, or missing global noise. These are technical exclusions ("missing data") from the initial population. Technical exclusions will be flagged, corrected whenever possible, and tracked at the level of the accountable entity.
[Response Ends]

sp.17. Provide details needed to calculate the denominator exclusions.

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 sp.11.

[Response Begins]

Exclusions

CT exams that cannot be placed into a CT category because they are simultaneous include exams of multiple body regions outside of four commonly encountered multiple region groupings are excluded. The four commonly encountered multiple region groupings are: (1) Simultaneous Chest and Abdomen and Pelvis; (2) Simultaneous Thoracic and Lumbar Spine; (3) Simultaneous Head and Neck Routine Dose; and (4) Simultaneous Head and Neck High Dose. Simultaneous exams of the abdomen and lower extremity are already included as a subset of exams included as part of the "Abdomen and Pelvis High Dose" category. Chest and cardiac are not considered separate body regions for purposes of determining whether the exam contains multiple body regions.

Technical exclusions

CT exams missing any of the four data elements required to calculate measure score are considered technical exclusions: CT category; size-adjusted radiation dose; global noise; birth date.
[Response Ends]

sp.18. Provide all information required to stratify the measure results, if necessary.

Include 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 in the Data Dictionary field.

[Response Begins]
The only stratification variable is the CT category, which is constructed using International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM) diagnosis codes and CPT® (Current Procedural Terminology) procedure codes from the billing entity’s claim (or other mapped fields in the electronic health record).

CT categories were constructed to reflect various body regions and different clinical indications for imaging, since different amounts of radiation and image quality are needed to create images sufficient for diagnosis depending on these factors. The framework for creating these categories took an image-quality informed approach, which first relied on categorizing CT exams into 10 body regions. In five of these regions (extremities, neck [including cervical spine], thoraco-lumbar spine [reflecting either thoracic spine or lumbar spine], combined chest-abdomen, and combined thoraco-lumbar spine [reflecting both thoracic and lumber spine]), clinical indications for scanning do not play a substantial role in altering the amount of radiation needed to produce required images; thus, there is a single CPT®-determined category for each of these body regions. In five other body regions (head, chest, cardiac, abdomen, and combined head and neck), clinical indications do affect the optimal radiation dose, thus these regions were sub-divided based on ICD-10-CM/CPT® defined clinical indications into low, routine, or high radiation dose categories. The “combined head and neck” category was divided into routine and high dose. The approach to determining low, routine, or high radiation doses within these categories was informed by: 1) a review of the published literature; 2) consultation with radiologists with specialty expertise; 3) input from a Technical Expert Panel; and 4) empirical evaluation of about 4.5 million consecutive CT exams from 161 imaging facilities that contribute to the UCSF International CT Dose Registry (January 1, 2016 to December 31, 2019). The categories had face validity as assessed by the Technical Expert Panel, and a manuscript describing this work is under resubmission review in *Radiology.* The strategy in creating the logic to assign exams to CT categories was to identify indications that were *exceptions* to the routine radiation dose category, rather than to identify every indication for scanning within the routine category. For example, lung cancer screening is the only defined indication for low-dose chest CT, and evaluation for suspected aortic rupture or dissection (or, more generally, a patient in acute shock) is the only defined indication for high-dose chest CT, leaving all other chest CTs in the routine-dose category. As in this example, all strata were constructed to mimic clinical decision-making regarding the most appropriate imaging protocol and its associated radiation dose range. The logic and code table for assigning body regions and indications to CT categories is provided in sp.11.

Size-adjusted radiation dose and global noise are assessed against thresholds specific to the CT category, as described

further below. However, the measure score is binary (in-range or out-of-range), and the total number/proportion of out-of-range exams is summed for a reportable entity without need for separate stratified calculation or reporting. The measure is not weighted by the stratum, but rather every CT exam contributes equally to overall score. An entity that performs CT exams within only a few strata has its exams judged against the thresholds for the exams that it performs.
[Response Ends]

sp.19. Select the risk adjustment type.

Select type. Provide specifications for risk stratification and/or risk models in the Scientific Acceptability section.
[Response Begins]
Statistical risk model
[Response Ends]

sp.20. Select the most relevant type of score.

Attachment: If available, please provide a sample report.
[Response Begins]
Rate/proportion
[Response Ends]

sp.21. Select the appropriate interpretation of the measure score.

Classifies interpretation of score according to whether better quality or resource use is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score
[Response Begins]
Better quality = Lower score
[Response Ends]

sp.22. Diagram or describe the calculation of the measure score as an ordered sequence of steps.

Identify the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period of data, aggregating data; risk adjustment; etc.

[Response Begins]
At a high level, the following steps occur for each CT exam assessed during the reporting period for the reporting entity:

1. The CT exam is assigned to a CT category using diagnosis (ICD-10-CM) and procedure (CPT®) codes.
2. The patient's size is calculated from DICOM (pixel) data included with the CT exam.
3. The size-adjusted radiation dose is calculated from DICOM data, including the Radiation Dose Structured Report (RDSR) and image pixel data, stored with the CT exam.
4. The global noise is calculated from DICOM (pixel) data stored with the CT exam.
5. The size-adjusted radiation dose and global noise are compared with allowable thresholds, and if either (or both) exceed the allowable thresholds, the CT exam is considered out-of-range (failed).
6. The measure score for the reporting entity is calculated as the proportion of out-of-range CT exams for the reporting entity.

As described in section sp.29, the measure derives standardized data elements from structured fields within the EHR and the radiology electronic clinical data systems including the Radiology Information System (RIS) and the Picture Archiving and Communication System (PACS).

In its existing framework, the eCQM cannot consume primary imaging data in its original format and thus cannot access the requisite data for measure calculation. UCSF and Alara Imaging, Inc. have developed software to access and process primary data elements from the electronic systems to calculate the three variables required by the measure – CT category, size-adjusted radiation dose, and global noise – which can then be ingested by the eCQM for calculating the measure score. The calculation of these variables is broadly described as "pre-processing."

This approach was tested across diverse EHR and PACS platforms. The software is installed at imaging facilities or hospitals within the firewall and functions as an edge device, drawing in data from the specified sources and calculating the variables that can be ingested by the eCQM in a manner that minimizes burden. The software can be fully integrated locally into existing data flows using QDM or FHIR or can be available as a web interface for organizations that do not desire a fully integrated solution.

Consecutive, diagnostic CT exams over one calendar year will be evaluated by the eCQM. These exams may be submitted prospectively in real-time or batch-submitted retrospectively (daily, weekly, monthly). The following steps take place to ingest and calculate the measure score on consecutive CT exams:

Ingestion – Edge Device

1. Radiology electronic clinical data systems record and store information related to medical imaging studies. EHRs record and store information related to the patient and medical imaging encounters.

2. Radiology electronic clinical data systems are configured to automatically forward relevant CT studies with included RDSR reports via DICOM protocols to the edge device. Once the CT study is forwarded to the edge device, the edge device queries the EHR via FHIR or direct API calls for additional information that is then linked to the related exam.

Ingestion – Web Interface

3. For sites not using the integrated edge device, information can be exported from the EHR and radiology electronic clinical data systems via custom reports such as FHIR resources, CCDA documents, and DICOM studies. Relevant information can then be uploaded by sites through a web application for measure calculation. This service will be provided at cost, or free, to minimize burden on providers.

Calculation

4. Software assesses the information for each CT exam for eligibility based on initial population assessment criteria and missing data. Missing data are flagged for the reporting entity and recovered when possible.

5. Remaining CT exams undergo *pre-processing* on the edge device software or web application, in which the three data elements needed for measure calculation are generated from primary data elements.

   a. CT category: The software categorizes the CT exam based on anatomic area (determined by the procedure (CPT®) codes on the exam claims data) and clinical indication (based on the diagnosis (ICD-10-CM) codes associated with the exam order).
   b. Size-adjusted radiation dose: The software calculates patient size from image pixel data and receives radiation dose from the Radiation Dose Structured Report (RDSR). The software uses these variables to perform risk adjustment of radiation dose based on patient size. The output of this process is size-adjusted radiation dose.
   c. Global noise: The software measures noise in pixel data on CT images. Noise varies by slice thickness, with thinner image slices having higher noise; thus, global noise is adjusted by slice thickness.

6. The eCQM receives all data elements.

7. The eCQM removes denominator exclusions (simultaneous CT exams of multiple body regions outside of four commonly encountered multiple region groupings).

8. For each individual CT exam, the eCQM compares size-adjusted radiation dose and global noise against allowable thresholds specific to the CT category. Exams exceeding dose or noise thresholds are considered failures (out-of-range).

9. The eCQM scores each CT exam in range (pass) or out-of-range (fail). The sum of all out-of-range exams constitutes the numerator for the measure at the patient or population level.

10. An overall measure score (i.e. proportion of CT exams that are out-of-range relative to all evaluated exams) is calculated and can be queried/aggregated at the level of the hospital.

For sites that wish to use existing EHR vendors for eCQM computation and submission, primary data elements are sent via the edge device or downloaded via the web interface for ingestion and storage by site EHRs either as a FHIR observation resource, or if FHIR is unavailable, through an integration with an EHR via API.

The measure score can be reported to CMS by the existing EHR vendor, or if preferred, the measure steward is also able to compute and submit measure results to CMS on behalf of sites. Either way, reporting will follow established CMS implementation guidelines.

Feedback will be provided to the hospital on the proportion of scans that are out-of-range and the reason these scans are out-of-range to encourage performance improvement.

[Response Ends]

sp.25. If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.

[Response Begins]
The measure is not based on a sample.
[Response Ends]

sp.28. Select only the data sources for which the measure is specified.

[Response Begins]
Electronic Health Data
Electronic Health Records
[Response Ends]

sp.29. Identify the specific data source or data collection instrument.

For example, provide the name of the database, clinical registry, collection instrument, etc., and describe how data are collected.

[Response Begins]
The measure derives standardized data elements from structured fields within the EHR and the radiology electronic clinical data systems including the Radiology Information System (RIS) and the Picture Archiving and Communication System (PACS). Primary imaging data stored in structured fields in the radiology electronic clinical data systems have been historically inaccessible using the existing eCQM framework. Thus, the eCQM cannot consume CT images and Radiation Dose Structured Reports (RDSR, which contain the radiation dose) in their original DICOM formats. These primary data, listed below, must be processed to create “calculated” data elements that can then be ingested by the eCQM. The measure developers have created software (available to all users to install locally by agreement, or made accessible through a web interface) to access and process primary data elements from these electronic systems to calculate variables that the eCQM uses to calculate the measure score.

The following primary data elements, their sources, and how they are used in the measure, are illustrated in Table sp-2 below. The steps for how these data elements are accessed, ingested, and processed by the eCQM are described in sp.22.

1. Diagnostic Study, Performed: Categorized CT Exams. All diagnostic CT exams performed during the measurement period, including the type of exam performed (derived from procedure (CPT®) codes associated with the exam bill) and the reason for study (derived from diagnosis (ICD-10-CM) codes associated with the exam order and with the exam bill). A validated algorithm uses combinations of diagnosis and procedure codes to generate the ***CT Dose and Image Quality Category*** ("CT category") that specifies the radiation dose and image quality thresholds for each CT exam. (CPT Copyright 2017 American Medical Association. All rights reserved. CPT® is a registered trademark of the American Medical Association.)

2. Diagnostic Study, performed: CT Studies with Radiation Dose Result. Radiation dose is derived from the Radiation Dose Structured Report (RDSR), a DICOM structured element generated by the CT machine for every exam, giving the total radiation dose delivered by the exam (measured as dose length product, mGy-cm). This is used to generate ***Calculated CT Size-Adjusted Dose*** ("size-adjusted radiation dose").

3. Diagnostic Study, performed: CT Studies with Image Quality Result. CT image pixel data are generated by the CT machine for every CT exam and stored as DICOM structured data. They are used to measure patient size (measured as diameter on mid-scan axial or coronal images, in mm), which is used in generating the final data element ***Calculated CT Size-Adjusted Dose***. They are also used to generate the final data element ***Calculated CT Global Noise*** ("global noise," measured in Hounsfield units).

4. Birth date, to confirm the patient is 18 years of age or older.

5. Supplemental data elements: payer, race, ethnicity, and sex.

Table sp-2. Primary data elements are accessed and combined to generate final data elements. "Radiology Electronic Clinical Data Systems" are the core information systems for data storage and practice management that are nearly universal in radiology practices, including the Picture Archiving and Communication System (PACS) and Radiology Information System (RIS).

Data source	Primary Accessed Data Element	Primary Accessed Data Element Code System	Calculated Data Element	Calculated Data Element Code System	Calculated Data Element Description
Electronic Health Record (EHR), or Radiology Electronic Clinical Data Systems (non-EHR)	Diagnostic Study, performed: CT Studies	ICD-10-CM	CT Dose and Image Quality Category	LOINC	Reflects the type of exam performed based on body region and clinical indication. Each CT category has a specific set of dose and image quality thresholds.
Radiology Electronic Clinical Data Systems (non-EHR)	Diagnostic Study Performed: CT Studies *Result attribute: Radiation Dose Structured Report (RDSR)*	DICOM	Calculated CT Size-Adjusted Dose	LOINC	Reflects the total radiation dose received during CT, risk-adjusted by patient size. The size-adjusted radiation

Data source	Primary Accessed Data Element	Primary Accessed Data Element Code System	Calculated Data Element	Calculated Data Element Code System	Calculated Data Element Description
Radiology Electronic Clinical Data Systems (non-EHR)	Diagnostic Study Performed: CT Studies *Result attribute: Image Pixel Data*	DICOM			dose thresholds vary by the CT category.
Radiology Electronic Clinical Data Systems (non-EHR)	Diagnostic Study Performed: CT Studies *Result attribute: Image Pixel Data*	DICOM	Calculated CT Global Noise	LOINC	Reflects the image quality (represented by global noise) of the CT. The global noise thresholds vary by the CT category. The measure adjusts global noise measurement by slice thickness.
Electronic Health Record (EHR)	Birth Date	LOINC	Birth Date	LOINC	MM-DD-YYYY, to confirm the patient is eligible

[Response Ends]

sp.30. Provide the data collection instrument.

[Response Begins]
No data collection instrument provided
[Response Ends]

Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate fields in the Scientific Acceptability sections of the Measure Submission Form.

