J Digit Imaging (2013) 26:217–226
DOI 10.1007/s10278-012-9515-1

Investigation of the Variability in the Assessment of Digital Chest X-ray Image Quality

Jacquelyn S. Whaley · Barry D. Pressman · Jonathan R. Wilson · Lionel Bravo · William J. Sehnert · David H. Foos

Published online: 1 August 2012

Abstract A large database of digital chest radiographs was developed over a 14-month period. Ten radiographic technologists and five radiologists independently evaluated a stratified subset of images from the database for quality deficiencies and decided whether each image should be rejected. The evaluation results showed that the radiographic technologists and radiologists agreed only moderately in their assessments. When compared against each other, radiologist and technologist reader groups were found to have even less agreement than the inter-reader agreement within each group. Radiologists were found to be more accepting of limited-quality studies than technologists. Evidence from the study suggests that the technologists weighted their reject decisions more heavily on objective technical attributes, while the radiologists weighted their decisions more heavily on diagnostic interpretability relative to the image indication. A suite of reject-detection algorithms was independently run on the images in the database. The algorithms detected 4 % of postero-anterior chest exams that were accepted by the technologist who originally captured the image but which would have been rejected by the technologist peer group. When algorithm results were made available to the technologists during the study, there was no improvement in inter-reader agreement in deciding whether to reject an image. The algorithm results do, however, provide new quality information that could be captured within a site-wide, reject-tracking database and leveraged as part of a site-wide QA program.

Keywords Image quality · quality assurance · quality control · digital radiography · computed radiography · algorithms · reject analysis · repeat analysis · image defect detection

Introduction

An estimated 182.9 million procedures were performed in US hospital radiology departments in 2010 using projection X-ray equipment with modalities including computed radiography (CR, 70 % of installed units), direct radiography (DR, 26 %), and traditional screen-film (4 %) [1]. For each captured radiograph, the technologist performing the procedure visually assesses the image for proper positioning and adequate exposure, and inspects the image for patient motion blur and other quality deficiencies that might impede diagnosis. Digitally captured images that satisfy established quality assurance (QA) criteria are accepted by the technologist (QA-accepted) and released to the Picture Archive and Communications System (PACS) for interpretation. Images that are deemed inadequate by the technologist are rejected (QA-rejected) and usually repeated. Reject rates, which are defined as the total number of rejected images divided by the total number of images acquired over an established time period, are used by most clinical sites as an integral component of an overall QA program. Reject rates for sites using digital radiography systems are reported to range between 3 % and 10 % [2–4].

The QA assessment process is performed visually and, thus, is inherently subjective [5, 6]. Some QA deficiencies that lead to the decision to reject an image are obvious such as a substantially cutoff lung field (Fig. 1) or the shadow of a necklace that obstructs the view of the spine. Some QA

J. S. Whaley (✉) · W. J. Sehnert · D. H. Foos
Clinical Applications Research, Carestream Health, Inc.,
1049 Ridge Road W.,
Rochester, NY 14615, USA
e-mail: jacquelyn.whaley@carestream.com

B. D. Pressman · J. R. Wilson · L. Bravo
Department of Imaging, S. Mark Taper Foundation Imaging
Center, Cedars-Sinai Medical Center,
8700 Beverly Blvd.,
Los Angeles, CA 90048, USA

