**National Quality Forum—Measure Testing (subcriteria 2a2, 2b2-2b7)**

**Measure Number** (*if previously endorsed*)**:** 0739

**Measure Title**: Radiation Dose of Computed Tomography (CT)

**Date of Submission**: 2/10/2014

**Type of Measure:**

|  |  |
| --- | --- |
| Composite – ***STOP – use composite testing form*** | Outcome (*including PRO-PM*) |
| Cost/resource | Process |
| Efficiency | Structure |

|  |
| --- |
| **Instructions**   * 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. * **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.** * **For outcome and resource use measures**, section **2b4** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed. * Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). |

|  |
| --- |
| **Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.**  **2a2.** **Reliability testing** [**10**](#Note10) 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b2.** **Validity testing** [**11**](#Note11) 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [**12**](#Note12)  **AND**  If patient preference (e.g., informed decisionmaking) 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). [**13**](#Note13)  **2b4.** **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 that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b5.** 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**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b6.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b7.** For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data),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 nonresponders) and how the specified handling of missing data minimizes bias.  **Notes**  **10.** 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).  **11.** 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.  **12.** 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.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.  **16.** 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. |

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

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

**1.1. What type of data was used for testing**? (*Check all the sources of data identified in the measure specifications and data used for testing the measure*. *Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*must be consistent with data sources entered in S.23*)** | **Measure Tested with Data From:** |
| abstracted from paper record | abstracted from paper record |
| administrative claims | administrative claims |
| clinical database/registry | clinical database/registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| other: Click here to describe | other: Click here to describe |

**1.2. If an existing dataset was used, identify the specific dataset** (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Several datasets have been used for testing of the measure including data from Individual institutions, collaborations of institutions, integrated health care systems, electronic medical records, data extracted from stored CT images (captured from the CT console at the time of scanning or harvested from the PACS (Picture Archiving Communication System - the computerized systems for reviewing and storing imaging data), printed CT images, or information stored in the medical record*. Smith-Bindman (JAMA Internal Medicine 2009, JAMA 2012), Miglioretti (JAMA Pediatrics 2013; JACR 2014) and Keegan (JACR 2014)* use various methods of data abstraction. Two manuscripts abstracting and summarizing dose using the NQF endorsed metric are included.

**1.3. What are the dates of the data used in testing**? January 1, 2008 – December 31 , 2013

**1.4. What levels of analysis** **were tested**? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

|  |  |
| --- | --- |
| **Measure Specified to Measure Performance of:**  **(*must be consistent with levels entered in item S.26*)** | **Measure Tested at Level of:** |
| individual clinician | individual clinician |
| group/practice | group/practice |
| hospital/facility/agency | hospital/facility/agency |
| health plan | health plan |
| other: facility | other: facility |

**1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

The measure has been tested in several settings: Group Health Research Institute, a large integrated Health System in the Pacific Northwest. CT examinations on over 10,000 examinations have been assembled and included in several publications *(Miglioretti, JACR 2014; Miglioretti JAMA Pediatrics 2013)*

The measure was tested in a consortium of integrated health care systems (n=6) and data were assembled for over 5000 CT examinations, and were published *(Smith-Bindman JAMA 2012)*

The measure was tested across the five University of California Medical Center, including over 100,000 CT examinations. The data has in part been published *(Keegan, JACR 2014)* and additional manuscripts were presented at national meetings (RSNA 2012) and are in preparation.

For all, analyses were done using consecutive sample of CT examinations within anatomic area, age and machine type strata as specified in this measure, or using a randomly selected subset of examinations and analyzed per measure specifications.

A quality improvement activity assembling data per the NQF specifications was approved by the Board of the American College of Radiology for PQRS credit.