- Measures must be tested for all the data sources and levels of analyses that are specified. If there is more than one set of data specifications or more than one level of analysis, contact NQF staff about how to present all the testing information in one form.
- All required sections must be completed.
- For composites with outcome and resource use measures, Questions 2b.23-2b.37 (Risk Adjustment) also must be completed.
- If specified for multiple data sources/sets of specifications (e.g., claims and EHRs), Questions 2b.11-2b.13 also must be completed.
- An appendix for supplemental materials may be submitted (see Question 1 in the Additional section), but there is no guarantee it will be reviewed.
- Contact NQF staff with any questions. Check for resources at the Submitting Standards webpage.
- For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for the 2021 Measure Evaluation Criteria and Guidance.

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.
2a. Reliability testing demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is

precise. For instrument-based measures (including PRO-PMs) and composite performance measures, reliability should be demonstrated for the computed performance score.
2b1. Validity testing demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For instrument based measures (including PRO-PMs) and composite performance measures, validity should be demonstrated for the computed performance score.
2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure;
AND
If patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately).
2b3. For outcome measures and other measures when indicated (e.g., resource use):
• an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; 14,15 and has demonstrated adequate discrimination and calibration
OR
• rationale/data support no risk adjustment/ stratification.
2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful 16 differences in performance;
OR
there is evidence of overall less-than-optimal performance.
2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.
2b6. Analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders) and how the specified handling of missing data minimizes bias.
2c. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:
2c1. the component measures fit the quality construct and add value to the overall composite while achieving the related objective of parsimony to the extent possible; and
2c2. the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.
(if not conducted or results not adequate, justification must be submitted and accepted)

Definitions
Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.
Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
Risk factors that influence outcomes should not be specified as exclusions.

With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v.$5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

Please separate added or updated information from the most recent measure evaluation within each question response in the Importance to Scientific Acceptability sections. For example:

<u>2021 Submission:</u>
Updated testing information here.

<u>2018 Submission:</u>
Testing from the previous submission here.

Reliability

2a.01. Select only the data sources for which the measure is tested.

[Response Begins]
Electronic Health Data
Electronic Health Records
[Response Ends]

2a.02. 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).

[Response Begins]
N/A – an existing dataset was not used
[Response Ends]

2a.03. Provide the dates of the data used in testing.

Use the following format: "MM-DD-YYYY - MM-DD-YYYY"

[Response Begins]
02-01-2020 – 04-15-2021
[Response Ends]

2a.04. Select the levels of analysis for which the measure is tested.

Testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.

Please do not select:

- *Clinician: Clinician*
- *Population: Population*

[Response Begins]
Facility
[Response Ends]

2a.05. List the measured entities 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.

[Response Begins]
Data were collected from each of the organizations and testing sites for approximately 4 weeks. Table 2a-1 provides data for the 16 hospitals. Four of the included health systems (8 included hospitals) are members of America's Essential Hospitals, an association representing 300 hospitals that care for the nation's vulnerable and provide vital services to communities, including caring for many patients with Medicaid. These organizations are noted as "safety net" in Table 2a-

Table 2a-1. Organizations and hospitals where field-testing was performed. Annual inpatient admissions and emergency department (ED) visits reported for each organization, including data from all hospitals where testing done, from the most recent year of available data (2018-2020). Inpatient discharges shown where inpatient admissions data is not available. Annualized CT exams estimated from 4 week testing.

(The numbers given in the leftmost column do not correspond to hospital numbers in the results section).

*	EHR	Location	Source of Data	Annual CT Exams	Annual Inpatient Admissions	Annual ED Visits	Urban/suburban/rural/safety net
1	Cerner	Huntsville, AL	Hospital	73,884	100,215	324,865	Urban, suburban, rura
2	Epic	Sacramento, CA	Hospital	38,520	29,841	73,194	Urban, suburban, rural, safety net
3	Epic	Irvine, CA	Hospital	32,112	22,142	55,000+	Urban, suburban, rural, safety net
4	Epic	San Diego, CA	Hospital	28,848	33,605	120,000	Urban, suburban, rural, safety net
5	Epic	Detroit, MI	Hospital	7,500	115,000	100,000 +	Urban, suburban, rural, safety net
6	Epic	Detroit, MI	Hospital	9,000			
7	Epic	Detroit, MI	Hospital	10,224			
8	Epic	Detroit, MI	Hospital	12,036			
9	EPIC	Detroit, MI	Hospital	28,008			
10	AllScripts	Greater NYC, NY	Hospital	28,932	303,729	865,260	Urban, suburban
11	AllScripts	Greater NYC, NY	Hospital	29,076			

*	EHR	Location	Source of Data	Annual CT Exams	Annual Inpatient Admissions	Annual ED Visits	Urban/suburban/rural/safety net
12	AllScripts	Greater NYC, NY	Hospital	22,068			
13	Epic	New York, NY	Hospital	32,520	154,662	521,382	Urban, suburban
14	Epic	New York, NY	Hospital	25,224			
15	Epic	New York, NY	Hospital	32,280			
16	Epic	New York, NY	Hospital	35,832			

*Cell left intentionally blank

References

American’s Safety Net Hospitals, https://essentialhospitals.org/americas-essential-hospitals/, accessed August 1, 2021

[Response Ends]

2a.06. Identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis), separated by level of analysis and data source; if a sample was used, describe how patients were selected for inclusion in the sample.

If there is a minimum case count used for testing, that minimum must be reflected in the specifications.

[Response Begins]
Consecutive CT exams were assembled from contributing testing sites for approximately 4 weeks without sampling. The distribution of CT exam by age and sex are shown in Table 2a-2 below. Each cell shows the proportion of CTs by hospital by sex and within each age strata. Data were not collected in adults ages 90 and older related to Institutional Review Board requirements. Race was not collected. All diagnoses that are associated with CT imaging are included and this includes most medical diagnostic groups.

Table 2a-2. Distribution of age and sex per hospital, in field-testing data.

Heading(H)	CT Exams	Sex Female	Sex Male	Age 18-20	Age 21-30	Age 31-40	Age 41-50	Age 51-60	Age 61-70	Age 71-80	Age 80-89
H1	625	0.63	0.37	0.00	0.06	0.09	0.10	0.14	0.21	0.19	0.20
H2	750	0.58	0.42	0.01	0.07	0.10	0.12	0.15	0.21	0.20	0.14
H3	852	0.54	0.46	0.02	0.05	0.07	0.12	0.16	0.23	0.18	0.17
H4	1003	0.52	0.48	0.01	0.06	0.08	0.10	0.19	0.23	0.21	0.11
H5	2334	0.51	0.49	0.02	0.07	0.10	0.13	0.18	0.25	0.17	0.09
H6	6157	0.53	0.47	0.03	0.07	0.09	0.12	0.18	0.20	0.18	0.12
H7	2710	0.53	0.47	0.01	0.08	0.08	0.11	0.19	0.18	0.17	0.17
H8	2101	0.47	0.53	0.01	0.10	0.15	0.13	0.19	0.18	0.13	0.12
H9	2690	0.54	0.46	0.01	0.08	0.10	0.13	0.18	0.20	0.17	0.13
H10	2986	0.49	0.51	0.01	0.07	0.10	0.10	0.21	0.22	0.17	0.13
H11	2411	0.57	0.43	0.01	0.03	0.05	0.09	0.20	0.29	0.23	0.11
H12	2423	0.52	0.48	0.02	0.06	0.09	0.09	0.15	0.21	0.22	0.16
H13	1839	0.53	0.47	0.02	0.08	0.13	0.13	0.18	0.18	0.14	0.14
H14	3210	0.48	0.52	0.02	0.07	0.10	0.11	0.19	0.23	0.18	0.09
H15	2676	0.47	0.53	0.02	0.09	0.12	0.13	0.20	0.19	0.16	0.10
H16	2404	0.48	0.52	0.01	0.06	0.09	0.12	0.21	0.25	0.17	0.09

[Response Ends]

2a.07. 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.

[Response Begins]

Data element validity

- CT category, size-adjusted radiation dose, and global noise were each validated on 37,172 CT exams from field-testing data.
- Global noise was validated using 740 exams from the Image Quality Study.

Measure score reliability was tested at the hospital level and included 16 hospitals.

Measure score validity was tested on a random sample of 7,000 CT exams (1,000 CT exams sampled per testing site).

Risk adjustment testing (including correlation between patient size and dose) was conducted using data on 6.5 million adult CT exams from the UCSF International CT Dose Registry.

Exclusions testing was completed on 42,211 exams from field-testing data, including 37,172 included in study, 3,349 technical exclusions (“missing data”), and 1,690 excluded as "uncommon multiple anatomic regions."

[Response Ends]

2a.08. List 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.

[Response Begins]
Social factors do not fit into the logic model described above, and are not known to affect radiation dose, because technical decisions on how to perform CT are made at the facility level rather than at the individual patient level. Given that this measure is an eCQM, no patient-reported data were collected. Therefore, social risk factors were not available and not analyzed.
[Response Ends]

Note: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a.07 check patient or encounter-level data; in 2a.08 enter "see validity testing section of data elements"; and enter "N/A" for 2a.09 and 2a.10.

2a.09. Select the level of reliability testing conducted.

Choose one or both levels.
[Response Begins]
Accountable Entity Level (e.g., signal-to-noise analysis)
[Response Ends]

2a.10. For each level of reliability testing checked above, describe the method of reliability testing and what it tests.

Describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used.

[Response Begins]
We estimated measure score reliability at the accountable entity 0.99 level using the intraclass correlation coefficient (ICC), a reliability coefficient that conceptually represents the true (between-entity) variance in a measure divided by the sum of true variance and error (within-entity) variance. We used randomly split samples for each accountable entity with 1,000 repetitions, applying a one-way random effects model, assuming that both entity effects and residual effects are random, independent, and normally distributed with mean 0. This approach corresponds to Case 1 or the ICC(1) in McGraw and Wong's seminal description of ICC reliability methods.(McGraw 1996) The Spearman-Brown prophecy formula was applied, in the usual manner, to adjust reliability from one-month test samples to the anticipated 12-month sample (i.e., (12*r)/(1 + (11*r)). (Frey 2018)

These ICC(1) estimates (bounded between 0 and 1) were then logit-transformed and used to model the linear relationship between entity volume and logit reliability. By ranking predicted reliabilities across the complete range of potential volumes, we estimated the volume threshold that would correspond to ICC(1)=0.9 for an accountable entity.

ICC(1) is abbreviated by ICC in the results below.

References

McGraw KO, Wong S . (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1(1), 30–46.

Spearman-Brown Prophecy Formula. In: Frey B, eds. The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation. Vol. 4. Thousand Oaks, CA: SAGE Publications, Inc.; 2018. Available at: https://methods.sagepub.com/reference/the-sage-encyclopedia-of-educational-research-measurement-and-evaluation/i19400.xml
[Response Ends]

2a.11. For each level of reliability testing checked above, what were the statistical results from reliability testing?

For example, provide the percent agreement and kappa for the critical data elements, or distribution of reliability statistics from a signal-to-noise analysis. For score-level reliability testing, when using a signal-to-noise analysis, more than just one overall statistic should be reported (i.e., to demonstrate variation in reliability across providers). If a particular method yields only one statistic, this should be explained. In addition, reporting of results stratified by sample size is preferred (pg. 18, NQF Measure Evaluation Criteria).

[Response Begins]
The estimated mean split-half ICC using 37,172 CT exams collected from 16 hospitals was 0.99 (after Spearman-Brown adjustment to a 12-month data collection period). The number of exams per hospital in the one month of data used for testing ranged from 625 to 6,157 (mean=2,323); predicted reliability for 12 months exceeded 0.99 for every hospital.

The number of CT exams obtained during inpatient hospitalizations (n=15) in the one month of testing data ranged from 134-1,568 (mean 715); thus the number of CT exams from inpatient settings per hospital is estimated to vary from 1,608-18,816 for a 12-month period. For the individual hospitals, the predicted reliability for 12 months of inpatient CT exams exceeded 0.99 for every hospital during the testing phase.

The number of CT exams obtained during hospital outpatient encounters, including emergency department encounters, in the one month of testing data ranged from 119-4,978 (mean 1,608); thus the number of CT exams from outpatient settings per hospital is estimated to vary from 1,428-59,736 for a 12-month period. For the individual hospitals, the predicte reliability for 12 months of outpatient CT exams exceeded 0.99 for every hospital during the testing phase.

Table 2a-3. Number of CT exams in inpatient and outpatient (including emergency department) settings at each hospital.

Heading(H)	Inpatient	Outpatient
H1	494	131
H2	631	119
H3	707	145
H4	134	869
H5	1398	936
H6	1179	4978
H7	544	2166
H8	140	1962
H9	1028	1662
H10	732	2254
H11	N/A	2411
H12	1568	855
H13	306	1533
H14	673	2537
H15	1075	1601
H16	834	1570

Based on the method described above, a minimum of 28 CT exams are required to achieve 90% reliability.

[Response Ends]

2a.12. Interpret the results, in terms of how they demonstrate reliability.

(In other words, what do the results mean and what are the norms for the test conducted?)

[Response Begins]
According to the scale developed by Koo and Li, an ICC estimate greater than 0.90 may be interpreted as excellent reliability. (Koo 2016) Based on the mean ICC of 0.99, after Spearman-Brown adjustment to a 12-month reporting period, the measure is reliable at the hospital level. Given the high volume of CT, virtually no hospitals would fall below the minimum denominator to achieve ICC >0.90.

<u>Reference</u>

Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016 Jun;15(2):155-63. Epub 2016 Mar 31. Erratum in: J Chiropr Med. 2017 Dec;16(4):346.
[Response Ends]

Validity Testing

2b.01. Select the level of validity testing that was conducted.

[Response Begins]
Patient or Encounter-Level (data element validity must address ALL critical data elements)
Accountable Entity Level (e.g. hospitals, clinicians)
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)
[Response Ends]

2b.02. 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.

[Response Begins]
<u>Patient/encounter-level (data element) validity</u>

CT category: The measure uses an algorithm to assign each CT exam to one of 18 CT categories based on the diagnosis associated with the exam order (codified in ICD-10-CM codes) and procedure performed (codified in CPT® codes). We used criterion validity to compare agreement between the CT category assigned using this method versus a gold standard method based on expert review of the complete medical record (including notes from the visit when the exam was ordered, information provided as free text with the test order, and information included in the final, dictated radiology report) for a sample of CT exams from UCSF Health System (alpha testing).