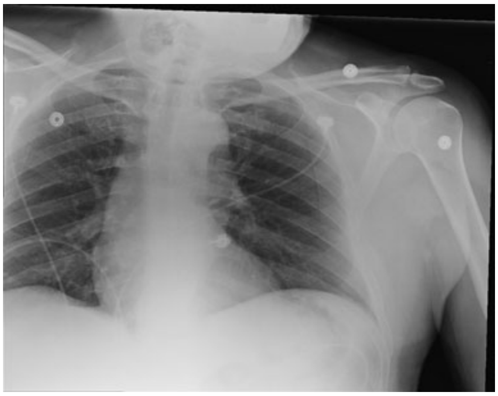

Fig. 1 Image showing substantially cutoff lung field

deficiencies are less obvious and require more careful judgment. For instance, the decision to reject a lateral chest exam because of spatially variant blurring caused by intra-exposure patient motion can be highly subjective (Fig. 2). The visibility of some quality deficiencies, such as poor contrast-to-noise ratio in the higher spatial frequencies, may be reduced or even masked for images that are presented on low-resolution modality console displays. The subjective nature of the visual QA assessment task can lead to inconsistencies among radiographic technologists in deciding whether to reject and repeat an image. This, in turn, can lead to variability in the quality of images that are delivered to the PACS for diagnostic and clinical interpretation. A recent study reported reject rates for chest X-rays that ranged from less than 1 % to greater than 8 % among technologists from one institution [7–9]. The individual technologists' reject rates were found to correlate with image quality, where those technologists with the higher reject rates delivered, on average, better quality radiographs to the PACS, while those technologists with lower reject rates produced poorer image quality radiographs. The concern with this particular finding is that higher reject rates necessarily imply a greater percentage of repeated exams and, therefore, increased radiation burden for patients. Moreover, higher repeat rates adversely influence workflow efficiency. To address these issues, clinical sites have been implementing QA programs such as enterprise-wide, reject-tracking, and analysis tools. These programs are intended to drive reject/repeat rates down while simultaneously maintaining, or even improving, overall image quality [10–12].

We have developed technology for the purpose of reducing the inherent subjectivity in performing visual QA assessments. The methodology has the potential to produce supplemental data that can be incorporated into an overall image QA program. The approach utilizes a series of computer-based, reject-detection algorithms. The concept is similar to computer-aided detection (CAD) for digital mammography, but instead of detecting and classifying potential cancer sites for the radiologist, the algorithms detect and classify QA deficiencies for the technologist. The algorithms can be applied at the point of capture, and they have the potential for providing additional information on the presence of QA deficiencies at the reject decision point. The goal is to provide ancillary information at the point of image capture to assist the technologist in cases where the quality deficiency is less obvious and to provide collateral data that, over time, may prove useful in performing reject analyses.

Many automated methods for disease detection, disease diagnosis, and materials-defect detection from radiographs have been investigated, using approaches such as statistical classifiers, fuzzy clustering, and neural networks [13–18]. We used a neural-network approach in developing algorithms to automatically predict the presence and degree of anatomy cutoff and patient motion in digital chest X-ray images. Exposure errors were detected using a feature-based, linear-model algorithm. The algorithms were evaluated by comparing algorithm-detected rejects against technologist-detected rejects and by comparing technologists' reject decisions with and without the software results.

Materials and Methods

More than 11,000 chest radiographs were retrospectively collected at the Cedars-Sinai Medical Center (CSMC) over

Fig. 2 Image showing intra-exposure patient motion

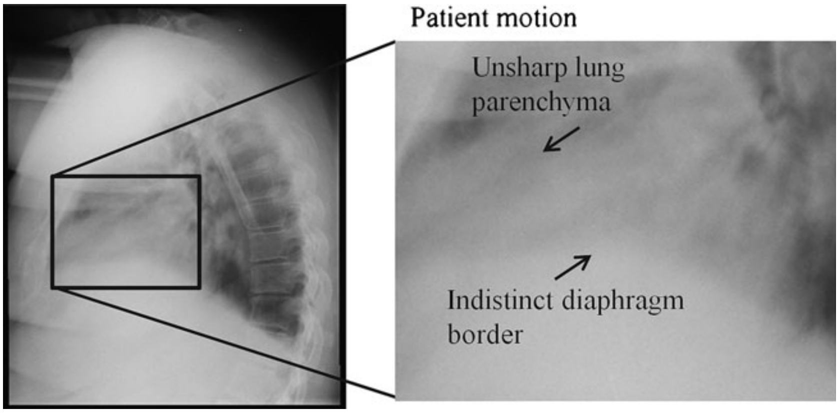

Table 1 Reject Reasons

Clipped anatomy	Over exposure
Double exposure	Patient motion
Foreign body	Plate erasure
Incorrect marker	Positioning
Other	Under exposure

a 14-month period from three DirectView 900 Series photostimulable storage phosphor CR systems (Carestream Health, Inc., Rochester, NY, USA). All QA-accepted and QA-rejected chest images were collected from these three systems. Each CR unit was configured to require the technologist acquiring the image to specify, via a check box, a reason for the rejection when an image was rejected during the QA step. Table 1 shows the ten reject reasons that were used in configuring these CR systems. Patient identification information was removed, and the de-identified images and support data were centrally archived. The case-collection protocol was approved by the CSMC Institutional Review Board.