*Miglioretti D, Johnson E, Vanneman N, Smith-Bindman R, al e. Use of Computed Tomography and Associated Radiation Exposure and Leukemia Risk in Children and Young Adults across Seven Integrated Healthcare Systems from 1994 – 2010. JAMA Pediatrics Published online June 10, 2013 joli:101001/jamapediatrics2013311 2013.*

*Miglioretti, YX Zhang, E Johnson, N Vanneman, R Smith-Bindman. Personalized Technologist Dose Audit Feedback for Reducing Patient Radiation Exposure from Computed Tomography. In press Journal of the American College of Radiology 2014.*

*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;169:2078-86.*

*Smith-Bindman R. Is computed tomography safe? N Engl J Med 2010;363:1-4.*

*Smith-Bindman R. Environmental causes of breast cancer and radiation from medical imaging: findings from the Institute of Medicine report. Arch Intern Med 2012;172:1023-7.*

*Smith-Bindman R, Miglioretti DL, Johnson E, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010. JAMA 2012;307:2400-9.*

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

The summary of CT dose has been done in children and adults, and using consecutive scans without exclusion (ie scans were not excluded on any individuals) and analyzed within strata. Because the measure is specified at the institutional level, there is no reason to exclude any individuals.

The strata for this measure include:

* Anatomic area strata: head, chest, abdomen/pelvis. These anatomic areas reflect approximately 85% of CT examination types in adults, and approximately 75% of CT examination types in children
* Age strata: infant (<1); small child (1-5); medium child (>5 - 10); large child (>10-15) and adult (>15)
* CT machine (manufacturer, type)

Part A: For adult categories, 100 patients within each strata provides minimum sample size, although larger sample size is acceptable

For child categories, 50 patients within each strata is more feasible and will provide adequate sample size. (One year of data should be extracted if the minimum cannot be met within a shorter time interval)

Part B: This measure includes all reports (no strata) and a minimum of 100 reports is adequate

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

There are no differences in the data or sample used for different aspects of testing.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2a2. RELIABILITY TESTING**

***Note****: If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter “see section 2b2 for validity testing of data elements”; and skip 2a2.3 and 2a2.4.*

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)  
 **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)  
 **Performance measure score** (e.g., *signal-to-noise analysis*)  
  
**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps―do not just name a method; what type of error does it test; what statistical analysis was used*)

The Proposed CT Dose measure calls for the collection of several metrics reflecting CT dose indices including DLP, CTDIvol, SSDE and Effective Dose. CTDI and DLP are calculated by all current CT scanners, and Effective dose reflects a combination of the dose measure the CT machine generates and a measure of how harmful that dose may be to the patient based on the site of the body that is radiated and the age of the patient. SSDE is a variation in the CTDIvol metric that takes into account patient size.

Two of the metrics (CTDIvol and DLP) are automatically generated by the CT manufacturers, and stored with CT machines, and thus the reliability tests were done to confirm that these data can be reliably accessed. Two of the metrics (SSDE and ED) must be calculated and thus the reliability tests were done to determine if there is consistency in how these are calculated.

CTDIvol and DLP measures have been widely used for over a decade in several other countries, are called for in a bill that is in effect in California and I have personal and recent experience collecting these dose Indices across 12 large institutions reflecting dozens of machines and thousands of patients Reliability of CT radiation dose metric abstraction (DLP and CTDIvol) was tested by our group in several ways. First, manual data abstraction of data recorded from the PACS system was repeated in two large samples (one at Group Health, and one at UCSF) where the data was abstracted by a single observer, yielding highly reliable measures between abstractions (i.e. the measures were concordant, nearly perfect Kappa statistics) Second, data were extracted via commercial software product using two different tools for extracting the data from the stored CT files in PACS, and these were reviewed by a medical physicist to ensure the data were correct. This was performed at five separate institutions, and found the electronically captured data was identical to the manual review, perfect Kappa statistics.

Reliability of the ED measurement was performed by calculating this measurement using several different statistical approaches (that different individuals and programs might use). This was conducted on a sample of 5000 CT examinations, conducted in individuals of different age and size and sex. The different approaches yielded highly consistent results; the kappa statistics were largely over 95%, and the only strata where there were disagreements between the various approaches, were for the calculation of effective dose in newborns, where depending on the technique used to estimate this metric, the results were divergent. However, few CT examinations are conducted in infants, and the other three metrics are valid in this group.

SSDE is a relatively new metric that tries to take into account patient size. In our published work *(Keegan JACR 2014)* it tracks in parallel to the other metrics, but has not undergone formal reliability testing in large cohorts of patients.