For field-testing (beta testing), we did not have access to complete medical records, so we developed a second referent standard that determines CT category based on natural language processing of DICOM elements in the CT imaging data, including the reason for study, protocol name, study description, and the full radiology report including history, imaging findings, and diagnosis. This second referent standard was compared to the gold standard medical record review in the same sample of UCSF Health System CT exams and found to be accurate (sensitivity = 0.92, specificity = 0.97).

Patient size: Methods for measuring patient diameter on CT images have been previously validated including measuring patient size on axial images (Cheng 2013) and on coronal images (Christianson 2012). We relied on this published work and tested how often this method generated clinically plausible and non-missing values for size in testing data.

Radiation Dose: The measure uses dose length product (DLP), which gives the total radiation imparted to the patient by the CT machine. This is a standardized data element, generated by virtually (>99%) all CT machines, is well validated and used broadly to reflect the radiation dose delivered to the patient. (Kanal 2017, Smith-Bindman 2019.) Further, DLP is currently used in benchmarking in the U.S. and internationally (ACR–AAPM–SPR: Practice parameter, European Commission, Radiation Protection No. 185, ICRP Publication135). The proposed measure adjusted DLP for patient size to ensure that differences in patient mix would not result in differences in measure scores across reporting entities. While there are other dose metrics used in some settings to measure radiation dose (such as size-specific dose estimate (SSDE) or effective dose), these are not suitable for a reliable quality measurement because they are not universally or automatically generated by the CT machine, do not reflect the total dose absorbed by the patient (the most clinically relevant measure), and would not adequately remove differences in measure score that are the result of patient case mix. We relied on this published work and tested how often this method generated clinically plausible and non-missing values for radiation dose in testing data.

Size-Adjusted Radiation Dose: We describe the validation of our method to risk-adjust radiation dose based on patient size in section 2b.26. In summary, when out-of-range rates are unadjusted for patient size, we observe failure rates that are strongly associated with size, with almost all failures occurring in larger patients. When failure rates are adjusted for size, there is no association. Using field testing data, we assessed whether we could calculate size-adjusted radiation dose within a plausible range and quantified missing data.

Global noise: The approach we used for measuring global noise in CT images was an adaptation of previously validated approaches. (Christianson 2017, Malkus 2017) These adaptations were motivated by the need to generate a summary value for global noise for the CT exam in exams with multiple scans, and to adjust for slice thickness, each validated in the Image Quality Study (described below). We also reviewed the literature for association between noise calculations in DICOM data and phantom measurements of noise and human readers' assessment of image quality. Next, using field-testing data, we assessed whether we could calculate global noise within a plausible range and quantified missing data.

We also calculated the correlation between global noise and physician dissatisfaction with image quality, a valid metric of quality as described and explained below, using data from the Image Quality Study (described below). Lastly, we explored the rate of physician dissatisfaction in CT exams that exceeded global noise thresholds. Dissatisfaction is defined as a physician rating CT image quality as "poor" or "marginally acceptable."

Thresholds for "out-of-range" values to define numerator: We used radiologists' satisfaction with CT images as a basis for establishing the maximum radiation dose and minimum image quality thresholds for each CT category. In clinical practice, radiologists are responsible for ensuring the images they interpret are of acceptable quality to allow them to make accurate diagnoses. If they are not satisfied with the image quality, they must ask that the exam be repeated.

Early in development of the proposed measure, we conducted an Image Quality Study to understand the relationship between radiation dose, global noise, and physician satisfaction. We first compiled a test set of 740 CT exams covering a wide range of anatomic areas and clinical indications. The test cases were sampled from the UCSF International CT Dose Registry and were selected from across the CT categories, and within each CT category, images were obtained across the entire observed dose distribution with over sampling of images at the low dose range where we suspected any issues with image quality would occur. CTs were selected from diverse organizations. 125 radiologists from diverse practice settings each graded 200 exams, resulting in 25,000 interpretations used to determine the thresholds for radiation dose and global noise. For each exam, the radiologist reader was asked to characterize the image quality on a four-point scale:

- **Excellent**: the images provide the needed information
- **Adequate**: the images are acceptable but not excellent; you would re-scan and change the parameters for a higher quality if it were easy to repeat, but if not, this is good enough

- **Marginally acceptable**: image quality is less than ideal and may compromise diagnostic quality; if the patient cannot easily be re-scanned you will interpret this, but would change parameters for future scans of this type
- **Poor**: image quality is not adequate for diagnosis and the scan should be repeated

Overall, 49% of exams were rated excellent, 40% adequate, 8% marginally acceptable, and 3% poor for clinical interpretation. Exams rated as excellent or adequate were considered of acceptable quality, and exams rated as either marginally acceptable or poor were considered unacceptable (to set generous thresholds favoring better image quality).

We used the radiologists' interpretations to set the thresholds for size-adjusted radiation dose and global noise. The maximum size-adjusted radiation dose threshold was set at the dose level within each CT category where 90% or more of radiologists graded the exam as acceptable quality (excellent or adequate). Doses above this level expose patients to harm without increasing image quality, as 90% of radiologists are already satisfied with the image quality. If a CT category had no observed threshold because radiologists were satisfied at every dose level, we used the median dose from the UCSF International CT Dose Registry as the threshold. This decision to use the median was based on extensive discussion with the Technical Expert Panel.

The minimum floor for image quality was set at the level where 25% or more of radiologists graded the exam as unacceptable (marginally acceptable or poor). Image quality at or below this level is considered inadequate. This threshold was discussed and agreed upon by the Technical Expert Panel, with the general view that, as images may be sent to many different radiologists to interpret within large practices, at least 75% should feel comfortable interpreting images with the quality level that is within range in this measure. If 25% or more of radiologists are uncomfortable with the quality of images, than the exam should be graded as unacceptable. Image quality is measured using global noise (Makkus 2017, Christiansen 2015) adjusted by slice thickness (Alshipli 2017), where higher global noise generally reflects worse quality. If a CT category had no observed noise threshold, we set the threshold based on the literature or based on closely related categories. (For example, the CT category cardiac low dose had no observed threshold; thus we used the observed threshold from the chest low dose category, which was observed). The approach to setting thresholds was influenced and strongly supported by our Technical Expert Panel.

Empirical validity testing: Gold standard comparison

Lastly, we validated the eCQM output (encounter-level validity) against medical record review using field testing data collected from electronic clinical data systems from 7 health systems. The "medical record review" is a human-reviewed indicator of whether the size-adjusted radiation dose or global noise of each sampled exam exceeds predetermined thresholds, thus constituting a “gold standard.”

Accountable entity-level (measure score) validity

Systematic assessment of face validity of measure score as an indicator of quality

We assessed measure score face validity through a 6-question poll to the Technical Expert Panel (TEP) assembled for the creation of this measure, administered by Co-Investigator Dr. Patrick Romano. The TEP represents a diverse group of clinicians (N=10), patient advocates (N=2), and leaders of medical specialty societies, payers, and healthcare safety and accrediting organizations. TEP members were identified by reaching out to key stakeholder organizations and advocates and identifying researchers who had contributed to the relevant literature.

The 6-question poll included the following face validity questions:

1. Do you agree that radiation dose is a relevant metric of quality for CT imaging? (to assess face validity of that data element)

2. Do you agree that image noise is a relevant metric of quality for CT imaging? (to assess face validity of that data element)

- We clarified during polling that this question was not assessing noise as a standalone metric, but as part of a balancing measure of radiation dose and noise.

3. Do you agree that size is an appropriate method for adjusting for radiation dose for a given indication? (to assess face validity of the risk-adjustment approach)

4. Do you agree that performance on this measure of radiation dose and image quality, adjusted for size, stratified by indication, is a representation of quality? (to assess face validity of the measure score)

5. Do you agree that if this measure is implemented in the CMS hospital programs that this measure is likely to lead to reductions in radiation dose while maintaining adequate image quality?

Technical Expert Panel members include:

- Niall Brennan, MPP, CEO, Health Care Cost Institute
- Jesse Chusid, MD, MBA, Chair, Department of Radiology Senior Vice President, Imaging Services, Northwell Health
- Melissa Danforth, Vice President of Health Care Ratings, The Leapfrog Group
- Tricia Elliot, MBA, CPHQ, Director, Quality Measurement, Joint Commission
- Mohammad Fakih, MD, MPH, Chief Quality Officer Clinical & Network Services, Ascension Healthcare
- Jeph Herrin, PhD, Adjunct Assistant Professor, Yale University
- Jay Leonard Lichtenfeld, MD, MACP, Independent Consultant, Formerly Deputy Chief Medical Officer American Cancer Society, Inc.
- Leelakrishna Nallamshetty, MD, Associate Chief Medical Officer, Radiology Partners
- Matthew Nielsen, MD, MS, Professor and Chair of Urology, UNC Gillings School of Global Public Health
- Debra Ritzwoller, PhD, Patient Advocate and Health Economist (Patient Representative)
- Lewis Sandy, MD, Executive Vice President, Clinical Advancement, UnitedHealth Group
- Mary Suzanne Schrandt, JD, Patient Advocate (Patient Representative)
- James Anthony Seibert, PhD, Professor, University of California, Davis
- Robert H. Sherrier, MD, Chief Consultant, Diagnostic Services, Veterans Affairs
- Arjun Venkatesh, MD, MBA, MHS, Associate Professor, Emergency Medicine, Yale School of Medicine

References

ACR–AAPM–SPR: Practice parameter for diagnostic reference levels and achievable doses in medical x-ray imaging. Revised 2018.

Cheng PM. Automated estimation of abdominal effective diameter for body size normalization of CT dose. J Digit Imaging. 2013 Jun;26(3):406-11.

Christianson O, Li X, Frush D, Samei E. Automated size-specific CT dose monitoring program: assessing variability in CT dose. Med Phys. 2012 Nov;39(11):7131-9.

Christianson O, Winslow J, Frush DP, Samei E. Automated Technique to Measure Noise in Clinical CT Examinations. AJR Am J Roentgenol. 2015 Jul;205(1):W93-9.

European Commission, Radiation Protection No. 185, European guidelines on diagnostic reference levels for paediatric imaging. 2018.

ICRP, 2017. Diagnostic reference levels in medical imaging. ICRP Publication 135. Ann. ICRP 46(1).

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;284(1):120-133.

Malkus A, Szczykutowicz TP. A method to extract image noise level from patient images in CT. Med Phys. 2017 Jun;44(6):2173-2184. doi: 10.1002/mp.12240. Epub 2017 Apr 25.

Alshipli M and Kabir NA 2017 J. Phys.: Conf. Ser. 851 012005.

[Response Ends]

2b.03. Provide the statistical results from validity testing.

Examples may include correlations or t-test results.

[Response Begins]
Patient/encounter-level (data element) validity

CT category: In alpha testing, we validated our method of assigning CT category based on diagnosis and procedure codes against a gold standard. The results, weighted by the distribution of CT categories in the UCSF International CT Dose Registry, were: sensitivity = 0.86 and specificity = 0.96 (n=978 CT exams).

When tested across 16 hospitals, the correct classification rate was 92% on average and varied from 88-97%.

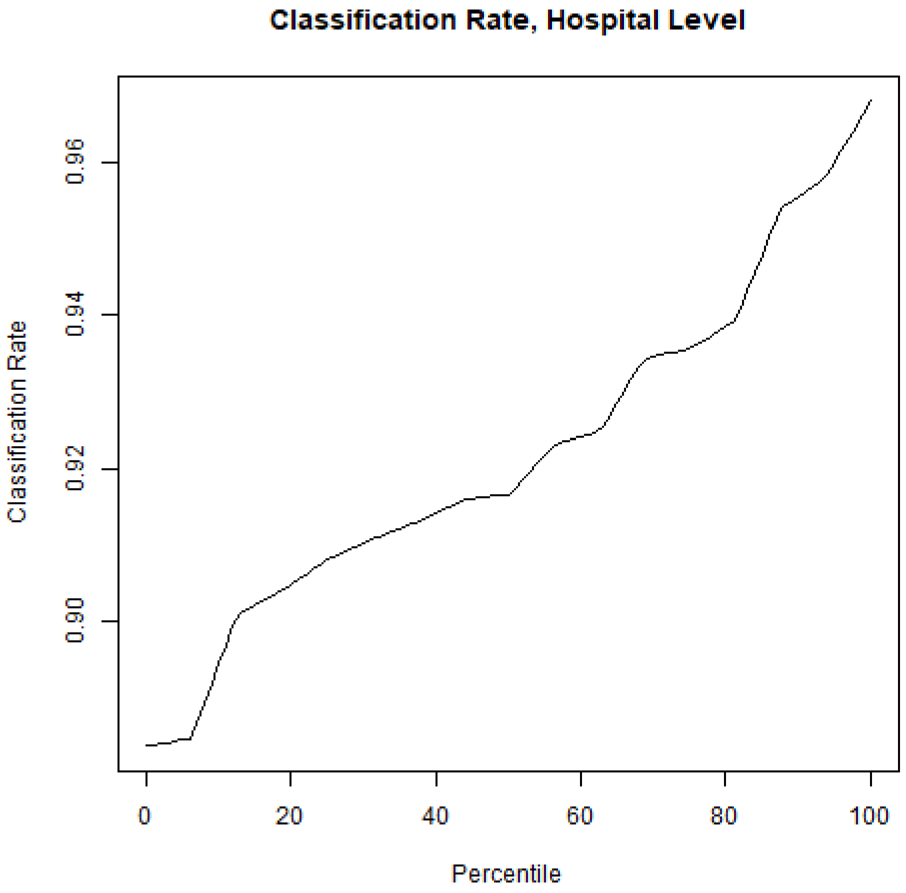


Figure 2b-1. Decile of correct classification rate, hospital level.

Size-Adjusted Radiation Dose: In field testing data, size-adjusted radiation dose could be calculated and was within plausible range for 99% of CT exams and was missing for 0.4% of exams.

Global Noise: Global noise measurements based on DICOM data are highly predictive of phantom measurements of noise and human readers' assessment of image quality (Christianson 2015.) Global noise could be calculated and was within a plausible range for 100% of CT exams in field-testing. Global noise was missing for 0.01% of examinations.

The correlation between noise and physician dissatisfaction with image quality is 0.37 overall based on the image quality study (n=727 CT exams).