Before running the reject-detection algorithms on the image database, it was necessary to filter out non-pertinent, QA-rejected images, including plate erasures; QA test captures, such as flat fields and test patterns; non-adult chest images; and images with reject reasons not relevant to the algorithms to be evaluated in the study. Images labeled as "Portable Chest AP" were retained; however, images labeled as "Chest AP" were set aside because a large percentage of these images were found to be non-chest exam types or QA test exposures that were mislabeled. The relatively high frequency of this particular mislabeling stemmed from the fact that "Chest AP" was configured as the default body part in the CR system. Thus, non-patient images that were acquired on the system and then discarded (rejected) were always labeled as "Chest AP." Also removed were pairs of QA-accepted images of the same patient having the same body part, view position, and accession number, which were captured only a short time apart. It was determined that these images comprised a two-view chest of a large patient. After filtering, the final database contained 10,606 adult chest images: 6,646 portable chest AP (6,613 accepted, 33 rejected), 2,286 chest lateral (2,279 accepted, 7 rejected), and 1,674 chest PA (1,636 accepted, 38 rejected).

Table 2 Reader Study Image Matrix

Category	1	2	3
Clipped anatomy, PA chests	8	25	16
Clipped anatomy, AP portables	21	72	13
Patient motion, laterals	20	10	0
Low exposure	3	9	3
Total	52	116	32

Table 3 RadLex Image Quality Rating Scale

Level	Title	Definition
1	Nondiagnostic	Little or no clinically usable diagnostic information, insufficient information to answer the primary clinical question
2	Limited	Not as much diagnostic information as is typical for an examination of this type, but likely sufficient to answer the primary clinical question
3	Diagnostic	Image quality that would be expected routinely when imaging cooperative patients
4	Exemplary	Image quality that can serve as an example that should be emulated

Three reject-detection algorithms were run on each image in the filtered database: patient-motion detection for lateral chests,[19] clipped-anatomy detection for AP portables and PA chests,[20] and low-exposure detection for all view positions. For motion detection, the lung region is segmented and divided into regions, and each region is assessed for evidence of localized motion. Motion is evaluated based on features extracted from edge images, both vertically and horizontally oriented, which are derived from the original image using directional band-pass filters. Features are extracted for each region within the horizontal and vertical edge images. The features include the moments of the edge values, the standard deviation, the mean deviation, mean local variation, the fraction of edges of magnitude exceeding a noise threshold, and the mean value of those significant edges. The presence of motion is established using these features as input into a support vector machine (SVM)[21] classifier that was previously trained to output a probability of motion blur. For the anatomy-clipping detection algorithm, spine and lung centerlines, and rib contours are first estimated. Leveraging these anatomical landmarks, five regions are established corresponding to the lung apex, the left and right peripheral lung borders, and the left and right costophrenic angles. Geometric and image content features are extracted for each region. The geometric features for the peripheral lung regions include the minimal distance of lung to the image border and the distance of rib edge to the image border. The geometric features for the costophrenic regions include the extent of the image of the anatomy and the distance from the end of lung centerlines

Table 4 Algorithm Detection Rates (Algorithm-Rejects/Total Images in Group) for Various Groups of Images in the Filtered Database

Technologist QA-action	Algorithm detection rate		
	AP portables	PA chests	Lateral chests
Accept	597/6,613 (0.09)	332/1,636 (0.20)	23/2,279 (0.01)
Reject	14/33 (0.42)	27/38 (0.71)	4/7 (0.57)

