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

**Measure Number** (*if previously endorsed*)**:** 0740

**Measure Title**: Participation in a Systematic Dose Index Registry

**Date of Submission**: Click here to enter a date

**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.    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.  **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)   * No exclusions   **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**  **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**.  All automated.  **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 |

Data files from the American College of Radiology Dose Index Registry (DIR) were used for testing. The files contain data on CT radiation dose indices in a standardized format, submitted from participating facilities using an automated method. Data on all CT exams performed at a facility are submitted to the DIR registry. Current lists of CT scanners from which data have been received are posted on the registry web-site at this link: <http://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/NRDR/DIR/scanners.pdf>

All data elements collected in the registry are posted on the Dose Index Registry website (<http://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/NRDR/DIR/DataElementsInDIRHeaderSR.pdf>) and are based on DICOM Radiation Dose Structured Report as published in DICOM Content Mapping Resource document PS 3.16-2009 (<ftp://medical.nema.org/medical/dicom/2009/09_16pu.pdf> ). For more information, please see <http://www.acr.org/Quality-Safety/National-Radiology-Data-Registry/Dose-Index-Registry/About-DIR>

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

The data sample includes all facilities currently participating in the American College of Radiology Dose Index Registry, which was launched in May 2011. Facilities must register to participate in the registry. A list of registered facilities is maintained. The participating facilities represent various types (such as academic, community), locations (such as metropolitan, rural), parts of the country, and sizes (based on procedure volume). The registry includes data from all major manufacturers of CT scanners, and from scanners of varying ages.

See here for lists of representative facilities (<http://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/NRDR/DIR/Representative%20List%20of%20DIR%20Facilities.pdf>) and here for scanners (<http://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/NRDR/DIR/scanners.pdf>).

**1.3. What are the dates of the data used in testing**? Click here to enter date range

Data from May 2011 – December 2013 are used to calculate for participation statistics and for testing main attestation measure.

Data from July 1, 2011-June 30, 2013 are used for calculating trends in dose indices; these trends provide supplemental information on the attestation measure.

**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: Click here to describe | other: Click here to describe |

**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*)

At the time of NQF endorsement (September 2011) there were 104 sites actively participating/submitting data. As of December 2013, there were 510 sites that had submitted data to the registry.

Number and characteristics of facilities participating and submitting data: Of the 510 facilities contributing data through 30 December 2013, over 44% were community hospitals and an additional 32.5% were free standing non-academic facilities. Academic facilities made up 13.6%, multispecialty clinics were 3.7% and children’s hospitals 3%. The locations types were diverse as well – 12.5% of the facilities were rural and one-half were metropolitan, and facilities were distributed all over the country. For more details please refer to Table 1 in the uploaded Appendix.

**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*)   
As of December 2013, the DIR included data on 6.9 million exams, covering adults and children, men and women. 53% of exams were on female patients, and 5% were on children. For analysis of trends, all exams at facilities that were in the registry between July 2011 and June 2013 were included. However, patient demographics do not influence any of the tests.

**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**.

The trend analysis sample only includes facilities that were continuously in the registry between July 2011 and June 2013.

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

**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*)

Since this measure is an attestation measure, reliability testing in terms of assessing repeatability of data elements is not applicable. However, reliability of the measure score (Y/N) was analyzed based on precision of registry data element completeness for participants in the registry (ACR DIR) used for source data. This is described below.

**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*)

Mean, median, and inter-quartile range of exams submitted per facility: On average, facilities had submitted 44,785 valid records per facility. Half the facilities contributed 32,984 records or more; 75% of the facilities contributed 15,557 or more records; and 25% of the facilities contributed 69,011 records or more.

Percent records with complete information on critical elements: Three quarters of participating facilities submitted data where 99% of the records had valid information for study description (standardized procedure name) and CTDIvol, which are critical for reporting on dose indices and comparisons. On average, facilities had 96% of their records with complete information on the critical elements, and the median facility had complete data on over 99.7% of its records.

Percent of records where exams were mapped to a standardized procedure name for meaningful comparison: On average, 81% of facilities’ records mapped to a valid standardized procedure name. For the median facility, almost 92% of the records mapped to a valid standardized procedure name. Records that are currently mapped may be associated with a standardized name in the future, and this standardized name will be applied retrospectively to all facility records.