Based on the field-testing data, there were few exams which exceeded the global noise thresholds. There were 4 CT categories with exams in which global noise exceeded the allowable threshold; average physician dissatisfaction rates for exams below and above thresholds for those CT categories are shown in the table below. For other CT categories, exams were not observed above the threshold.

Table 2b-1. Dissatisfaction rates for CT exams below and above the global noise threshold, and the proportion of exams above threshold, for CT categories with exams in which global noise exceeded allowable thresholds.

*	Dissatisfaction rate for exams below noise threshold	Dissatisfaction rate for exams above noise threshold	Proportion of exams above noise threshold
Chest Low Dose	0.20	0.47	0.05
Chest Routine Dose	0.11	0.28	0.03
Cardiac High Dose or Chest High Dose	0.11	0.35	0.03
Thoracic or Lumbar Spine	0.13	0.40	0.07

*Cell Intentionally left blank

Empirical validity testing: Gold standard comparison

The results of the medical record review were compared with the results of the eCQM computation by selecting a sample of exams (N=7,000) representative of exams generated by the 16 hospitals across 7 health systems. The out-of-range results (measure score) from the medical record review and the eCQM computation were identical with no discrepancies between the two approaches, indicating a correct and robust implementation of the measure logic.

Accountable entity-level (measure score) validity

Systematic assessment of face validity of measure score as an indicator of quality

No TEP members abstained from voting. The results were as follows:

1. Do you agree that radiation dose is a relevant metric of quality for CT imaging?

- 100% agreement

2. Do you agree that image noise is a relevant metric of quality for CT imaging?

- 100% agreement

3. Do you agree that size is an appropriate method for adjusting for radiation dose for a given indication?

- 100% agreement

4. Do you agree that performance on this measure of radiation dose and image quality, adjusted for size, stratified by indication, is a representation of quality?

- **100% agreement**

5. Do you agree that if this measure is implemented in the CMS hospital programs that this measure is likely to lead to reductions in radiation dose while maintaining adequate image quality?

- 100% agreement

[Response Ends]

2b.04. Provide 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?)

[Response Begins]
Patient/encounter-level (data element) validity

The measure algorithm assigns CT category with 92% accuracy when compared to a validated referent standard.

Size-adjusted radiation dose and global noise have face validity as metrics of quality, as assessed by our Technical Expert Panel, and could be calculated with plausible ranges for virtually all exams in field-testing. Moderate correlation between global noise and physician dissatisfaction with the quality of CT images, another valid quality indicator, supports global noise as a proxy measurement of image quality. and for CT categories where there were exams exceeding global noise thresholds, physician dissatisfaction for those out-of-range exams was considerable (28-47%).

The eCQM computed identical results for a sample of 7,000 CT exams, compared to medical record review.

Accountable entity-level (measure score) validity

100% of our Technical Expert Panel supported the face validity of the measure score, agreeing unanimously that *"performance on this measure of radiation dose and image quality, adjusted for size, stratified by indication, is a representation of quality."*

These results provide evidence that the measure as specified is a valid representation of quality, and the measure score accurately differentiates good performance from poor performance.
[Response Ends]

2b.05. 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 in Importance to Measure and Report: Gap in Care/Disparities.

[Response Begins]
We consider it clinically meaningful to be able to detect entities whose prevalence of "out-of-range" exams (either by size-adjusted dose or by noise) is at least 5 percentage points above or below the average national performance. For testing purposes, this threshold refers to out-of-range prevalence values above 38% or below 28%.

To compute the minimal sample size necessary to be able to detect such out-of-range prevalence with 0.8 power, 0.05 level of significance, we use the equations

$$0.8 = \Pr[Z < z_{0.025} - H(0.33,0.38) * \mathrm{sqrt}(N_{high})]$$
$$0.8 = \Pr[Z > z_{0.025} - H(0.33,0.28) * \mathrm{sqrt}(N_{low})]$$

Where Z is a normally distributed random variable, $z_{0.025}$ is the 2.5th percentile of a normally-distributed random variable, N_{high} is the minimal required sample size to detect an out-of-range rate of 38%, N_{low} is the minimal required sample size to detect an out-of-range rate of 28%, and

$$H(x,y)=2 * \arcsin(\mathrm{sqrt}(x)) - 2 * \arcsin(\mathrm{sqrt}(y))$$

We then compared these estimated values of N_{high} and N_{low} against the observed distribution of entity-specific volumes in our test data, adjusted to a 12-month reporting period.

Finally, we empirically estimated the distribution of measure scores across the entities that participated in pilot testing, and assessed the statistical significance of their observed values, relative to the national average prevalence of "out-of-range" exams (33%).

[Response Ends]

2b.06. Describe the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities.

Examples may include 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.

[Response Begins]
The required sample size to detect deviation of 5 percentage points above the mean (i.e., 38%, with 80% power) is 566. The required sample size to detect deviation of 5 percentage points below the mean (i.e., 28%, with 80% power) is 524. Essentially 100% of participating hospitals would meet this requirement, based on our pilot data.

The empirically observed distribution of measure scores from our test data is shown in Table 2b-2 below. At the hospital level (n=16), we were able to identify 7 hospitals with significantly better than average performance, based on the 95% confidence intervals surrounding the estimated values. These hospitals had "out-of-range" prevalence between 24% and 30%. We were able to identify 7 hospitals with significantly worse than average performance; these hospitals had "out-of-range" prevalence between 37% and 57%. These results are shown graphically in Figure 2b-3.

Table 2b-2: Measure score values with confidence intervals by hospital. The average width of these confidence intervals is 4 percentage points.

*	N	Proportion Out-of-Range	Lower Confidence (95%)	Upper Confidence (95%)
H2	750	0.24	*0.21*	*0.28*
H6	6157	0.24	*0.23*	*0.25*

*	N	Proportion Out-of-Range	Lower Confidence (95%)	Upper Confidence (95%)
H13	1839	0.27	*0.25*	*0.29*
H16	2404	0.27	*0.25*	*0.28*
H9	2690	0.29	*0.27*	*0.31*
H14	3210	0.29	*0.27*	*0.31*
H12	2423	0.30	*0.28*	*0.32*
H1	625	0.35	0.31	0.39
H3	852	0.36	0.33	0.39
H8	2102	0.37	*0.35*	*0.39*
H7	2710	0.37	*0.35*	*0.39*
H10	2986	0.37	*0.35*	*0.39*
H5	2334	0.40	*0.38*	*0.42*
H11	2411	0.43	*0.41*	*0.45*
H15	2676	0.44	*0.42*	*0.46*
H4	1003	0.57	*0.54*	*0.60*

*Cell intentionally left blank

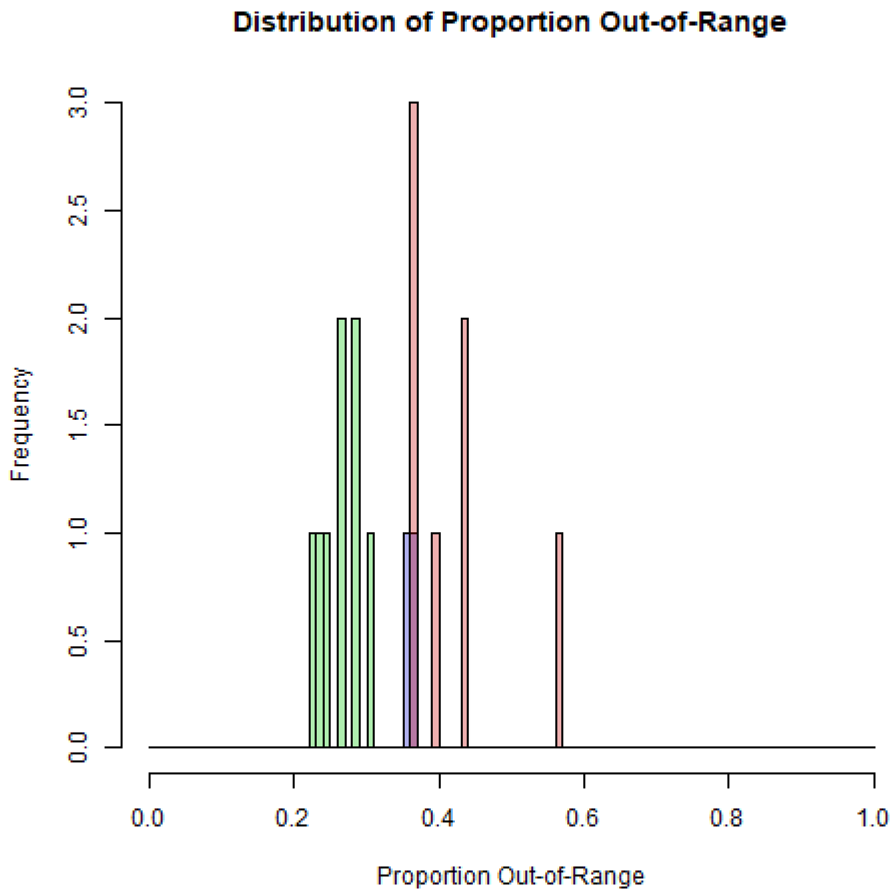


Figure 2b-2. Measure score distributions for hospitals overall, color-coded as follows: confidence interval lies above 33% (red); contains 33% (blue); and lies below 33% (green).

[Response Ends]

2b.07. Provide 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.

In other words, what do the results mean in terms of statistical and meaningful differences?

[Response Begins]
Of the hospitals assessed, all are expected to meet the sample size requirements for discriminating meaningful differences in performance if one year of data is provided.

Of the hospitals assessed, all of those with observed measure scores at least 5 percentage points from the mean (33%) had confidence intervals not containing 33%, indicating high ability in practice to detect clinician groups and hospitals that deviate meaningfully from the mean. Of the hospitals assessed, 6 had an observed deviation from 33% out-of-range prevalence of less than 5 percentage points. Of the remaining 10 hospitals with detectable difference, all had confidence intervals not containing 33%, indicating high ability in practice to detect hospital which deviate from the mean.

We have an ability to detect even smaller differences in measure score performance (less than 5 percentage points), and over time this could be reported to further drive quality improvements.
[Response Ends]

2b.08. Describe the method of testing conducted to identify the extent and distribution of missing data (or non-response) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders). Include how the specified handling of missing data minimizes bias.

Describe the steps—do not just name a method; what statistical analysis was used.

[Response Begins]
Of exams submitted for testing, 3,349 were removed from analysis due to missing data (compared with 37,172 which had full data). Missing data can come in one of the following forms:

1) Missing radiation dose (due to missing Radiation Dose Structured Report, RDSR)

2) Missing patient diameter (failure of diameter calculation algorithm)

3) Missing global noise (failure of noise calculation algorithm)

Exams can also be excluded if the patient's age is missing, though patient age was available for all exams in testing data.

To assess the potential impact of missing data on measure scores, we first estimated the percentage of CT scans with missing data at the accountable entity level and identified the extent to which missing data were concentrated at a small number of accountable entities.

Next, we compared the distributions of CT category and patient diameter between CT scans that would be excluded due to missing data (defined as any scan with missing radiation dose, missing patient diameter, or missing global noise) and CT scans that would be retained in the analysis ("non-missing data"). Due to the large sample size of our testing data, we expect even modest, clinically insignificant differences in these distributions to be statistically significant. Thus, rather than perform statistical testing, we focus on the clinical significance of: (1) differences in probability distribution of CT categories between missing and non-missing data; and (2) differences in patient diameter deciles between missing and non-missing data. If data are "missing at random," then the distributions of both CT category and patient diameter should be similar between the CT scans with missing data and those with non-missing dose data.
[Response Ends]

2b.09. Provide the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data.

For example, provide results of sensitivity analysis of the effect of various rules for missing data/non-response. If no empirical sensitivity analysis was conducted, identify the approaches for handling missing data that were considered and benefits and drawbacks of each).

[Response Begins]
Of the 3,349 CT scans removed due to missing data, 93% were removed due to missing radiation dose and 7% were removed for other reasons. The tables below show missing data rates at the accountable entity level and compare the distributions of CT categories and patient diameters (size) between scans with missing data and scans with non-missing data.

Table 2b-3. Number of missing data (by type) across hospitals.

Hospital	Non-Missing Sample Size	Missing Radiation Dose	Missing Global Noise	Missing Patient Diameter
H1	625	1	0	0
H2	750	0	0	0
H3	852	21	0	0
H4	1003	63	0	0
H5	2334	142	0	0
H6	6157	1761	3	0
H7	2710	372	0	0
H8	2102	12	0	0
H9	2690	26	0	0
H10	2986	290	4	0
H11	2411	2	0	0
H12	2423	2	0	0
H13	1839	0	0	0
H14	3210	408	0	233
H15	2676	4	0	0
H16	2404	5	0	0

Table 2b-4. Probability distributions of CT category among missing data group and among non-missing data group.

*	Non-Missing Data	Missing Data
Abdomen and Pelvis Low Dose	2%	2%
Abdomen and Pelvis Routine Dose	22%	23%
Abdomen and Pelvis High Dose	5%	4%
Chest Low Dose	1%	1%
Chest Routine Dose	13%	12%
Cardiac Low Dose	3%	1%
Cardiac Routine Dose	9%	12%
Cardiac High Dose or Chest High Dose	0%	0%

*	Non-Missing Data	Missing Data
Thoracic or Lumbar Spine	1%	1%
Simultaneous Thoracic and Lumbar Spine	0%	0%
Simultaneous Chest and Abdomen	10%	11%
Head Low Dose	3%	2%
Head Routine Dose	16%	15%
Head High Dose	0%	0%
Neck or Cervical Spine	3%	3%
Simultaneous Head and Neck Routine Dose	7%	8%
Simultaneous Head and Neck High Dose	0%	0%
Extremity	3%	3%

*Cell left intentionally blank

Table 2b-5. Deciles of patient diameter (in millimeters) of head exams (including CT categories Head Low Dose, Head Routine Dose, Head High Dose, Simultaneous Head and Neck Routine Dose, and Simultaneous Head and Neck High Dose) among missing data group and among non-missing data group. Values shown on patient effective diameter in millimeters.

Percentile	Non-Missing Data	Missing Data*
10%	131	128
20%	145	147
30%	154	155
40%	160	163
50%	166	169
60%	171	174
70%	176	177
80%	182	183
90%	195	193

*Exams with missing patient diameter were excluded from this specific analysis.

Table 2b-6. Deciles of patient diameter (in millimeters) of trunk exams (all exams not represented in the “head exams” table above) among missing data group and among non-missing data group. Values shown on patient effective diameter in millimeters.