**2a2.3. For each level of testing checked above, what were the statistical results from reliability testing**? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Highly reliable, Kappas > 95%

**2a2.4 What is your interpretation of the results in terms of demonstrating reliability**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

Highly reliable, Kappas > 95%

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b2. VALIDITY TESTING**

**2b2.1. What level of validity testing was conducted**? (*may be one or both levels*)  
 **Critical data elements** (*data element validity must address ALL critical data elements*)

**Performance measure score**

**Empirical validity testing** **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

**2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

The four dose parameters (metrics) included in this measure reflects slightly different aspects of dose, and each was included because it provides a unique reflection of dose and can be used to improve quality and safety.

These dose parameters specified in this measure primarily reflect the dose that comes out of the machine and the dose that the patient is exposed to and *dictate the absorbed organ doses to the patient*. Absorbed doses (these are the doses a patient actually receives) will vary by sex and weight, but are *primarily* determined by the doses that come out of the machine. These dose metrics are highly correlated with the doses patients receive; higher DLPs, CTDIs, SSDE and Effective doses are associated with higher absorbed dose to the patient’s organs and higher patient detriment (harm). If these doses were lowered patients would be exposed to lower doses of radiation, have correspondingly lower absorbed organ doses and would be expected to have less detriment from these exposures to radiation. While patient absorbed doses are important, they are difficult to quantify.

However, the dose parameters themselves are vitally important as they 1) closely reflect organ doses and 2) are precisely those measurements that the technologist and physician can influence to lower doses. That is why these measures were chosen for this metric. Estimating absorbed organ doses might be a more precise way to compare doses between two examinations on two patients, however, this is simply not practical. It is much more complicated to estimate these parameters, there are over 30 different organs where these doses can be compared and it does not make sense to measure because the technologist cannot directly influence these measures, and there would be practical way to compare facilities as there ware so many organ doses to compare. Using organ dose might add a very small amount more precision for an estimate of an individual patient, but its not clear that its relevant or possible to measure and compare at the facility level. Thus organ dose was not proposed as a practical or useful metric for patient safety assessment.

The output of radiation from the machine is far simpler to measure and in fact is the important variable, as this is what the radiologist and the technologist can influence.  The measures are primarily proposed to reflect the average CT dosing at the institutional level and small variations in patient size will average out across institutions.

We have conducted comparison of each of the dose metrics with measures of absorbed dose among a sample of 10,000 CT examinations and the correlations are high (> 90%). Further, the correlation within the metrics are also high. Details of this comparison were provided at the time of consideration of this measure when it was first endorsed. The organ doses were calculated by Dr. Choonsik Lee, PhD an Investigator in the Radiation Epidemiology Branch, in the Division of the Cancer Epidemiology and Genetics at the National Cancer Institute. His research includes the development of dosimetry databases and Monte Carlo dose calculations using human models that permit estimating absorbed radiation dose that takes into account patient weight. His method for estimating organ doses has been validated against direct measurement.

**2b2.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)

The different metrics are highly correlated *(see Keegan JACR 2014, Attachment of UCSF CT DOSE Report).* The metrics are highly correlated with absorbed doses.

**2b2.4. What is your interpretation of the results in terms of demonstrating validity**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

The metrics are valid and meaningful and will reflect a facilities average CT doses

Of note, these are not patient level metrics and for an individual patient does not provide information about whether the dose that were used was appropriate.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b3. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b4***](#section2b4)

The measure is meant as a tool for facilities to understand their practice. It is not a tool to assess the appropriateness of dose for an individual patient, as patient and facility level factors can influence the choice of dose. Since this measure reflects an average of all patient groups, no individual patients need to be excluded.

**2b3.1. Describe the method of testing exclusions and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

**2b3.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

**2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis.*  *Note:* ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b5***](#section2b5)***.***

**2b4.1. What method of controlling for differences in case mix is used?**

**No risk adjustment or stratification**

**Statistical risk model with** Click here to enter number of factors **risk factors**

**Stratification by** three anatomic areas and 5 age groups **risk categories**

**Other,** Click here to enter description

**2b4.2. 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 (case mix) is not needed to achieve fair comparisons across measured entities**.