Table 5 Fraction of Study Images Rated as "Reject" by Each Technologist

Reader session	Category	Images	Technologist										
			1	2	3	4	5	6	7	8	9	10	Avg.
1	1	52	0.02	0.00	0.02	0.04	0.00	0.02	0.02	0.00	0.00	0.00	0.01
2	1	52	0.04	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
1	2	116	0.51	0.25	0.41	0.31	0.13	0.28	0.11	0.24	0.16	0.15	0.26
2	2	116	0.57	0.32	0.37	0.41	0.23	0.28	0.14	0.30	0.05	0.11	0.28
1	3	32	1.00	0.88	0.97	0.97	0.72	0.97	0.77	0.97	0.80	0.80	0.89
2	3	32	1.00	1.00	0.97	0.97	0.68	0.97	0.88	0.97	0.47	0.81	0.87

to the bottom edge of the image. Image content features for all regions include the average and minimal intensity values, and the total direct exposure area. These features are input into another SVM classifier to determine a probability of clipping for each region. For both motion and clipped anatomy, an image is flagged as a reject if any one of the region probabilities for that image exceeds a predefined probability value. The threshold for low-exposure detection was established by the senior radiographic technologists at CSMC to be 0.20 mR for an average exposure to the CR plate within the anatomical region. The 0.20-mR threshold corresponds to the CR vendor's specific exposure index of about 1,300 [22].

After the reject-detection algorithms were applied to the images in the filtered database, the images were partitioned into two classes according to the algorithm output. An image was classified as an algorithm-reject if the algorithms flagged the image for rejection; otherwise, the image was classified as an algorithm-accept.

Each image was subsequently binned into one of three categories: (1) QA-accepted by technologist and an algorithm-accept, (2) QA-accepted by technologist but an algorithm-reject, and (3) QA-rejected by technologist and an algorithm-reject. Images that fell into a fourth category of QA-rejected by the technologist but algorithm-accepted were not included. Only a small subset of cases (33 images) fell into this category. Upon inspection, it was discovered that most images in this category suffered from such severe degradation that it caused the reject-detection algorithms to generate erroneous output. In an operational scenario, we felt that it would be highly improbable that any of these images would be QA-accepted.

Table 6 Fraction of Study Images Rated as "Reject" by Each Radiologist

Category	Number images	Radiologist					
		1	2	3	4	5	Avg.
1	52	0.00	0.02	0.02	0.02	0.00	0.01
2	116	0.03	0.19	0.07	0.17	0.07	0.11
3	32	0.23	0.94	0.63	0.81	0.53	0.63

From the filtered database, a sequential sampling scheme was used to attempt to populate the reader study database with one half of the reader study images coming from categories 1 and 3 and the remaining images coming from category 2. A total of 200 images were selected for use in a reader study based on the categorizations described above. In actuality, the final breakdown was dictated by the composition of the filtered database and availability of images among the respective categories. Table 2 shows the matrix of images selected to represent each category for the reader study.

Ten technologists each evaluated all 200 images twice, in two independent reading sessions. The sessions were separated in time by 45 days to limit the influence of memory effects. For the first reader session, images were presented to each technologist on a 1,280 × 1,024 resolution, 19-inch flat panel color display (Elo TouchSystems, Berwyn, PA, USA). The technologist reading sessions were conducted with ambient lighting conditions consistent with the environment under which QA is performed at CSMC. The graphical user interface on the study software was designed to mimic the interface on the CR system with which the technologist was familiar. The exposure indicator, body part, view position, and reason for the exam were also presented with the image. The technologists were allowed to perform basic image manipulation functions, as required, such as window width and level adjustments, zoom, and pan. Image display order was independently randomized for each reader.

For each image, the technologists provided an image quality rating based on the RadLex [23] scale and corresponding definitions (Table 3). The ratings were recorded electronically via a separate window on the display interface. For any image rated as Nondiagnostic (RadLex level 1), the image was automatically categorized as a reject, and the technologists were prompted to indicate a reason for the rejection based on the criteria shown in Table 1. If the technologists scored an image as Limited (RadLex level 2), they were prompted to further indicate whether they would accept or reject the image; if

Table 7 Cohen's Kappa Coefficients for the Technologist Reader Group (First Reader Session)