Percentile	Non-Missing Data	Missing Data*
10%	190	203
20%	230	234
30%	250	255
40%	266	271
50%	281	285
60%	294	299
70%	309	314
80%	327	331
90%	353	356

*Exams with missing patient diameter were excluded from this specific analysis.

[Response Ends]

2b.10. Provide 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.

In other words, 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 was conducted, justify the selected approach for missing data.

[Response Begins]
Our results show that only 8% of CT scans from our test sites reported missing data, meaning that the impact of missing data on the measure overall is low. The majority of CT scans with missing data do not have radiation dose available, but do have CT category, global noise, and patient diameter available.

Most accountable entities had very little missing data, indicating that the problem of "missing data" is within the capacity of accountable entities to resolve. Therefore, the developer recommends that "missing data" rates should be tracked, and entities should be expected to reduce their "missing data" rates to zero over time. For example, the hospital with the highest missing radiation dose data (H6) came on board rather late in our testing period. Thus, unlike other sites, they did not have sufficient time to modify their CT machines to save the radiation dose structured report (RDSR), the digitized, structured summary providing the total radiation output during the CT exam. Many CT machines require such modification to save RDSRs; this is discussed elsewhere in this application. This site reported that if it had started earlier, they probably could have adjusted their systems and thus would have had less missing radiation dose data. Because our testing period was only one month in duration, there was insufficient time for all sites to modify their systems to save all RDSR (radiation dose) data.

Finally, assessment of the distributions of CT category and patient diameter among missing data shows that they are very similar to those in non-missing data, and thus missing data are very unlikely to bias results at the accountable entity level.
[Response Ends]

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 eCQMs). 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.

2b.11. Indicate whether there is more than one set of specifications for this measure.

[Response Begins]
No, there is only one set of specifications for this measure
[Response Ends]

2b.12. 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. Indicate what statistical analysis was used.

[Response Begins]
[Response Ends]

2b.13. Provide the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications.

Examples may include correlation, and/or rank order.

[Response Begins]
[Response Ends]

2b.14. Provide your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications.

In other words, what do the results mean and what are the norms for the test conducted.

[Response Begins]
[Response Ends]

2b.15. Indicate whether the measure uses exclusions.

[Response Begins]
Yes, the measure uses exclusions.
[Response Ends]

2b.16. Describe the method of testing exclusions and what was tested.

Describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used?

[Response Begins]
The only exams submitted subject to exclusion are exams scanning an "uncommon" combination of multiple body parts. "Common" combinations of body parts are sorted into one of the CT Dose and Image Quality Categories – for example, Simultaneous Chest and Abdomen, Simultaneous Thoracic and Lumbar Spine, Simultaneous Head and Neck Routine Dose, and Simultaneous Head and Neck High Dose. These uncommon combinations of multiple body parts are not part of the population of interest, and thus our measure has no mechanism for computing whether their radiation dose or global noise are out-of-range. The impact of these exclusions thus cannot be precisely calculated. We will, however, assess a range of possible impacts, comparing the performance score of each hospital in our testing data under three circumstances:

1) Performance score calculated as intended by our proposed measure.

2) Performance score if uncommon combinations of multiple body parts were hypothetically included, and they were all out-of-range.

3) Performance score if uncommon combinations of multiple body parts were hypothetically included, and none were out-of-range.
[Response Ends]

2b.17. Provide the statistical results from testing exclusions.

Include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores.

[Response Begins]
Across the testing data, there were a total of 1,690 exams scanning uncommon combinations of multiple body parts, compared to 37,172 exams that were included for analysis.

Table 2b-7. Number of exclusions (and inclusions) in each of the 16 hospitals.

*	Included CT exams	Excluded CT exams, Number	Excluded CT exams, %
H1	625	23	3.5%
H2	750	30	3.8%
H3	852	26	3.0%
H4	1003	24	2.3%
H5	2334	85	3.5%
H6	6157	308	4.8%
H7	2710	172	6.0%
H8	2102	112	5.1%
H9	2690	144	5.1%
H10	2986	158	5.0%
H11	2411	33	1.4%
H12	2423	114	4.5%
H13	1839	100	5.2%
H14	3210	151	4.5%
H15	2676	106	3.8%
H16	2404	104	4.1%

*Cell Intentionally left blank

Table 2b-8. Measure scores of the 16 hospitals, under each of the three hypothetical circumstances.

*	Observed failure rate	Theoretical failure rate (excluded data all out-of-range)	Theoretical failure rate (excluded data not out-of-range)
H1	35%	38%	34%
H2	24%	27%	23%
H3	36%	38%	35%
H4	57%	58%	56%
H5	40%	42%	38%
H6	24%	27%	23%
H7	37%	41%	35%
H8	37%	40%	35%

*	Observed failure rate	Theoretical failure rate (excluded data all out-of-range)	Theoretical failure rate (excluded data not out-of-range)
H9	29%	33%	28%
H10	37%	40%	35%
H11	43%	44%	43%
H12	30%	33%	29%
H13	27%	31%	25%
H14	29%	32%	28%
H15	44%	46%	42%
H16	27%	30%	26%

[Response Ends]

2b.18. Provide your interpretation of the results, in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results.

In other words, the value outweighs the burden of increased data collection and analysis. Note: If patient preference is an exclusion, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion.

[Response Begins]
The choice to exclude uncommon combinations of multiple body parts is due to a lack of sufficient data that would allow us to construct a reasonable out-of-range threshold for such exams, resulting in their removal from the population of interest. The results of 2b.17 indicate that the prevalence of exclusions is small enough that their impact on performance scores is clinically insignificant.
[Response Ends]

2b.19. Check all methods used to address risk factors.

[Response Begins]
Statistical risk model with risk factors (specify number of risk factors)
Stratification by risk category (specify number of categories)
The measure has 18 risk categories.
1 risk factor (patient size)
[Response Ends]

2b.20. If using statistical risk models, provide detailed risk model specifications, including the risk model method, risk factors, risk factor data sources, coefficients, equations, codes with descriptors, and definitions.

[Response Begins]
The means by which a CT examination is determined to be “out-of-range" with respect to radiation dose is measured by observing whether its patient size-adjusted radiation dose exceeds a pre-determined evidence-based threshold. The value of this size-adjusted radiation dose is calculated with the following equation for any given exam:

$$D_A = D_R * \exp(-(d-d_k) * \beta_k)$$

Where...

$D_A$is the size-adjusted radiation dose of the exam

$D_R$is the radiation dose of the exam, without adjustment

d is the diameter of the anatomic area being examined

d_k is the “expected diameter” of the CT category associated with the exam. This “expected diameter” is equal to the median diameter of all exams associated with the CT category in the UCSF International CT Dose Registry containing 6.5 million exams from 161 institutions.

β_k is the “size-adjustment coefficient” of the CT category associated with the exam. This “size-adjustment coefficient” is the slope parameter of a collection of log-transformed linear regression models fit using the UCSF Registry. A total of 18 models were fit, each using data from one of the CT Dose and Image Quality Categories. The models are parametrized such that, in the kth model and associated dataset, for the jth observation, from the ith hospital, we define:

$$\log(\{D_R\}_{ij}) = \{\beta_0\}_k + \beta_k * d_{ij} + \{z_i\}_k + \varepsilon_{ij}$$

Where D_R and d are respectively the radiation dose without adjustment and diameter of the anatomic area being examined, β_0 is an intercept term, z is a random effect indicating variation due to the hospital at which the exam was performed, and ε is the residual variation. We restrict the value of $\beta_k$to be greater than 0; when it is less than 0, it is set to 0 and no adjustment is performed. For the estimated values of β_k across CT categories (strata), please see 2b.30 below.

The intended interpretation of D_A is the “expected radiation dose of the exam if the diameter of the anatomic area being examined were equal to the population-level median.”

[Response Ends]

2b.21. If an outcome or resource use measure is not risk-adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (i.e., case mix) is not needed to achieve fair comparisons across measured entities.

[Response Begins]
N/A – the outcome is risk adjusted
[Response Ends]

2b.22. Select all applicable resources and methods used to develop the conceptual model of how social risk impacts this outcome.

[Response Begins]
Published literature
[Response Ends]

2b.23. Describe the conceptual and statistical methods and criteria used to test and select patient-level risk factors (e.g., clinical factors, social risk factors) used in the statistical risk model or for stratification by risk.

Please be sure to address the following: potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10 or other statistical tests; correlation of x or higher. Patient factors should be present at the start of care, if applicable. Also discuss any “ordering” of risk factor inclusion; note whether social risk factors are added after all clinical factors. Discuss any considerations regarding data sources (e.g., availability, specificity).

[Response Begins]
A comprehensive review of the published literature was performed to inform the design of this measure, including the identification of patient-level or exam-level risk factors. This review included all of the literature cited by the American College of Radiology (ACR) in its #3621 submission to NQF, as well as additional literature not cited by the ACR. The UCSF measure development team has actively contributed to this literature. Only patient and machine factors present at the

start of care were considered in this review. Because the current measure was designed as an eCQM, we do not have the ability to test risk factors that were not supported by our conceptual model and literature review.

Because decisions are made at the level of patient groups, rather than individual patients, the logic model does not include varying technical parameters for individual patients. To the extent they have been studied, social factors including race/ethnicity and socioeconomic status are not predictive of radiation dose for CT exams. Messenger et al. (2016) used a cohort of 3442 CTs for calcium scoring to assess the relationship between effective dose (dose length product multiplied by a fixed conversion factor) and a variety of patient characteristics including age, sex, ethnic group, and body mass index. Each continuous independent variable was converted into categories, and the means of each category was reported. They reported no substantial differences between effective dose and any categorical/categorized patient characteristic, except age among those >75 years old.

There is a potential concern that the age of CT machines may be associated with increased radiation dose, as newer machines sometimes offer dose reduction software. Theoretically, this could lead to higher doses and poorer performance on the measure in safety-net settings that may have older machines. However, there is no evidence to support a strong association between CT machine factors, including the age of the machine, and increased radiation dose. (Catalano 2007) In a study of over 2 million CT exams from 151 institutions, including 290 machines from the four largest machine manufacturers and 49 machine models, Smith-Bindman et al. evaluated the contribution of machine characteristics to radiation dose variation. (Smith-Bindman 2019). They observed statistical significance for nearly all variables assessed due to large sample size, but the effect sizes for patient sex and radiation dose, and patient age and radiation dose, were both negligible. The effect size of patient size, measured using effective diameter, was large and substantial in all anatomic areas studied. For chest exams, for example, one standard deviation increase in effective diameter was associated with an increase of 36% in effective dose. For abdomen exams, this effect size was 47%. No patient or machine characteristics explained the variability of effective dose to any notable extent. The authors concluded that differences in observed dose were almost entirely associated with how institutions used the machines, reflecting different choices of technical scanning parameters and not the machines themselves.

Another study showed, among institutions performing low-dose CT exams for lung cancer screening, a significant proportion of institutions and patients had doses that exceeded guideline-recommended dose levels. However, the type of institution, including whether the hospital was a public hospital, was not associated with the radiation dose used. (Demb 2019.) Lastly, several analyses are underway using data from the UCSF International CT Dose Registry demonstrating that optimized doses have been observed across all machine makes and models in the Registry, regardless of machine characteristics.

References

Catalano C, Francone M, Ascarelli A, Mangia M, Iacucci I, Passariello R. Optimizing radiation dose and image quality. Eur Radiol. 2007;17 Suppl 6:F26-32.

Demb J, Chu P, Yu S, Whitebird R, Solberg L, Miglioretti DL, Smith-Bindman R. Analysis of Computed Tomography Radiation Doses Used for Lung Cancer Screening Scans. JAMA internal medicine 2019;179(12):1650-1657. doi: 10.1001/jamainternmed.2019.3893

Freeman K, Strauchler D, Miller TS. Impact of socioeconomic status on ionizing radiation exposure from medical imaging in children. J Am Coll Radiol. 2012 Nov;9(11):799-807. doi: 10.1016/j.jacr.2012.06.005. PMID: 23122347.

Hou, J.K., Malaty, H.M. & Thirumurthi, S. Radiation Exposure from Diagnostic Imaging Studies Among Patients with Inflammatory Bowel Disease in a Safety-Net Health-Care System. Dig Dis Sci 59, 546–553 (2014). https://doi.org/10.1007/s10620-013-2852-1

Messenger B, Li D, Nasir K, Carr JJ, Blankstein R, Budoff MJ. Coronary calcium scans and radiation exposure in the multi-ethnic study of atherosclerosis. Int J Cardiovasc Imaging. 2016 Mar;32(3):525-9. doi: 10.1007/s10554-015-0799-3. Epub 2015 Oct 29. PMID: 26515964.

Strauchler D, Freeman K, Miller TS. The impact of socioeconomic status and comorbid medical conditions on ionizing radiation exposure from diagnostic medical imaging in adults. J Am Coll Radiol. 2012 Jan;9(1):58-63. doi: 10.1016/j.jacr.2011.07.009. PMID: 22221637.

Smith-Bindman R, Wang Y, Chu P, et al. International variation in radiation dose for computed tomography examinations: prospective cohort study. BMJ. 2019;364:k4931.
[Response Ends]

2b.24. Detail the statistical results of the analyses used to test and select risk factors for inclusion in or exclusion from the risk model/stratification.

[Response Begins]
Based on the logic model and literature review described above, only one risk factor (patient size) was selected for inclusion in the risk model. The logic model and literature review do not support inclusion of any other risk factors. This decision was endorsed by our Technical Expert Panel, as described above.
[Response Ends]

2b.25. Describe the analyses and interpretation resulting in the decision to select or not select social risk factors.

Examples may include prevalence of the factor across measured entities, availability of the data source, empirical association with the outcome, contribution of unique variation in the outcome, or assessment of between-unit effects and within-unit effects. Also describe the impact of adjusting for risk (or making no adjustment) on providers at high or low extremes of risk.

[Response Begins]
Our decision to not include social risk factors was based on review of the literature and finding no empirical evidence supporting the influence of social risk factors (including provider-level proxies for social risk factors, such as machine characteristics) on radiation dose. Providers who see a disproportionate number of patients from disadvantaged backgrounds, or in safety-net settings which may have older CT machines, are not expected to fail the measure more frequently because of these factors.
[Response Ends]

2b.26. 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). Provide the statistical results from testing the approach to control for differences in patient characteristics (i.e., case mix) below. If stratified ONLY, enter "N/A" for questions about the statistical risk model discrimination and calibration statistics.