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

The high percent of records with complete data indicate that participation in the registry is indicative that feedback reports generated by the registry are meaningful to facilities for quality improvement.

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

**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)*.

We conducted 5 tests of validity:

1. Hypotheses that the measure score indicates quality; scores are different for groups known to have differences in quality assessed by another method:

We assessed performance rates for measures included within the ACR DIR to see if there were trends indicating that participation results in improved quality, e.g. decreased dose indices. We selected three high volume examinations – head without contrast, chest without contrast, and abdomen-pelvis with contrast. For facilities that were in the registry for every six month period between July 2011 and June 2013, we calculated the median dose index values (CTDIvol and DLP). This median represents the typical dose indices for that protocol at that facility. We calculated the mean value of the dose indices across all included facilities for each period and plotted these charts.

We perform the trends analysis on CTDIvol and DLP and not the Size-Specific Dose Estimate (SSDE) because not all facilities send us localizer images on all exams to support estimation of SSDE. We provide feedback on SSDEs for all facilities that provide us with data to support that estimation, and we are encouraging facilities to move in the direction of SSDE because that is the measure most closely related to patient dose.

1. Correlation of measure scores with another valid indicator of quality for the specific topic:

We reviewed other quality improvement activities that similarly assess quality based on dose index registry participation.

1. Participation trends:

We evaluated changes in participation in the ACR DIR to determine potential trends in increased uptake and usage of the registry. We plot both the number of facilities and the number of examinations in the registry over time.

Participating facilities who mapped their study descriptions to standard procedure names receive feedback reports every six months comparing dose indices by exam for each facility to corresponding dose indices at similar facilities. Separate reports are generated for exams performed on adult patients (older than 18) and exams performed on pediatric patients (age 18 or younger). Reports contain comparisons of a facility median dose index value for an exam with corresponding values at facilities across the registry, facilities of same type (such as academic, community), in similar locations (such as metropolitan, rural), and in the same census region. Dose indices in the report include CTDIvol, DLP, and starting in September 2012, Size-Specific Dose Estimates (SSDE). Detailed charts are provided in pdf format for exams with high volumes at a large number of facilities. Additional information on lower volume exams are provided on a spreadsheet. Links to sample reports are provided below.

Also facilities have access to their own data at all times and can view how their data compare to registry averages.

1. Face validity:

We reviewed publications, quality improvement programs and other activities that provide evidence that participation in registries does improve performance, or that identify measurement of dose indices such as CTDIvol, DLP, and SSDE is a valid approach to improving quality.

1. Case studies:

We gathered case studies from participating facilities of how registry data were used for quality improvement.

**2b2.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)  
A. Hypothesis that measure score indicates quality:

The trend charts from the analysis are included in the Appendix as Figure 1. For the four six-month periods between July 2011 and June 2013, 105 facilities had data in all periods for head exams 103 had data for chest exams, and 107 had data for abdomen pelvis exams.

For head exams, the average CTDIvol across participating facilities decreased by 7.8% and the DLP by 6.4% during the two year period. For abdomen pelvis exams, the average CTDIvol across participating facilities decreased by 17.3% and the DLP by 16.5% for the same period. For abdomen pelvis exams, the average CTDIvol across participating facilities decreased by 13% and the DLP by 11% for this period. None of these are statistically significant because there is substantial variability across facilities in their median dose indices, but these to point to movement in the right direction.

C. Participation trends

The number of participants and number of exams in the registry have grown since the launch of the registry in May 2011. See Figure 2 for illustration of this growth.

Sample semi-annual reports provided to facilities are posted here:

Adult patients: <http://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/NRDR/DIR/DIRSampleReport.pdf>

Pediatric patients: <http://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/NRDR/DIR/DIRPedSampleReport.pdf>

Facilities also receive detailed color-coded spreadsheet reports that help them identify high-priority for review and optimization, and a spreadsheet report with trends to help them determine which protocols have changes in dose indices over time. An extract from these spreadsheet reports is included in the appendix as Figure 3.

Figure 4 includes examples of live reports available to facilities to examine their own data at any time between semiannual reports.