Technologist	1	2	3	4	5	6	7	8	9	10
1	–	0.51	0.61	0.57	0.29	0.58	0.38	0.47	0.38	0.33
2	0.51	–	0.61	0.66	0.48	0.58	0.45	0.74	0.50	0.59
3	0.61	0.61	–	0.69	0.40	0.66	0.43	0.61	0.47	0.43
4	0.57	0.66	0.69	–	0.48	0.66	0.50	0.62	0.60	0.58
5	0.29	0.48	0.40	0.48	–	0.44	0.37	0.57	0.55	0.53
6	0.58	0.58	0.66	0.66	0.44	–	0.45	0.65	0.55	0.46
7	0.38	0.45	0.43	0.50	0.37	0.45	–	0.55	0.50	0.59
8	0.47	0.74	0.61	0.62	0.57	0.65	0.55	–	0.65	0.67
9	0.38	0.50	0.47	0.60	0.55	0.55	0.50	0.65	–	0.62
10	0.33	0.59	0.43	0.58	0.53	0.46	0.59	0.67	0.62	–

rejection was indicated, the technologists were again prompted to enter a reason. Any image rated as either Diagnostic (RadLex level 3) or Exemplary (RadLex level 4) was automatically categorized as an accepted image. For the second session, technologists performed the same procedure for rating images as in the first session; but in the second session, the technologists were also provided with the results from the reject-detection algorithms. The presentation sequence was randomized independently from the first session, so each reader viewed the images in an order different from the first session.

Separately, five radiologists each provided RadLex ratings and reject reasons, as required, for the same set of 200 images. For the radiologist reader session, the images were displayed on a 1,536×2,048 resolution diagnostic grayscale display (National Display Systems, Inc., Morgan Hill, CA, USA) under reduced ambient lighting consistent with the environment in the radiology reading rooms at CSMC. The sequence of image presentation was again randomized independently for each of the readers. The radiologist readers were instructed to rate each image according to what they would have directed the technologist to do under an operational scenario. The radiologists' RadLex ratings were categorized as reject or accept using the same method used for categorizing the technologists' ratings.

Distributions of accept and reject responses were generated for each reader, reading session, reject reason, and image category. Agreement among technologists for each reading session, among radiologists, and between technologists and radiologists was quantified using Cohen's kappa coefficient [24]. Changes in agreement among technologists between reader sessions 1 and 2 were computed. Proportion analyses were also performed for each technologist reader, comparing each reader's reject decisions between the first and the second session.

Algorithm performance was quantified using precision analysis,[25] where truth was established by a consensus of the technologist readers. For each of the 200 images used in the study, a label of (technologist) consensus reject was assigned if the image was rated by at least one-half of the ten technologist readers as a reject. Precision (p), which is the fraction of algorithm-rejects that were also consensus rejects, was calculated for each combination of view position and QA action, where the QA action was represented by the accept or reject

Table 8 Change in Cohen's Kappa Coefficients Between the First and Second Technologist Sessions

Technologist	1	2	3	4	5	6	7	8	9	10	Avg.	*P*
1	–	−0.09	0.08	−0.11	−0.14	−0.02	−0.01	−0.06	0.20	0.00	−0.02	0.65
2	−0.09	–	0.00	−0.12	−0.11	−0.10	−0.08	0.01	0.16	0.05	−0.03	0.35
3	0.08	0.00	–	0.06	−0.09	0.11	−0.09	−0.06	0.23	0.00	0.03	0.47
4	−0.11	−0.12	0.06	–	−0.03	0.00	0.02	−0.06	0.36	0.07	0.02	0.67
5	−0.14	−0.11	−0.09	−0.03	–	−0.15	−0.01	−0.06	0.15	0.04	−0.04	0.20
6	−0.02	−0.10	0.11	0.00	−0.15	–	−0.11	−0.02	0.21	−0.10	−0.02	0.62
7	−0.01	−0.08	−0.09	0.02	−0.01	−0.11	–	0.05	0.08	−0.07	−0.02	0.30
8	−0.06	0.01	−0.06	−0.06	−0.06	−0.02	0.05	–	0.34	0.12	0.03	0.53
9	0.20	0.16	0.23	0.36	0.15	0.21	0.08	0.34	–	0.17	0.21[a]	<0.001
10	0.00	0.05	0.00	0.07	0.04	−0.10	−0.07	0.12	0.17	–	0.03	0.31

Differences were not detectable for nine of the ten technologists. Technologist 9 showed significantly lower kappa coefficients based on the responses from session 2 as compared with responses from session 1.