Validation testing should be conducted in a data set that is separate from the one used to develop the model.

[Response Begins]
The purpose of this model is to account for the need for higher radiation doses to adequately image larger structures and patients. Size-adjustment is intended to eliminate bias that would otherwise result from exogenous variation in the size distribution of patients across accountable entities. Literature review and several rounds of expert panel discussions identified no other relevant confounders at the patient level. This is not a predictive model intended to adjust for patient characteristics in predicting patient outcomes, so traditional metrics of classifier performance (i.e., c statistic, receiver operating characteristic curve, precision-recall curve) are not appropriate.

Accordingly, we validate the adequacy of the risk-adjustment method detailed in 2b.20 by fitting a comparable model:

$$\log(\{D_A\}_{ij}) = \{\beta_0\}_k + \beta_k * d_{ij} + \{z_i\}_k + \varepsilon_{ij}$$

Where all variables above are defined as they were in 2b.20. If the size-adjustment were adequate, we would expect the R-squared of the above model to be close to zero. That is, we expect there to be no relationship between patient size and size-adjusted radiation dose. This R-squared should be close to zero whether the above model is fit using the same data set as the one used to acquire D_A, or using a synthetic data set generated by randomly sampling (with replacement) from the data set used to fit the model. We randomly generated 100 synthetic data sets (of the same size as the Registry) to test the adequacy of our method for acquiring D_A.

Note that, a priori, we do *not* expect the above model to have an R-squared value close to zero (or to remove all differences between observed and expected dose values) when it is fit on a randomly-selected hospital, or on any other population whose practices may not be representative of the general population. This is because some clinicians, clinician groups, or hospitals may systematically overdose some patient size groups (relative to national norms) while dosing other patient size groups in a manner consistent with national norms.

[Response Ends]

2b.27. Provide risk model discrimination statistics.

For example, provide c-statistics or R-squared values.

[Response Begins]

Prior to size-adjustment, the (marginal) R-squared of the models described in 2b.20 differ by CT category, though the magnitude of the association is notable only in Abdomen, Extremities, and Simultaneous Chest and Abdomen and Pelvis exams.

Table 2b-9. Marginal R-Squared by CT category before size-adjustment.

CT category	Marginal R-Squared
Abdomen and Pelvis Low Dose	0.29
Abdomen and Pelvis Routine Dose	0.15
Abdomen and Pelvis High Dose	0.07
Chest Low Dose	0.08

CT category	Marginal R-Squared
Chest Routine Dose	0.10
Cardiac Low Dose	0.06
Cardiac Routine Dose	0.07
Cardiac High Dose or Chest High Dose	0.00
Thoracic or Lumbar Spine	0.05
Simultaneous Thoracic and Lumbar Spine	0.03
Simultaneous Chest and Abdomen and Pelvis	0.18
Head Low Dose	0.03
Head Routine Dose	0.01
Head High Dose	0.00
Neck or Cervical Spine	0.04
Simultaneous Head and Neck Routine Dose	0.01
Simultaneous Head and Neck High Dose	0.00
Extremity	0.22

After size-adjustment, the (marginal) R-squared of the models described in 2b.26 are uniformly close to zero (<0.01). There is negligible variation across the 100 synthetic data sets used to obtain these results, confirming that the risk-adjustment models remove bias due to patient size. The discrimination performance (i.e., c statistic) of these models is not relevant, because their purpose is to remove bias due to a single known confounder, not to maximize prediction of the outcome.

[Response Ends]

2b.28. Provide the statistical risk model calibration statistics (e.g., Hosmer-Lemeshow statistic).

[Response Begins]
The outcome of our size-adjustment is the size-adjusted dose-length product, a continuous variable. The purpose of this model is to account for the need for higher radiation doses to adequately image larger structures and patients. Size-adjustment is intended to eliminate bias that would otherwise result from exogenous variation in the size distribution of patients across accountable entities. Literature review and several rounds of expert panel discussions identified no other relevant confounders at the patient level. Accordingly, following traditional Hosmer-Lemeshow methods, we sorted all CT exams by patient size (as this is the only risk factor in our risk-adjustment models), and estimated observed and size-adjusted doses, as well as the probability of an exam being classified as “out-of-range," across these size deciles. These differences can be interpreted in the same manner as the differences between observed and expected risk levels from a decile plot analysis, but without a global goodness-of-fit statistic.
[Response Ends]

2b.29. Provide the risk decile plots or calibration curves used in calibrating the statistical risk model.

The preferred file format is .png, but most image formats are acceptable.

[Response Begins]
We present the expected dose length product by patient diameter, before and after adjustment. We present results separately for the three largest CT categories, head routine (Table 2b-10), chest routine (Table 2b-11), and abdomen and pelvis routine (table 2b-12.)

Table 2b-10. Dose Length Product by Patient Diameter – Head Routine Dose Exams.

Size Category (Deciles)	Mean Dose Length Product (Unadjusted)	Mean Dose Length Product (Size-Adjusted)	Proportion Out-of-Range (Unadjusted)	Proportion Out-of-Range (Size-Adjusted)
1st	800	879	0.22	0.29
2nd	856	892	0.23	0.27
3rd	873	897	0.25	0.28
4th	887	902	0.26	0.28
5th	905	912	0.28	0.29
6th	923	920	0.30	0.30
7th	943	930	0.33	0.31
8th	966	941	0.36	0.32
9th	1001	960	0.40	0.35
10th	1083	976	0.50	0.39

Table 2b-11. Dose Length Product by Patient Diameter – Chest Routine Dose Exams.

Size Category (Deciles)	Mean Dose Length Product (Unadjusted)	Mean Dose Length Product (Size-Adjusted)	Proportion Out-of-Range (Unadjusted)	Proportion Out-of-Range (Size-Adjusted)
1st	340	638	0.26	0.47
2nd	311	424	0.23	0.38
3rd	338	413	0.28	0.38
4th	369	414	0.33	0.40
5th	402	417	0.39	0.41

Size Category (Deciles)	Mean Dose Length Product (Unadjusted)	Mean Dose Length Product (Size-Adjusted)	Proportion Out-of-Range (Unadjusted)	Proportion Out-of-Range (Size-Adjusted)
6th	444	427	0.46	0.43
7th	491	438	0.54	0.45
8th	550	451	0.64	0.48
9th	640	468	0.74	0.52
10th	863	492	0.85	0.54

Table 2b-12. Dose Length Product by Patient Diameter – Abdomen and Pelvis Routine Dose Exams.

Size Category (Deciles)	Mean Dose Length Product (Unadjusted)	Mean Dose Length Product (Size-Adjusted)	Proportion Out-of-Range (Unadjusted)	Proportion Out-of-Range (Size-Adjusted)
1st	507	993	0.22	0.52
2nd	524	778	0.23	0.45
3rd	580	760	0.28	0.45
4th	646	764	0.35	0.46
5th	721	775	0.43	0.48
6th	811	793	0.53	0.51
7th	917	810	0.65	0.54
8th	1046	822	0.77	0.58
9th	1218	817	0.88	0.60

Size Category (Deciles)	Mean Dose Length Product (Unadjusted)	Mean Dose Length Product (Size-Adjusted)	Proportion Out-of-Range (Unadjusted)	Proportion Out-of-Range (Size-Adjusted)
10th	1551	742	0.95	0.52

[Response Ends]

2b.30. Provide the results of the risk stratification analysis.

[Response Begins]
As described in 2b.27. The observed relationship between patient size and radiation dose differs by CT category, meaning a different risk-adjustment coefficient (different β_k) was required for each CT category. The table provided in 2b.27 show the specific results by CT category. These β_k values are as follows:

Table 2b-13. Risk-adjustment coefficients by CT category.

CT Category	β_k
Abdomen and Pelvis Low Dose	0.009
Abdomen and Pelvis Routine Dose	0.008
Abdomen and Pelvis High Dose	0.006
Chest Low Dose	0.005
Chest Routine Dose	0.009
Cardiac Low Dose	0.006
Cardiac Routine Dose	0.007
Cardiac High Dose or Chest High Dose	0.000
Thoracic or Lumbar Spine	0.003
Simultaneous Thoracic and Lumbar Spine	0.003
Simultaneous Chest and Abdomen	0.007
Head Low Dose	0.011
Head Routine Dose	0.006
Head High Dose	0.000
Neck or Cervical Spine	0.004
Simultaneous Head and Neck Routine Dose	0.000
Simultaneous Head and Neck High Dose	0.000
Extremity	0.008

There are four CT categories (Cardiac High Dose or Chest High Dose, Head High Dose, Simultaneous Head and Neck Routine Dose, Simultaneous Head and Neck High Dose) where the value of β_k was less than 0 at initial fitting of the model in 2b.20. In all four of these categories, no adjustment was performed, but the relationship between patient diameter and non-adjusted dose length product was nonetheless minimal, as shown by the R-squared values in section 2b.27.

As sample sizes in the UCSF Registry are very large, non-zero values of β_k are highly statistically significant, with confidence intervals imperceptibly narrow.

[Response Ends]

2b.31. Provide your interpretation of the results, in terms of demonstrating adequacy of controlling for differences in patient characteristics (i.e., case mix).

In other words, what do the results mean and what are the norms for the test conducted?

[Response Begins]
The table provided in 2b.27 shows that in some CT Categories, the radiation dose is associated with patient diameter, reflecting the clinical practice of using higher radiation doses to penetrate higher-diameter body structures. The fact that the R-squared values in 2b.27 are consistently close to zero after adjustment, and the much weaker relationship between patient diameter and dose length product after adjustment in 2b.29, shows that the adjustment was adequately conducted. Size adjustment does not completely remove the apparent relationship between size and dose in our beta testing data, because the estimated coefficients shown in 2b.30 were derived from a separate registry database that is over 100 times larger than the test data. When these coefficient estimates are applied to any selected set of hospitals, some residual association may be found if some entities overdose certain size groups (relative to national norms) while dosing other patient size groups in a manner consistent with national norms.
[Response Ends]

2b.32. Describe any additional testing conducted to justify the risk adjustment approach used in specifying the measure.

Not required but would provide additional support of adequacy of the risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed.

[Response Begins]
N/A
[Response Ends]

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.

3.01. Check all methods below that are used to generate the data elements needed to compute the measure score.

[Response Begins]
Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)
Coded by someone other than person obtaining original information (e.g., DRG, ICD-10 codes on claims)
[Response Ends]

3.02. Detail to what extent the specified data elements are available electronically in defined fields.

In other words, indicate whether data elements that are needed to compute the performance measure score are in defined, computer-readable fields.
[Response Begins]
ALL data elements are in defined fields in a combination of electronic sources
[Response Ends]

3.03. 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 data elements not from electronic sources.

[Response Begins]
N/A
[Response Ends]

3.05. Complete and attach the NQF Feasibility Score Card.

[Response Begins]
Attached
[Response Ends]

Attachment: Feasibility_scorecards_Facility.xlsx

3.06. 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.

[Response Begins]
There were minimal difficulties surrounding data availability, although issues of missing data are discussed below.

Feasibility scorecards were completed for each EHR system tested: Epic (N=5), Cerner (N=1), Allscripts (N=1). These EHR systems are used across inpatient and hospital outpatient settings. We tested feasibility at the health system level rather than at the hospital level because EHR and other electronic clinical data systems did not differ between hospitals within health systems, and data were collected at the level of the health system, rather than separately for each hospital.

The feasibility scorecard assesses our ability to access the Data Elements in structured fields in electronic clinical data sources (*including both EHR and non-EHR sources*). The results were the same across all EHR systems:

Availability: All primary-access data elements were available and accessible in structured fields in either the EHR or the radiology electronic clinical data systems, including the Radiological Information System (RIS) and the Picture Archiving and Communication System (PACS). The three final data elements – CT category, size-adjusted radiation dose, and global noise – were generated through pre-processing and available for measure score calculation in each system tested.

Accuracy: All data elements have a high likelihood of being correct since they are either entered by a provider into the EHR (typically through text mapping to a code lookup table, or with assistance from a professional coder) for purposes of billing (e.g., ICD-10-CM and CPT® codes, date of birth) or generated by the CT machine itself (RDSR and image pixel data).

Standards: all data elements are structured using nationally accepted vocabularies. Primary-access data elements use code systems ICD-10-CM, CPT®, DICOM, and LOINC. Final data elements are mapped to LOINC codes:

- CT Dose and Image Quality Category: LOINC, 96914-7
- Calculated CT Size-Adjusted Dose: LOINC, 96913-9
- Calculated CT Global Noise: LOINC, 96912-1

Workflow: Once the measure software is implemented, there is no impact on clinician workflow. All data elements are generated during the ordinary course of care or through pre-processing, and no manual abstraction is required.

Missing data

During testing there was some missing data for 8% of exams, and over 90% of the missing data were related to radiation dose. We believe that the issue of missing radiation data is for the most part entirely solvable and within the control of accountable entities. The missing radiation data is not related to an entity's hardware except in very rare situations in which very old machines are used to perform the exam; rather, it is almost entirely a software and data storage issue. The radiation dose data is stored within the Radiation Dose Structured Report (RDSR), a digitized, structured summary of the total radiation output associated with the performance of the CT exam. The RDSR is produced with every CT scan and CMS incentivizes the creation of the RDSR by paying a lower reimbursement for CT scans that do not produce an RDSR. The issue that can arise is that some entities may not *save and store* the RDSR. There is a widespread campaign organized by the American College of Radiology to encourage entities to save and store RDSR information, and the practice is growing. Sites that do not currently save the RDSR in their radiology information systems will need to invest time and resources in modifying their systems to be able to do so. We calculated the amount of time this requires as part of the testing and it was quite modest, as described below and in Table 3-1. Although sites may require vendor support, this work is not excessively burdensome. One of our testing sites went from saving 0% to 96% of their machines' RDSRs in a week's time with remote support from Siemens. Another site with mostly General Electric CT machines increased saving from 10% to 65% within a month, adjusting one machine at a time.

The measure steward will closely monitor missingness at the accountable entity level and report these numbers to the entities, which will be expected to fix the issue within a reasonable period of time. If missingness doesn't resolve to near-zero by the time of NQF Maintenance, we will consider revising the measure to establish a missing data threshold beyond which exams with missing data will be treated as out-of-range (i.e. failed).