**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?*)

1. Hypotheses that measure score indicates quality:

The decrease in dose indices over time indicates that participation in the DIR is associated with improvements in performance quality with respect to radiation dose optimization among participating facilities. While we cannot measure the direct quality differences between participants and non-participants, the trend suggests that the registry provides a valuable infrastructure to help facilities improve protocols. This infrastructure supports the implementation of accountable quality measures such as percent of exams with dose indices within desired target ranges. We intend to submit accountability measures for consideration by NQF in the next few months. But even with the accountability measures, there is a need for a simple attestation measure for facilities to be able to build the infrastructure to report on more advanced measures.

1. Correlation of measure score with other similar indicator:

The correlation of this measure with requirements and recommendations from other organizations points to the value of this measure.

* As of December 2013, the American Board of Radiology has recorded 20 diplomates that have indicated using the ACR DIR Practice Quality Improvement Project for Part IV requirements.
  + The Joint Commission Sentinel Alert Issue 47 – Radiation risks of diagnostic imaging, August 24 2011 <http://www.jointcommission.org/sea_issue_47/>

Actions suggested by The Joint Commission:

Health care organizations can reduce risks due to avoidable diagnostic radiation by raising awareness among staff and patients of the increased risks associated with cumulative doses and by providing the right test and the right dose through effective processes, safe technology and a culture of safety.

In addition, The Joint Commission [endorses the creation of a national registry to track radiation doses as the start of a process to identify optimal and reference doses].

* The Joint Commission Standards: Revised Requirements for Diagnostic Imaging Services; prepublication December 30, 2013 <http://www.jointcommission.org/assets/1/6/PREPUB-12-20-2013-DiagImaging_AHC.pdf>

Standard PI.02.01.01

The organization compiles and analyzes data.

Elements of Performance for PI.02.01.01

A 6. For organizations that provide diagnostic computed tomography (CT) services: The organization compiles and analyzes data on patient CT radiation doses and compares it with external benchmarks, when such benchmarks are available.

1. Participation trends:

The increasing numbers of facilities participating in this voluntary registry points to the value provided by it for quality improvement. The number of exams and facilities increased in each six month period despite there being no reporting requirement of obvious monetary incentive for participation. The sample reports demonstrate some of the tools received by participants that help them optimize radiation doses.

1. Face validity:

Recognized quality and safety programs and organizations that identify tracking dose indices as a means for improving quality indicate that the measure score can be used to distinguish good from poor quality. Programs we identified are shown below.

* + CMS Physician Quality Reporting System (PQRS) 2014 Measures Group – Optimizing Patient Exposure to Ionizing Radiation (OPEIR) includes a similar measure on reporting to a dose index registry. [2014 PQRS Measure Groups Specifications, Release Notes, Getting Started with 2014 PQRS Measures Groups, 2014 Quality-Data Code Categories, and 2014 PQRS Measures Groups Single Source Code Master](http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/Downloads/2014_PQRS_MeasuresGroupSpecs_ReleaseNotes_SupportingDocs_12132013.zip) ; 2014 Physician Quality Reporting System (PQRS) Measures Groups Specifications Manual 12/13/2013, page 321
  + American Board of Radiology – Approved Maintenance of Certification Part IV Practice Quality Improvement Project.

<http://www.theabr.org/moc-dr-pqi-projects> and [Dose Index Registry Example](http://www.theabr.org/theABR/sites/all/themes/abr-media/pdf/Dose_Index_Registry_Example.pdf)

* + U.S. Food and Drug Administration Initiative to Reduce Unnecessary Radiation Exposure from Medical Imaging

<http://www.fda.gov/Radiation-emittingProducts/RadiationSafety/RadiationDoseReduction/default.htm>

Tracking Radiation Safety Metrics – Dose Registries

Dose registries would enable facilities to compare their radiation doses to those delivered in other facilities for the same exam, and such comparisons over time could assist in optimizing patient radiation doses for medical imaging. The goals of tracking imaging exams and the associated radiation exposure include: (1) providing information at the point-of-care for the referring practitioner (i.e. supporting justification); (2) promoting development and use of diagnostic reference levels (DRLs) (i.e. supporting optimization); (3) providing information for assessment of radiation risks; and (4) establishing a tool for use in research and epidemiology.