[a] Statistically significant

decision made by the technologist at the time the image was captured. Recall analysis was not performed because the selection of images chosen for the study omitted those images that were QA-rejected but algorithm-accepted. Leaving out such images would result in an inflated estimate of recall value that would provide little insight when attempting to generalize to the larger population.

Results

Algorithm-detection rates for images in the filtered database are shown in Table 4.

Detection rates are shown separately for images that were QA-accepted and QA-rejected and for each view position. The reader study results are summarized for each technologist in Table 5. The values in the table represent the fraction of images within each category that were rated as reject by each technologist and are reported for each of the two reader session. A summary of the radiologist results is shown in Table 6.

Values for Cohen's kappa coefficient (i.e., a measure of agreement) among technologists in the reader session are shown in Table 7.

Changes in coefficient values between the first and second technologist reader sessions are shown in Table 8. Each row of data in Table 8 was averaged, and the resulting mean was tested against zero using a *t*-statistic.

Table 9 shows values for Cohen's kappa coefficient for the radiologist group. Inter-reader agreement among technologists and radiologists is shown in Table 10.

The results from the test comparing the proportion of the 200 images that were rated as reject between the first and second reading sessions are shown for each technologist in Table 11. A positive value for the difference indicates the technologist rejected a greater number of study images in the second session when compared with the number of images rejected during the first session. A negative value indicates the technologist rejected a greater number of images in the first session when compared with the number of images in the second session.

Algorithm performance was quantified in terms of precision (Table 12). Based on the 200 images used in the reader study, precision values were calculated for each view position and relevant image category. To understand how the calculations were performed, consider the following example. Twenty-five PA chests were included in the reader study from category 2 (from Table 2). As was previously described, category 2 represents images that were QA-accepted but algorithm-rejected. When evaluated by the ten technologists in the reader study, 5 of these 25 images (20 %) were rated by one-half or more of the technologists as a reject, i.e., 5 images were consensus rejects. Thus, algorithm precision, when run against the population of PA chests that are QA-accepted by the technologist at the time of capture, is estimated to be 0.20. This figure of merit can be interpreted to mean that 20 % of PA chest images that are accepted by the technologist at the time of image capture, but which are flagged by the detection algorithms as a reject, would be rejected by a majority of independent technologists. Note that precision cannot be calculated for category 1 images because none of these images were rejected by the algorithm.

Table 9 Cohen's Kappa Coefficients for the Radiologist Reader Group

Radiologist	1	2	3	4	5
1	–	0.23	0.39	0.29	0.49
2	0.23	–	0.49	0.59	0.43
3	0.39	0.49	–	0.54	0.44
4	0.29	0.59	0.54	–	0.60
5	0.49	0.43	0.44	0.60	–

Table 10 Cohen's Kappa Coefficients Between Radiologist and Technologist Reader Groups

Technologist	Radiologist				
	1	2	3	4	5
1	0.09	0.44	0.19	0.40	0.23
2	0.18	0.60	0.35	0.40	0.37
3	0.12	0.52	0.29	0.39	0.21
4	0.17	0.63	0.41	0.54	0.31
5	0.37	0.50	0.53	0.52	0.37
6	0.14	0.59	0.36	0.51	0.29
7	0.21	0.51	0.49	0.58	0.37
8	0.19	0.58	0.49	0.56	0.32
9	0.29	0.52	0.48	0.58	0.44
10	0.28	0.54	0.52	0.49	0.45

Discussion

The reject-detection algorithms flagged a surprisingly large number of QA-accepted images from the filtered database (chest PA, 20 %; AP portables, 9 %; laterals, 1 %). The magnitude of these percentages motivated the preferential selection of images from category 2 for inclusion in the reader study. More than half (116) of the 200 reader study images were selected from this category. For each of the three reader sessions, category 1 images were consistently rated as acceptable (0.99 accepted). Category 3 images were consistently rejected by technologists during each of the two reader sessions (0.88 rejected). The radiologist reader group also rejected the majority of category 3 images, although at