Burden and workflow changes (time and cost of data collection)

Interviews were conducted by the UCSF measure development team with representatives from all 7 health systems that served as measure field-testing sites (including site PIs, PACS administrators, and IT and radiology-IT staff). In these interviews, we explored the burden to physicians and staff in terms of hours, cost, complexity, and changes in workflow.

While the implementation imposed no burden on clinicians, it affected staff (mostly IT) workflow. The structured interviews centered around four main topics, and we provide the average and range in time reported for each task across all testing sites in Table 3-1. The reported burden decreased over time as the UCSF team became more adept at troubleshooting and advising the testing sites. All testing sites reported that if the testing were repeated, the hours required would be lower in subsequent rounds. The average cost per hour of the personnel working on the project was estimated by testing sites as $50. Thus, testing was completed at an average cost of $3250 per health system. This level of implementation effort is similar to the burden for other eCQMs, and generally less than the effort involved in participating in national registries.

Table 3-1. Range and average number of hours required, per task group, across all testing sites.

Step	**Range (hours)**	**Average (hours)**
Server/software set up • Building the server (virtual machine) to house the software edge device • Installing the software and troubleshooting	3-40	11.5
Migration of imaging exams to server • Directing the PACS to send CT exam data to the software • Monitoring the data transfer	1-20	6.3
Extracting diagnostic (ICD10) and procedure (CPT) codes and sending to software • Identifying data sources and building queries • Running queries and performing quality control	1-25	10.1
Saving the Radiation Dose Structured Report (RDSR) in PACS • The RDSRs are universally created by the CT machines • This data element is not universally saved nor stored • The process of saving the RDSR varies by manufacturer and needed to be implemented across all scanners within each network	1-50	37.5
Total (Reflects total reported hours at each testing site)	**8-65**	**65.4**

[Response Ends]

Consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

3.07. Detail 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),

Attach the fee schedule here, if applicable.

[Response Begins]
There are no fees for users submitting their eCQM data to CMS programs.

As described in sp.22, the measure requires access to and processing of primary data elements from the EHR and radiology electronic clinical data systems into variables that can be ingested by the eCQM for measure score calculation. The steward's software to ingest this data and calculate the measure is freely available, with a license agreement described below that prevents reselling by other companies. The specifications of the measure (e.g., code lists, risk model coefficients, radiation dose and noise thresholds, and required algorithms) are in the public domain. Should they choose, other vendors may also develop their own software to implement the measure specifications using the information included in this submission.

Consistent with other eCQMs, this measure can be reproduced and distributed, without modification, for noncommercial purposes (e.g., use by healthcare providers in connection with their practices). Commercial use is defined as the sale, licensing, or distribution of the measure for commercial gain, or incorporation of the measure into a product or service that is sold, licensed, or distributed for commercial gain. All commercial uses or requests for modification must be approved by Alara Imaging, Inc. and are subject to a license at the discretion of Alara Imaging, Inc.

Limited proprietary coding is contained in the measure specifications for convenience. Users of the proprietary code sets should obtain all necessary licenses from the owners of these code sets. Alara Imaging, Inc. disclaims all liability for use or accuracy of any third-party code contained in the specifications. CPT(R) contained in the measure specifications is copyright 2004-2021 American Medical Association. LOINC(R) is copyright 2004-2021 Regenstrief Institute, Inc. Due to technical limitations, registered trademarks are indicated by (R) or [R].
[Response Ends]

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.

Extent to which intended audiences (e.g., consumers, purchasers, providers, policy makers) can understand the results of the measure and are likely to find them useful for decision making.

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 demonstrating performance improvement.

4a.01.

Check all current uses. For each current use checked, please provide:

Name of program and sponsor

URL

Purpose

Geographic area and number and percentage of accountable entities and patients included

Level of measurement and setting

[Response Begins]
Not in use
This is a new measure submitted for initial endorsement. It is not currently in use in any program.
[Response Ends]

4a.02. Check all planned uses.

[Response Begins]
Payment Program
Quality Improvement with Benchmarking (external benchmarking to multiple organizations)
Quality Improvement (internal to the specific organization)
[Response Ends]

4a.03. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing), explain why the measure is not in use.

For example, do policies or actions of the developer/steward or accountable entities restrict access to performance results or block implementation?

[Response Begins]
N/A – this is a new measure
[Response Ends]

4a.04. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes: used in any accountability application within 3 years, and publicly reported within 6 years of initial endorsement.

A 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.

[Response Begins]
This measure is intended for use in three CMS programs:
1. Hospital Inpatient Quality Reporting Program (IQR), which seeks to drive quality improvement through measurement and transparency by publicly displaying data to help consumers make more informed decisions about their health care.
2. Hospital Outpatient Quality Reporting Program (OQR), which is a pay for quality data reporting program implemented by CMS for outpatient hospital services. Quality outcomes are publicly available.
3. Medicare and Medicaid Promoting Interoperability Program for Eligible Hospitals and Critical Access Hospitals (CAHs) (PI), which seeks to advance certified electronic health record technology (CEHRT) utilization, further reduce burden, and increase interoperability and patient access to their health information. Eligible hospitals and critical access hospitals earn points by reporting on measures in each of these areas.

This measure would apply to all program-eligible hospitals that perform diagnostic CT in inpatient and hospital outpatient (including emergency department) settings. Measurement is at the hospital level.
We will submit this measure to the CMS MUC List in 2022, for consideration in the Hospital programs. If adopted by CMS, the first assessment period would be January 1, 2024 through December 31, 2025.

CMS publicly reports outcomes of its IQR and OQR programs on its *Care Compare* website, which reports information on more than 100 quality measures for over 4,000 hospitals nationwide and allows consumers to compare hospital performance across many conditions. As media coverage of radiation overuse has proven this to be an important safety issue to patients and the public, the UCSF measure development team believes there is strong interest and benefit in public reporting of this measure.

[Response Ends]

4a.05. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

Detail 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.

[Response Begins]
This measure is designed to not only monitor performance but also provide feedback to achieve a meaningful reduction in radiation doses. Though the measure score itself only reflects an aggregated out-of-range rate across all CT categories, the edge device software (described in sp.22) generates stratified feedback to users allowing them to make decisions to improve their performance. The feedback highlights CT categories of poor performance so that sites can see exactly where they need to take corrective action to improve their radiation doses or image quality. While the measure is reported at the accountable entity level, the feedback can be provided at multiple levels, such as the individual clinician, clinician group, facility, imaging center, or hospital level, making the feedback exceedingly actionable.

Also, the feedback will evolve over time in response to user demand. For example, some of our testing sites have asked for optimized protocols to help them achieve in-range radiation dose targets; thus, this is under development.

Alara Imaging, Inc. – our partner in software development – is working with our testing sites to pilot this educational feedback that will be provided during implementation. The testing sites are receiving this information free of charge.
[Response Ends]

4a.06. Describe the process for providing measure results, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

[Response Begins]
Sites that provided data for measure testing were convened by video conference call to review their performance on the measure. Sites were able to view their CT exams by CT category and compare (1) their allocations of exams across CT categories relative to the UCSF Registry, (2) a pass ("in-range") rate for exams across each CT category, and (3) a weighted score that combines the frequency and pass rate to assess the CT categories that need the most attention for overall measure score improvement. Sites are also receiving detailed feedback *by CT protocol* in terms of the technical parameters they used in comparison with sites that have the lowest doses/lowest measure score. This provides highly actionable information to modify practice.
[Response Ends]

4a.07. Summarize the feedback on measure performance and implementation from the measured entities and others. Describe how feedback was obtained.

[Response Begins]
Verbal feedback was provided by site participants on the video calls. More detail on this feedback is provided in 4a.08 below.
[Response Ends]

4a.08. Summarize the feedback obtained from those being measured.

[Response Begins]
Feedback from sites often reflected a recognition and understanding for why radiation doses were particularly high. For example, one site that failed a number of exams in the Head Routine Dose category routinely uses three phase scans for this type of scan, an approach that deviates from industry norms and leads to unnecessarily high doses.

Some sites had generally high radiation doses across a number of categories, while others struggled with only one or two high-volume categories. For the sites that had targeted issues, there was an interest in not only ascertaining which imaging protocols were leading to failure in the measure, but also a desire for guidance on alternative protocols to administer in order to optimize dose while maintaining adequate image quality.
[Response Ends]

4a.09. Summarize the feedback obtained from other users.

[Response Begins]
There has been general interest from sites that were not included as testing partners to obtain the type of feedback provided to testing sites. It is often the case that sites are unsure how their doses and image quality compare to peers and there is demand for solutions that can help provide this guidance and tailored feedback in a structured way.
[Response Ends]

4a.10. Describe how the feedback described has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

[Response Begins]
UCSF, as the measure developer, has been working with 26 health care organizations and 161 imaging facilities for 10 years on ways to assess radiation dose and provide feedback to help organizations improve quality and safety of CT imaging. This work has included a randomized controlled trial of different approaches to audit feedback and education (described at length in 4b.01). The feedback we've received from both Registry and field-testing sites in the form of surveys, interviews, webinars, forums for sharing best-practices, and informal conversations have influenced the development and the specification of the measure. For example, the CT categories were revised several times based on feedback from imaging facilities. The measure was defined to include a 100% sample of CT exams so as not to have selected exams submitted. The approach of providing feedback on the measure score – e.g. to provide feedback at the

level of specific machine and on individual patients whose doses exceed thresholds – all came from input from our testing partners. While measure will be scored and reported at an aggregated level, the feedback was requested to be far more nuanced to make it actionable.
[Response Ends]

4b.01. You may refer to data provided in Importance to Measure and Report: Gap in Care/Disparities, 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, provide an explanation. 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.

[Response Begins]
This is a new measure not previously in use. Thus, while empirical performance improvement data are not yet available, previous research suggests educational feedback of the kind delivered through measure implementation, described at length in 4a.05, can help reduce excessive radiation doses in CT while preserving diagnostic utility. In a randomized controlled trial involving roughly 1 million CT exams from 100 imaging facilities across 6 countries, Smith-Bindman et al. studied the impact of multicomponent educational feedback on radiation doses used in CT imaging. (Smith-Bindman 2020) This included audit feedback with targeted suggestions, participation in a quality improvement collaborative, and best-practice sharing. Together, these interventions achieved 23-58% reductions in the proportion of high-dose exams, based on organ dose, with no observed change in image quality. Audit feedback alone, comparing radiation doses with those of other facilities, also reduced the proportion of high-dose exams and mean doses, but with a smaller magnitude.

Prior to this randomized trial, smaller, single-center, and/or observational studies reached the same conclusion that educational feedback such as audits reduces radiation doses. The Luxemburg Ministry of Health implemented an audit of radiation doses in its CT imaging departments and observed reductions in the 75th percentile of dose of 18-75%, for all body regions, which were sustainable over time. (Tack 2014). A small, controlled pilot examining the effect of personalized dose audit reports and education directed at radiology technologists within a US health system similarly lowered patients' radiation exposure in CT imaging. (Miglioretti 2014). Another interventional study across the University of California system deployed radiation dose audits and best practice sharing, resulting in considerable dose reductions: a 19% and 25% decrease in mean effective dose for chest and abdomen exams, respectively, and a reduction in the number of exams exceeding allowable benchmarks by 48% and 54% for chest and abdomen, respectively. (Demb 2017).

References

1. Demb J, Chu P, Nelson T, et al. Optimizing Radiation Doses for Computed Tomography Across Institutions: Dose Auditing and Best Practices. JAMA Intern Med. 2017;177(6):810-817.
2. Miglioretti DL, Zhang Y, Johnson E, et al. Personalized technologist dose audit feedback for reducing patient radiation exposure from CT. J Am Coll Radiol. 2014;11(3):300-308.
3. Smith-Bindman R, Chu P, Wang Y, et al. Comparison of the Effectiveness of Single-Component and Multicomponent Interventions for Reducing Radiation Doses in Patients Undergoing Computed Tomography: A Randomized Clinical Trial. JAMA Intern Med. 2020 May 1;180(5):666-675.
4. Tack D, Jahnen A, Kohler S, et al. Multidetector CT radiation dose optimisation in adults: short- and long-term effects of a clinical audit. Eur Radiol. 2014;24(1):169-175.

[Response Ends]

4b.02. Explain any unexpected findings (positive or negative) during implementation of this measure, including unintended impacts on patients.

[Response Begins]
Field testing involved retrospective data collection to capture baseline performance at testing facilities. Since no intervention took place, there were no unintended impacts on patients.

We learned early on in field-testing that the Radiation Dose Structured Report (RDSR) was initially unavailable for many CT exams at all testing sites. This issue is described at length in section 3.06. The RDSR is a digitized, structured summary, automatically generated by the CT machine, providing the total radiation output for each CT exam. Though federal law requires CT machines *generate* the RDSR, there is no mandate that facilities *save* the report, and most of our testing sites were unaware the report was not saved. We worked with our sites to modify their systems to save the RDSR, ultimately capturing 94% of dose reports. Nationwide, awareness of this issue is growing, and more facilities are saving the RDSR. Regulatory solutions should be considered upon measure implementation to ensure this trend continues.

Given the relationship of radiation dose and image noise, there is concern that dose reduction will result in deteriorated image quality. Theoretically, this reduces the diagnostic utility of CT images and could harm patients by requiring repeated scanning (thus doubling the dose). However, we did not see this play out in our testing data. Out-of-range measure scores due to inadequate image quality (i.e. excessive global noise) were exceedingly rare, with less than 1% of exams, on average, across all reporting entities. This was to some degree expected, given the earlier Image Quality Study, in which radiologists graded 3% and 8% of exams as “poor” or “marginally acceptable” image quality, respectively (this is described at length in the Validity Testing section 2b.02). This finding supports a considerable opportunity to reduce radiation doses without impacting quality. Since field-testing captured only about four weeks’ worth of CT data, we did not observe trends in image quality. The measure steward will monitor out-of-range rates annually to determine if image quality is worsening due to declining radiation doses and determine if thresholds should be adjusted or if a subsequent Image Quality Study of radiologist satisfaction should be repeated.
[Response Ends]

4b.03. Explain any unexpected benefits realized from implementation of this measure.

[Response Begins]
Testing this measure prompted many sites to learn of the problem of Radiation Dose Structured Reports not being saved in their PACS systems and to implement corrective changes. Beyond that, it is too early to identify other unexpected benefits.
[Response Ends]

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.