<http://www.fda.gov/Radiation-EmittingProducts/RadiationSafety/RadiationDoseReduction/ucm299368.htm>

1. Case studies:

Facility representatives contributing case studies – two speakers at the 2013 ACR Annual Informatics and Registries Summit contributed case studies describing their facility’s experience using the registry data to improve quality (See links to speaker presentations here: <http://www.acrinformatics.org/DataRegistriesForumAgenda.aspx>). Additionally, one facility representative provided experience summaries for inclusion in the ACR case studies in “Imaging 3.0” (more information at <http://www.acr.org/Advocacy/Economics-Health-Policy/Imaging-3/Case-Studies> ).

**Case Study 1: (imaging 3.0)**

**Demographics**

477-bed general medical and surgical hospital

**Previous Approach to Optimizing Patient Exposure to Radiation Dose**

The hospital relied almost exclusively on vendor education, physics evaluations, and radiologist protocols to establish radiation dose. They had “no true analytic data to set goals for lowering dose.”

**Study Approach to Optimizing Patient Exposure to Radiation Dose**

Apply CT software (that claimed) to reduce dose up to 40%, and compare the facility to one without the new technology. A committee was created to promote cohesive work within the department through the transitions in protocol. They met quarterly to review and analyze DIR data.

**Results**

* Since the hospital started using the DIR, they stopped performing delayed scans of the kidneys on patients 18 and younger- those who are at greatest risk for radiation (before the DIR, these additional images were standard for all Abdomen/Pelvis CTs).
* Staff radiologists verified that their dose-saving reconstruction algorithms have reduced dose by 30-40%, while maintaining image quality.
* A reduction in CTA Head and Neck studies was implemented after the DIR showed that the dose for these specific studies was slightly above the national average.

**Conclusion**

* Protocols were improved “across the board” by viewing the aggregate data.
* DIR provides a powerful tool for protecting patients by comparing results with national standards.
* The DIR is an indispensable tool for achieving a status of As Low As Reasonably Achievable (ALARA) for images.

**Case Study 2:**

**Demographics**

A 666-bed academic hospital with a level 1 trauma center.

**Previous Approach to Optimizing Patient Exposure to Radiation Dose**

Retrospective instead of proactive approaches to dosage levels and mistakes in studies (three case studies presented in results section).

**Study Approach to Optimizing Patient Exposure to Radiation Dose**

49,500 CT exams were performed in 2012; down 23% from 2010. The main concern was CT dose, and dose management (using the NCDR DIR) became the primary focus for safe and economic reasons. Proactive applications of dose monitoring were applied, including a dose alert case study as well as a size-specific DRL (WIP) case study.

**Results**

Retrospective Case 1: CTPA study

* Comparison of effective dose for 800 patients
* Very time-consuming (before DIR: 3-4 months, post-DIR: <2 hours)

Retrospective Case 2: CT Chest without contrast vs. CT Chest with contrast

* From January-June, 2013: CT Chest without contrast dosage was less than or equal to peers
* From January-June, 2013: CT Chest with contrast dosage was approximately 20% higher than some peers

Retrospective Case 3: Evaluation of ASIR’s performance

* The target dose reduction rate for CT of the Chest/Abdomen/Pelvis with contrast was 20%, however, the actual dose reduction rate achieved was approximately 18%.

Proactive Case 1: Dose Alerts

* CT operators were alerted when scan settings exceeded dose thresholds prior to acquisitions (inputs were from the institutional DIR report, e.g. 75th percentiles)

Proactive Case 2: Size-specific DRL (WIP)

* Factors affecting the CT technique and dose were found to include:

1. Anatomy/pathology
2. Patient size/habitus
3. Age: adult vs. pediatric
4. Indication: screening vs. follow-up
5. Availability of dose management technologies: ATCM, SIR, special scan modes, etc.

**Conclusion**

The root cause for the (approximately 20%) higher CT Chest with contrast dosage from January-June, 2013 was that radiologists at the facility had been prescribing the same protocol, despite the existence of a lower dose protocol.

DIR is an extremely valuable tool of quality assurance and patient safety for participating facilities, with usage not limited to retrospective application.