Table 11 Comparison of the Proportion of Images Rated as "Reject" by Technologists Between Reader Sessions 1 and 2

Technologist	1	2	3	4	5	6	7	8	9	10
Proportion rejected (first session)	0.47	0.29	0.41	0.35	0.19	0.34	0.20	0.30	0.22	0.21
Proportion rejected (second session)	0.50	0.35	0.39	0.39	0.24	0.31	0.22	0.33	0.10	0.20
Difference	−0.03	−0.06	−0.02	−0.04	−0.05	0.03	−0.02	−0.03	0.12[a]	0.01
P value	0.62	0.23	0.71	0.38	0.26	0.58	0.69	0.50	<0.01[a]	0.72

[a] Statistically significant difference

a lesser percentage than the technologists (0.63 rejected). The images from category 2 generated variable responses from technologists during both reader sessions. Among the ten technologists that participated in the reader study, the fraction of images that individual readers rejected ranged between 0.11 and 0.51 (average 0.26) during the first session and between 0.05 and 0.57 (average 0.28) during the second session. The radiologists also provided mixed responses for images in this category, rejecting between 0.03 and 0.19 of the 116 images (average, 0.11). As a group, the reject rate for the radiologists was less than half that of the technologists for category 2 images, and was 29 % less for category 3 images. We hypothesize that the on-average lower reject rates for the radiologist group when compared with the technologist group stems from differences in the intrinsic criteria used when making reject decisions. The radiologists seemingly weighted their decisions more heavily on whether image quality was sufficient for the intended diagnostic purpose, while the technologists weighted their reject decisions more heavily on objective technical quality factors. This assertion is supported by observations of the types of images where the radiologists and technologists disagreed. For example, the technologists rejected, on average, 10 of the 12 low-exposure images from categories 2 and 3, while the radiologists, on-average, accepted all 12. A key factor used by technologists in deciding whether to reject an image for low exposure is the exposure index, which is an objective technical criterion for quality. Although a value for exposure index may be set as a threshold and used to make binary decisions of accept or reject, the interpretability of images generally decreases gracefully as the signal-to-noise ratio, and thus exposure index, is reduced. As such, radiologists may perceive diagnostic interpretability as sufficient for images having exposure index values falling below the threshold for some indications. To highlight another example, approximately 12 % of the images from category 2 were rejected by the technologist reader group for clipped anatomy but were accepted by the radiologist group. It is evident by inspection of these images that while clipping of the lung field exists, the degree of lung-field clipping is slight (Fig. 3a, b). The presence of clipping in these images is corroborated by the reject-detection algorithms that independently identified the quality deficiency from an objective technical perspective. Given that the lung fields were nearly fully imaged, the diagnostic quality of these images for the respective indications was seemingly judged to be sufficient by the radiologists, even if quality was limited from an objective technical perspective.

Table 12 Algorithm Precision (*p*)

Category	QA-action	AP portables	PA chests	Lateral chests
2	Accept	0.14	0.20	0.68
3	Reject	0.92	1.00	1.00

Reader agreement was assessed using Cohen's kappa coefficient. Coefficients were calculated separately for the technologist reader group for the first and second reading sessions, for the radiologist reader group, and between the radiologist and technologist reader groups. Using the interpretation of kappa shown in Table 13,[26] the technologist inter-reader agreement data shown in Table 7 ranged from fair to substantial, with 4 instances of fair, 29 of moderate, and 12 of substantial. Technologist inter-reader agreement was essentially unchanged in the second session, with the exception of technologist 9. Technologist 9 agreed less ($p<0.01$) with the rest of the technologists in the second session, with the on-average kappa coefficient changing from 0.54 (moderate) to 0.32 (fair). The proportion of images that were rejected (Table 11) by this technologist increased ($p<0.01$) in session 2 versus in session 1, while the proportions of rejected images for the other technologists were essentially unchanged between the two sessions. Inter-reader agreement among radiologists as a group (Table 9) was comparable to that of the technologists as a group. The instances of kappa coefficients ranged from fair (2 instances) to moderate (8 instances). Agreement between the radiologist and technologist reader groups (Table 10) was lower than the agreement within either the technologist or radiologist reader groups, with nearly half the kappa coefficients falling within the slight (7 instances) to fair (16 instances) interpretation ranges and the remaining coefficients falling within the moderate interpretation range (26 instances). The lower inter-reader agreement between groups is consistent with the inter-group reject rate differences previously discussed.