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

The radiation doses used for CT vary by anatomic area (head, chest, abdomen and pelvis) and age (adult versus various child age groups.) *Smith-Bindman (JAMA Int Med 2009; JAMA 2012); Miglioretti JAMA Pediatrics 2013, JACR 2014).* There are further the categories that have been used in radiation safety programs for data collection In Europe and the UK. While there are other factors that influence radiation doses, these are relatively minor in comparison to these groups, and it is not feasible to collect or report data into smaller stratifications. Nor does the measure lose validity by not stratifying by clinical indication or patient size (see below.)

Why it is not important to stratify for clinical indication or protocol

The way CT scans are conducted should vary by why the patient is being scanned. For example, a search for occult malignancy may require very different parameters than an assessment for bleeding for trauma. Unfortunately there is currently no standardization for either categorizing the indications for imaging, nor for deciding the best way to image given a patients suspected problem, nor for categorizing the protocols that are used. I am currently leading a project to standardize the protocols we use for imaging across the five University of California Medical Centers and for each indication, the different institutions have adopted dozens of different ways to image patients with similar clinical questions. Smith-Bindman (*JAMA Int Med 2009*) highlighted the issue of differing ways to imaging a very standard problem and the resulting radiation dose. For example, while some institutions chose to image patients with suspected stroke using a standard 2 mSv CT, others routinely use a high dose CT where the doses were on average 20, and some facilities used 58 mSv. There are almost no standards for defining how to image different clinical questions and the profound variation reflects physician preferences, and the promotion of certain protocols by the manufacturers, rather than evidence that the higher dose protocols are more accurate or diagnostic or truly needed based on evidence. As an example, in a NEJM article (*Smith-Bindman 2010*) images were included from a patient who underwent two chest CT examinations for the same clinical indication at the same institution one year apart. The patient had a 1.5 mSv dose study on one occasion and a 15.9 mSv study on the second occasion. Both studies were done for exactly the same reason of the surveillance of a pulmonary nodule, and both were done within a single institution and within the setting of a clinical trial where what was done should be standardized. Thus there is profound variation in how studies are conducted, even in the few situations where the reason for imaging is known and guidelines exist. Further, there are often financial incentives that drive the decision to image using repeated imaging protocols versus single imaging protocols (even though the former could lead to doses that are twice as high as the latter). For example, there were recent reports that some facilities use double imaging protocols (with and without contrast) for conducting Chest CT, thereby double billing and double radiating the patient, in a setting where doing two scans is considered rarely necessary. Thus while some facilities were using double scanning in 1% of patients, others were using this in 80% of patients, and CMS has concluded that this reflected overuse of CT(see <http://www.nytimes.com/2011/06/18/health/18radiation.html?_r=1>)

Anatomic area, rather than specific indication or protocol, will actually provide the patient with the information they want to know – i.e. if I go to a facility, how high or low will my dose be. It will also allow facilities to identify where they need to explore their doses in greater detail to assess why they are outside the normative range – is it that the are using too high doses within a protocol or using high dose protocols too often. The way the measure is currently written the choice of protocol will be reflected within the facilities metrics, whereas if dose were reviewed only within protocol, the facility that chooses to use high dose studies and repeated studies on most of its patients would appear fine.

Why it is not important to adjust for patient size

Weight will contribute to the variation in dose used for CT, and if individual patients were compared, it would be extremely important to assess weight when deciding about optimum ways to set up CT scans. Differences in weight may account for a 1-3 fold difference in the radiation used. Dr. Huda has published several relevant recent papers showing that doses vary up to 2 fold based on patient weight . “Radiation related cancer risks in a clinical patient population undergoing cardiac CT” AJR 2011 and “Estimating cancer risks to adults undergoing body CT examinations” Radiation Protection Dosimetry 2011. However, its important to point out that it is in no way established exactly how to increase doses for larger patients – i.e. there is no clear standard. A recent and interesting article found that machines that automatically adjust for patient weight seem to be giving too much radiation so that the organ doses increase even more so than does the weight (Israel, G. M., Cicchiello, L., Brink, J. and Huda, W. Patient size and radiation exposure in thoracic, pelvic, and abdominal CT examinations performed with automatic exposure control. Am. J. Roentgenol. 195, 1342–1346 (2010).