Exam-specific, size-specific DRLs are more appropriate, and ACR is the best resource to lead this long-term vision.

**Case Study 3:**

**Demographics**

Women and children’s 200 bed facility averaging 75,000 annual exams, part of larger 4 hospital and vascular institute health system.

**Previous Approach to Optimizing Patient Exposure to Radiation Dose**

In 2011 the established Dose Reduction Committee determined data at the facility to be limited, and identified CT Abdomen as an opportunity for dose reduction. In 2012 data was again found to be limited, and the Dose Reduction Committee identified CT Paranasal Sinuses without contrast as a new opportunity for dose reduction.

**Study Approach to Optimizing Patient Exposure to Radiation Dose**

2011: The Dose Reduction Committee formulated the following plans of action to address dose reduction for CT Abdomen studies:

1. Slice thickness and pitch were to be reviewed
2. Middle of software upgrade where CARE KV would be installed
3. Optimize workflow and protocols to better utilize CareDose and CarekV
4. Hesitant to change too much too fast

**Results**

2012 Data: Data was benchmarked against the NRDR DIR; 209 participants in 2012 data across 7 U.S. Children’s hospitals, broken into age groups of 0-2, 3-6, 7-10, 11-14, and 15-18 years.

CT Head Brain Without IV Contrast: CTDIvol Per Scan (Age Group 7-10):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | N | Median | Mean | St Dev |
| Facility | 4 | 22 | 22 | 2 |
| All DIR Sites | 1433 | 45 | 43 | 17 |

CT Abdomen/Pelvis With IV Contrast: CTDIvol Per Scan (Age Group 3-6):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | N | Median | Mean | St Dev |
| Facility | 4 | 17 | 18 | 2 |
| All DIR Sites | 379 | 4 | 6 | 8 |

* From 2011-2013 the facility decreased CTDIvol by >75% in CT Abdomen/Pelvis with IV contrast in 3-6 year olds.
* Reduction of CTDIvol of 50% in CT Head Angio with IV contrast in 11-14 year olds.
* From 2011-2012, the facility experienced an overall reduction in cases that were overexposed, and increase in cases that were underexposed, and continuation of dose reduction education for staff.

**Conclusion:**

2012 data showed the facility was able to exceed the national benchmark demonstrating lower radiation dose for Median CT Head Brn without IV contrast, CTDIvol per scan.

Challenges faced by the facility included:

1. Mapping (protocol naming)
2. CT scanner software upgrades
3. Now able to lock CT protocols
4. Communicating CT changes to all techs
5. Getting agreement by all radiologists
6. CT Technologists identified inconsistencies in protocols (varied by physician preference)

2012: The Dose Reduction Committee formulated the following plans of action to address dose reduction for CT Paranasal studies:

1. Safire installation
2. Review CT Paranasal protocol and make recommendations

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

**2b3. EXCLUSIONS ANALYSIS**

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

**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)***.***

Risk adjustment is not directly applicable to the attestation measure on participation in a registry.

Measures within the registry on dose indices and estimates for exams are risk adjusted by stratification of facilities, standardization of procedure names, and adjustment for patient size with the use of size-specific dose estimates (SSDE).

In the reports, summary dose indices are stratified by facility type, facility location, and census division. When facilities receive their reports, they can compare their performance on a procedure or protocol with the performance of other facilities similar to their type, location, census division, and across all registry participants. The stratification scheme accounts for differences in patient populations across geographies, and differences in the types of patients seen at different types of facilities, for example, patients with advanced disease may be more likely to be seen at academic centers. In the absence of detailed information on patient indication or case history, facility characteristics serve as a proxy for differences in patients receiving the same procedure or exam.

Procedure names are standardized to a standard terminology, RadLex (<http://playbook.radlex.org>), to ensure comparability across facilities. Without standardization, procedure names vary widely across and within facilities, making it impossible to perform any comparisons for performance assessment.

Variations in patient sizes account for necessary and justifiable differences in radiation exposure. To account for this, we calculate and compare size-specific dose estimates for all body exams, using guidelines from the American Association for Physicists in Medicine.