If you are updating a maintenance measure submission for the first time in MIMS, please note that the previous related and competing data appearing in question 5.03 may need to be entered in to 5.01 and 5.02, if the measures are NQF endorsed. Please review and update questions 5.01, 5.02, and 5.03 accordingly.

5.01. Search and select all NQF-endorsed related measures (conceptually, either same measure focus or target population).

(Can search and select measures.)
[Response Begins]
3621: 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
2820: Pediatric Computed Tomography (CT) Radiation Dose
[Response Ends]

5.02. Search and select all NQF-endorsed competing measures (conceptually, the measures have both the same measure focus or target population).

(Can search and select measures.)
[Response Begins]
[Response Ends]

5.03. If there are related or competing measures to this measure, but they are not NQF-endorsed, please indicate the measure title and steward.

[Response Begins]
Two existing process measures in the CMS Merit-based Incentive Payment System (MIPS) program are related (not competing) in that they address patient safety related to radiation exposure in CT imaging:

1. Optimizing Patient Exposure to Ionizing Radiation: Count of Potential High Dose Radiation Imaging Studies: Computed Tomography (CT) and Cardiac Nuclear Medicine Studies (CMIT # 2286, steward: American College of Radiology)
2. Radiation Consideration for Adult CT: Utilization of Dose Lowering Techniques (CMIT # 2570, stewards: American College of Radiology, American Medical Association-Physician Consortium for Performance Improvement, National Committee for Quality Assurance)

There are three process measures related to CT in the CMS Hospital Outpatient Quality Reporting Program, but none directly addresses radiation dose:

1. Head CT or MRI Scan Results for Acute Ischemic Stroke or Hemorrhagic Stroke who Received Head CT or MRI Scan Interpretation Within 45 Minutes of ED Arrival (CMIT # 918, steward: Centers for Medicare & Medicaid Services)
2. Cardiac Imaging for Preoperative Risk Assessment for Non-Cardiac Low-Risk Surgery (CMIT # 1367, steward: Centers for Medicare & Medicaid Services)
3. Abdomen Computed Tomography (CT) Use of Contrast Material (CMIT # 2599, steward: Centers for Medicare & Medicaid Services)

[Response Ends]

5.04. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s), indicate whether the measure specifications are harmonized to the extent possible.

[Response Begins]
Yes
[Response Ends]

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

[Response Begins]
Measure 2820 was developed by the same UCSF measure development group as the current proposed measure. It calls for imaging facilities to assess their radiation doses in children against published benchmarks, and it provides a framework to improve doses exceeding benchmarks. In contrast, the proposed new measure is specified in adults. Measure 2820 was a first-generation pediatric measure, and the new measure is a second-generation adult measure that incorporates stratification by clinical indication, adjustment by patient size, and image quality. The UCSF team plans to update measure 2820 in a subsequent review cycle to include stratification for clinical indication and an assessment of image quality and will reflect harmonization with the newly proposed measure.

Measure 3621, developed by the American College of Radiology, is also focused on reducing radiation doses for CT, but the outcomes and target populations are different. The denominator of measure 3621 includes CT exams in all patients who have undergone three specific types of CT scans: single phase CT abdomen-pelvis exams with contrast, single phase CT chest exams without contrast, and single phase CT head/brain exams. This means patients who may have undergone *multi-phase* abdomen, chest and head scans are not included. In contrast, the proposed new measure's denominator is nearly all diagnostic CT exams in adults. Thus, the proposed measure inherently considers the clinician's subjective choice of imaging protocol (e.g. whether to assign a patient to a single or multi-phase abdomen exam), which is the single most important predictor of radiation dose. Measure 3621 does not account for this high impact decision, assessing dose only after the selection of a single phase exam is made. This difference impacts the meaningfulness of the measures. Measure 3621 stratifies by protocol, in essence comparing single phase CT abdomen-pelvis exams with contrast to other single phase CT abdomen-pelvis exams with contrast, regardless of the reason for scanning. Assessing doses in this way, without considering the underlying indication, ignores the variation stemming from protocol selection and fails to identify patients who require a particular protocol, such as single phase abdomen, but who instead received much higher doses through unnecessary multi-phase exams. Most high radiation doses are a result of using multi-phase protocols, and yet these exams are not included in Measure 3621.

In effect, the denominator of measure 3621 is not stable; in some practices this might represent a large portion of patients who underwent CT, whereas in others it might be very few. In the UCSF International CT Dose Registry, which includes over 6.5 million CT scans from 161 hospitals and imaging 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 measure 3621 does not reflect a patient population who *require* these exams, but rather reflects the variable decisions of radiologists to assign patients to different imaging protocols. This is not a hypothetical problem but one that would be expected to occur frequently and miss the most egregious radiation overdosing. A physician group that uses multiphase scanning for most of their CT exams will deliver inappropriately high doses to many patients, but this will not be assessed, flagged, or failed by measure 3621.

An important difference between the measures is that the proposed measure assesses radiation dose *according to thresholds determined by the underlying clinical indication for imaging*, while Measure 3621 uses the average observed dose in the ACR registry for these protocols, without consideration if the doses are appropriate for the underlying indication. Radiation doses should be assessed based on the intent and clinical question of the provider ordering the scan, not on the radiologist's choice of protocol. Nonetheless, Measure 3621 can contribute to dose optimization and potentially encourage physicians to lower radiation doses for single-phase exams.

A final advantage of the proposed measure is that it includes assessment of image quality as a means of protecting the diagnostic value of CT imaging from unintended consequences of excessive dose reduction.

We believe the data collection burden would be nominal if sites choose to report on both measures. In terms of harmonization, both measures utilize data generated during the standard course of clinical care, either by clinicians or CT machines; no human abstraction is required. Both measures use the same radiation dose metric (dose length product)

and use effective diameter as a metric of patient size. In the future, the ACR may require the RDSR, and when they do, the measures will be harmonized on this data source. However, complete harmonization is not possible due to the fundamentally different approaches; for example, the proposed measure uses diagnosis (ICD-10-CM) and procedure (CPT®) codes associated with the exam to assign the CT category, while measure 3621 determines exam type using DICOM data from the CT exam including study description and body region. As an eCQM, our measure is designed to minimize the burden of data collection. As described in section 3.06, the bulk of the cost and effort is in set-up, but minimal effort for staff (no effort for clinicians) is required on an ongoing basis.
[Response Ends]

5.06. Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality). Alternatively, justify endorsing an additional measure.

Provide analyses when possible.

[Response Begins]
There are no competing measures

Measure 3621 is related. As described in 5.05, the proposed measure is different than, and improved upon Measure 3621 in the following ways:

(1) It assesses radiation doses by clinical indication, thereby allowing consideration for the *reason* for imaging.

(2) Similarly, it assesses radiation dose *according to thresholds determined by the underlying clinical indication for imaging,* rather than to observed doses without consideration if the doses are appropriate for the underlying indication.

(3) The proposed measure's denominator includes nearly all diagnostic CT exams in adults. Thus, the proposed measure inherently considers the clinician's subjective choice of imaging protocol (e.g. whether to assign a patient to a single or multi-phase abdomen exam), which is the single most important predictor of radiation dose.

(4) Includes assessment of image quality as a means of protecting the diagnostic value of CT imaging from unintended consequences of excessive radiation dose reduction.
[Response Ends]

Appendix

Supplemental materials may be provided in an appendix.: Available in attached file
Attachment: 1074QDM_Bonnie_screenshot.jpg
Attachment: 1074QDM_Bonnie_test_cases.xlsx
Attachment: FHIR_testing_synthetic_patients.png
Attachment: FHIR_testing_eCQM_code_output.png

Contact Information

Measure Steward (Intellectual Property Owner) : Alara Imaging
Measure Steward Point of Contact: Mazonson, Nathan, nate@alaracare.com
Measure Developer if different from Measure Steward: University of California, San Francisco
Measure Developer Point(s) of Contact: Smith-Bindman, Rebecca, rebecca.smith-bindman@radiology.ucsf.edu
Smith-Bindman, Rebecca, Rebecca.smith-bindman@ucsf.edu
Stewart, Carly, carly.stewart@ucsf.edu

Additional Information

1. Provide any supplemental materials, if needed, as an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be collated one file with a table of contents or bookmarks. If material pertains to a specific criterion, that should be indicated.

[Response Begins]
Available in attached file
[Response Ends]

Attachment: 1074QDM_Bonnie_screenshot.jpg
Attachment: 1074QDM_Bonnie_test_cases.xlsx
Attachment: FHIR_testing_synthetic_patients.png
Attachment: FHIR_testing_eCQM_code_output.png

2. List the workgroup/panel members' names and organizations.

Describe the members' role in measure development.

[Response Begins]
Project leadership:

Rebecca Smith-Bindman, MD, Principal Investigator (University of California San Francisco). Dr. Smith-Bindman has had overall responsibility for leading the project, from measure conceptualization through development, implementation, and testing. She supervised all project staff and led the development of the reporting software, the risk adjustment strategy, the measurement of image quality, and alpha and beta testing. Lastly, she directed the Technical Expert Panel and ensured integration of their feedback into the measure.

Marc Kohli, MD, Co-Investigator (University of California San Francisco). Dr. Kohli contributed his expertise in medical informatics, clinical workflow within Radiology and EHR, standards in imaging, and knowledge of data extraction from electronic radiology data to measure development, specifications, testing, and implementation.

Patrick Romano, MD, MPH, Co-Investigator (University of California Davis). Dr. Romano oversaw UC Davis' participation in the project, with a specific focus on supporting the development, testing, refinement, and validation of detailed technical specifications for the proposed measures. He also advised and supported the UCSF team through submissions to the CMS Measure Under Consideration List and National Quality Forum.

Andrew Bindman, MD, Advisor (Kaiser Foundation Health Plan). Dr. Bindman was formerly a Co-Principal Investigator with the University of California San Francisco. He initially shared overall responsibility for the project with Dr. Smith-Bindman, specifically contributing to developing measure concepts, specifications, and the risk adjustment strategy. Following his move to Kaiser in the fall of 2020, he stayed on the project in an advisory capacity.

Technical Expert Panel members include:

- Niall Brennan, MPP, CEO, Health Care Cost Institute
- Jesse Chusid, MD, MBA, Chair, Department of Radiology, Senior Vice President, Imaging Services, Northwell Health
- Melissa Danforth, Vice President of Health Care Ratings, The Leapfrog Group
- Tricia Elliot, MBA, CPHQ, Director, Quality Measurement, Joint Commission
- Mohammad Fakih, MD, MPH, Chief Quality Officer Clinical & Network Services, Ascension Healthcare
- Jeph Herrin, PhD, Adjunct Assistant Professor, Yale University
- Jay Leonard Lichtenfeld, MD, MACP, Independent Consultant, Formerly Deputy Chief Medical Officer American Cancer Society, Inc.
- Leelakrishna Nallamshetty, MD, Associate Chief Medical Officer, Radiology Partners

- Matthew Nielsen, MD, MS, Professor and Chair of Urology, UNC Gillings School of Global Public Health
- Debra Ritzwoller, PhD, Patient Advocate and Health Economist (Patient Representative)
- Lewis Sandy, MD, Executive Vice President, Clinical Advancement, UnitedHealth Group
- Mary Suzanne Schrandt, JD, Patient Advocate (Patient Representative)
- James Anthony Seibert, PhD, Professor, University of California, Davis
- Robert H. Sherrier, MD, Chief Consultant, Diagnostic Services, Veterans Affairs
- Arjun Venkatesh, MD, MBA, MHS, Associate Professor, Emergency Medicine, Yale School of Medicine

[Response Ends]

3. Indicate the year the measure was first released.

[Response Begins]
N/A - this is a new measure
[Response Ends]

4. Indicate the month and year of the most recent revision.

[Response Begins]
N/A - this is a new measure
[Response Ends]

5. Indicate the frequency of review, or an update schedule, for this measure.

[Response Begins]
The measure steward will review measure specifications annually to ensure they remain appropriate to the measure's concept or logic. In particular, the steward will monitor performance annually to determine if the specified radiation dose and image quality thresholds remain appropriate. For example, if radiation doses overall are reduced, the steward will assess if the radiation dose thresholds should change accordingly. Or if dose reduction leads to a concern about image quality, the steward will determine if another Image Quality Study assessing physician satisfaction with CT images is needed.

The steward will also continue to update the algorithm for CT category assignment as diagnosis and procedure codes are created or retired.
[Response Ends]

6. Indicate the next scheduled update or review of this measure.

[Response Begins]
N/A – this is a new measure
[Response Ends]

7. Provide a copyright statement, if applicable. Otherwise, indicate "N/A".

[Response Begins]
Copyright (C) 2021 Alara Imaging, Inc. All Rights Reserved.

Alara Imaging, Inc. is not responsible for any use of the Measure. Alara Imaging, Inc. makes no representations, warranties, or endorsement about the quality of any organization or physician that uses or reports performance measures and Alara Imaging, Inc. has no liability to anyone who relies on such measures or specifications.

The Measure can be reproduced and distributed, without modification, for noncommercial purposes (e.g., use by

healthcare providers in connection with their practices). Commercial use is defined as the sale, licensing, or distribution of the Measure for commercial gain, or incorporation of the Measure into a product or service that is sold, licensed or distributed for commercial gain. All commercial uses or requests for modification must be approved by Alara Imaging, Inc. and are subject to a license at the discretion of Alara Imaging, Inc.
[Response Ends]

8. State any disclaimers, if applicable. Otherwise, indicate “N/A”.

[Response Begins]
The Measure is not a clinical guideline, does not establish a standard of medical care, and has not been tested for all potential applications.

Alara Imaging, Inc, the University of California San Francisco, and its members and users shall not be responsible for any use or accuracy of the Measure or any code contained within the Measure. THE MEASURE AND SPECIFICATIONS ARE PROVIDED “AS IS” WITHOUT WARRANTY OF ANY KIND.

Limited proprietary coding is contained in the Measure specifications for convenience. Users of the proprietary code sets should obtain all necessary licenses from the owners of these code sets. Alara Imaging, Inc. disclaims all liability for use or accuracy of any third-party code contained in the specifications. CPT® contained in the Measure specifications is copyright 2004-2021 American Medical Association. LOINC® is copyright 2004-2021 Regenstrief Institute, Inc.
[Response Ends]

9. Provide any additional information or comments, if applicable. Otherwise, indicate “N/A”.

[Response Begins]
N/A
[Response Ends]