We have assessed the association between weight and the doses used, and presented at the initial submission of this metric, with an explanation of why it is not important to adjust for weight. When we compared the radiation dose used among patients in the top quartile of weight, to the radiation dose used in the bottom quartile of weight, the average doses increased by a factor of less than 2. For example among adult patients age 25 and older in the lowest quartile of weight (i.e. those under 152 lbs) the mean DLP among patients who underwent an abdominal and pelvic CT was of 781 . Among patients in the largest quartile of weight (ie those between 220 and 425 lbs, reflecting a mean weight twice as high), the mean dose was 1282 DLP or around 60% higher. However, within each of the weight groups, there was much more dramatic variation within group, then between groups. For example, among the smallest patients (those <25%) the range in dose between the 1st and 99th distribution was 54 – 1890 (40 fold variation between the highest and lowest group), and in variation in the highest quartile of weight was 352 – 2885 (8 fold variation). Thus the variation in dose based on weight was small in comparison to weight based on other factors (such as physician and facility preferences).

These weight differences are not relevant at the facility level, as while patient size may influence dose by 2 fold (between the smallest and largest patients) other factors, can influence the dose by up to 100 fold (based on our data), and these factors, rather than individual patient weight, will drive the facility level dose indices measures. Even if a facility had ALL patients of a size <25%, versus all patients over the 75% the influence would be very modest.

However, while I do not believe including weight would influence a facility’s measures, there have been several recent publications which provide simple ways to account for size when reporting radiation dose, and including one of these metrics in the measure may allow greater adoption of the measure by various stakeholders. These measures essentially have determined for a fixed amount of machine dose, how the absorbed dose to the patient varies by their size; larger patient will tend to have a lower adjusted dose (because the same dose is spread out in their larger body) whereas a smaller patient will have a correspondingly larger dose (because the same dose is distributed in a small volume of tissue.) Using these adjustment factors, it is possible to get a more precise estimate of the dose absorbed by the patient based on the machine output and a conversion factor based on the patient’s size. The SSDE measurement ( AAPM Report 204, Size Specific Dose Estimates in Pediatric and Adult Body CT Examinations) is now included in this measure and is a measure that accounts for patient size.

In our work across the University of California Medical Centers (as part of the UCDOSE, PI Smith-Bindman), among > 100,000 CT scans, there was no difference in facility level conclusions about performance when any of the metrics were used (i.e. SSDE, CTDIvol, DLP and ED) all characterized facilities the same.

**2b4.4. What were the statistical results of the analyses used to select risk factors?**

**2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps―do not just name a method; what statistical analysis was used*)

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***If stratified, skip to*** [***2b4.9***](#question2b49)

**2b4.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):

**2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:  
  
**2b4.9. Results of Risk Stratification Analysis**:

**2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)

**2b4.11.** **Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

**2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Comparing institutional performance to benchmarks permits identification of outlying performance. Because the metric is based on summarizing dose for a large number of individuals (> 100 within each strata) and comparison to benchmarks, the comparisons are stable at identifying outlying performance. In the attached document (UCSF CTDOSE Report), we illustrate the result of comparing institutions (using t-tests and quantile regression) using the NQF measure format. Basically, facilities can be identified and compared with benchmarks, and stable estimates of facilities with outlying performance can be identified. *See Miglioretti 2014 JACR, Keegan JACR 2014*

While the generation of averages will permit the comparison of facilities to benchmarks, the measure does not specify cutoffs or how the comparisons would be judged. These can be set based on the clinical or quality improvement needs of a facility, organization, etc.

**2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

**2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

Outlying institutions can be easily identified

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

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

**Note***: This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

**2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different data sources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)

**2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)

**2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps―do not just name a method; what statistical analysis was used*)

The measure calls for collecting consecutive scans so that participants cannot choose their best or most optimum dose metrics to quantify. The data will be available, or can be calculated from essentially all (>95%) of CT scans

**2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g.,**results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

**2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)