**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** Click here to enter number of categories **risk categories**

**Other,** Adjustment is not applicable to the attestation measure. Measures within the registry are stratified on the multiple dimensions described in Table 1.

**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**.

The measure is an attestation measure for participation, and does not use patient level information. As a result, risk adjustment or stratification are not really applicable to the measure itself. As described above, measures within the registry are appropriately standardized and stratified for comparison.

**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 idea of risk adjustment by stratification is to find a way of comparing apples to apples. In DIR, the participating facilities vary in their patient populations and the types of indications for which patients receive imaging procedures. But the registry has no data on indication or patient history; patient data are limited to age and gender. Patient populations vary by facility characteristics, and therefore, we use facility characteristics as a proxy for case mix. DIR provides a way for facilities to compare themselves to similar facilities as well as to all DIR facilities nationwide. Facilities can take a step by step approach to optimize their dose indices, possibly starting with protocols most different from their peers and attempting to be closer to regional peers before attempting more challenging targets.

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

**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*)**:**N/A.

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

**2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:  
N/A.

**2b4.9. Results of Risk Stratification Analysis**:

N/A.

**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*)  
N/A. There is no stratification or adjustment applicable to the attestation measure.

**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)*

As this is an attestation measure, we do not have any data on non-participants and are unable to measure differences in quality with non-participants. But trend data on dose indices indicate decreases over time in radiation exposure for facilities in the registry. Trend results for three high volume exams are presented above as part of validity testing. See Figure 1 for trend charts.

Sample reports demonstrate variability in dose indices that support the need for dose monitoring. However, the participation measure does not directly address any metrics monitored within the registry. The benefit of the participation measure is that it helps develop the infrastructure to implement meaningful measures on dose optimization and identify meaningful benchmarks.

**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*)  
Our internal data indicate that there are approximately 8,500 facilities performing CT in the United States, and our participation of 510 facilities, or our registered base (including facilities in the process of getting set up for data submission) is a little over 658. While some of the facilities not in the registry undoubtedly have dose monitoring measures in place, there is still a large gap in the use of standard infrastructure to facilitate meaningful standardized comparisons.

Among facilities that are in the registry, there is a trend towards decreasing dose indices over time. While this is not any indication of whether facilities in the registry are better than the ones that are not, this does point to the opportunity for improvement for facilities in the registry.

**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?*)

Opportunities for improvement:

As we expand participation and more facilities start to use standardized reports with size-adjusted estimates for standard procedure names, we are now able to develop accountability measures for radiation dose monitoring that steer participants towards evidence-based benchmarks. We intend to submit some on these accountability measures to NQF in the next few months. Such measures may include the use of pediatric protocols for pediatric patients, attaining radiation dose indices within target ranges for high- volume exams, and the use of standardized procedure names for meaningful comparison and communication.

In the next year, we anticipate adding computed and digital radiography exams to the registry, and subsequently, fluoroscopy and nuclear medicine. The goal of the registry is to include all exams that use ionizing radiation. As we add each imaging modality, the registry participation measure will continue to provide support for the development of infrastructure. As the data collection process for each modality matures, we will be able to develop additional measures to monitor and reduce variations in performance quality. For example, as we collect sufficient data on patient sizes and radiation dose indices for each exam, we can develop meaningful benchmark ranges.

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

**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*)  
 Since the only method and data source for the measure is the ACR Dose Index Registry, comparability is not a significant issue for assessing participation.

However, the manner of data acquisition at a participating facility prior to submission does vary. Some facilities send data directly from scanners, and others send data from local PACS (Picture Archiving and Communication Systems). There are differences in scanner manufacturers and models and the age of scanners. Scanners have different capabilities – some generate DICOM radiation dose structured reports (RDSR), others generate screen images with dose information, and yet others do not display any dose information. The dose screen images vary across scanner manufacturers and models. The registry software can accept data from all major manufacturers of CT scanners, General Electric, Siemens, Phillips, and Toshiba, and can process radiation dose structured reports and screen images. In addition to the registry software, some facilities use commercial or third-party radiation dose monitoring software such as Radimetrics, DoseMonitor, Radiance, Aware, or DoseWatch. The registry software cannot obtain dose index measures from scanners that do not even generate screen images, but some of the third-party solutions can.

As mentioned previously, exams are labeled using different schemes for procedure names and there is a great deal of variability.

Despite different sources, all records are converted to DICOM radiation dose structured report (RDSR) format before submission to the Dose Index Registry. The data contain a standard set of DICOM data elements as specified in the data dictionary. The same mapping tool is provided to all facilities to standardize exam names. A description of the tool and its use can be found here: <http://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/NRDR/DIR/MappingToolUserGuideDIR.pdf>.

The 510 facilities participating in the Dose Index Registry are used as a data sample to illustrate the extent of variation in local systems.

**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*)

As evidence that the DIR is able to accept data from multiple scanners, a list of scanners by type and software version (where available) in the registry is tabulated. (<http://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/NRDR/DIR/scanners.pdf>)

There is variability in the exam names received from multiple facilities. In the report generated on data submitted between January and June 2013, 19439 distinct study descriptions mapped to 908 standardized procedure names. One standard exam name, “CT ABDOMEN WITH IV CONTRAST” was associated with at least 198 distinct study descriptions. On average, 81% of facilities’ records mapped to a valid standardized procedure name. For the median facility, almost 92% of the records mapped to a valid standardized procedure name. Records that are currently mapped may be associated with a standardized name in the future, and this standardized name will be applied retrospectively to all facility records.

**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*)

As described in the reliability section above, a very large percent of submitted data are complete. Also, all data are obtained in the same format. The data are fairly consistent across sources and it is not particularly productive to investigate comparability.

To the extent that scanners are different, the DIR received valid data from a wide range of scanners. The DIR was able to receive exam names with a large amount of variability and standardize the information for meaningful comparison.

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

**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*)

Of the 510 facilities with valid data submission in the registry, 104 were in the registry at the time that Measure 0740 was endorsed in September 2011.

Mean, median, and inter-quartile range of exams submitted per facility: On average, facilities had submitted 44785 valid records per facility. Half the facilities contributed 32984 records or more; 75% of the facilities contributed 15557 or more records; and 25% of the facilities contributed 69011 records or more.

Records submitted by facilities were analyzed for missing data on critical elements.

**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*)

Three quarters of participating facilities submitted data where 99% of the records had valid information for study description (standardized procedure name) and CTDIvol, which are critical for reporting on dose indices and benchmarks. On average, facilities had 96% of their records with complete information on the critical elements, and the median facility had complete data on over 99.7% of its records.

Percent of records where exams were mapped to a standardized procedure name for comparison and benchmarking: On average, 81% of facilities’ records mapped to a valid standardized procedure name. For the median facility, almost 92% of the records mapped to a valid standardized procedure name. Records that are currently mapped may be associated with a standardized name in the future, and this standardized name will be applied retrospectively to all facility records.

**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*)

Operational use has taught us a great deal about the variability across facilities. Some examples are as follows:

* We were aware early on about the wide range of procedure names across and even within facilities to describe the same exam. As we worked through the process of standardized names, we learned what kinds of assistance and tools facilities needed to be able to implement the standardization.
* The experiences of our facilities helped the lexicon developers (RadLex) implement user-friendly interfaces and continues to inform the development of a complete and useful lexicon.
* We were aware of differences in scanner output with regards to dose information, but the extent of variation between scanners was larger than we had anticipated. We learned that we needed to use our relationships with manufacturers to seek their assistance in testing the software that we use for data collection. We also sought advice from manufacturers when trouble-shooting the data acquisition process and used their support in helping their customers get started with data submission.
* We have learned that our facilities are our best resource in solving operational issues and we recruit their assistance in helping other participating facilities get started with data submission. Facilities often understand hands-on operational issues much better than registry staff or manufacturer representatives.
* We continue to learn more about facility operations that we do not fully accommodate and our oversight committee will help us find consensus solutions. For example, some perfusion exams may have multiple monitoring scans to monitor contrast before performing a scan of the body part of interest and this may lead to misrepresentation of the dose index for an exam. Some facilities may split and rename an exam such as CT of the chest, abdomen, and pelvis and may have different body parts interpreted by different physicians; the dose index screen in this case may reflect the whole exam but the exam name may only describe one of the body parts. Our committee is working on solutions to minimize misinterpretation.