Precision was used as a figure of merit to assess algorithm performance. Precision was found to be nearly 100 %

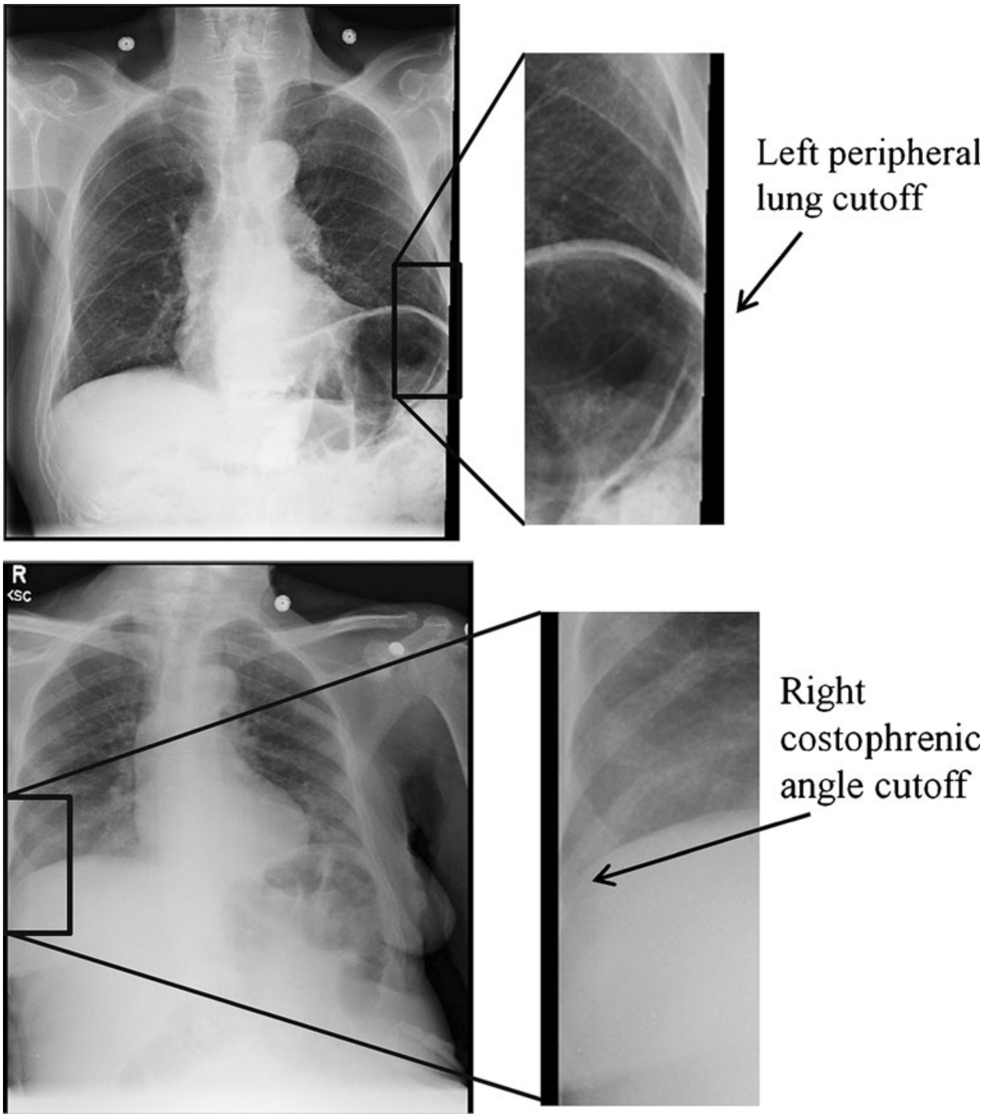

Fig. 3 Examples of clipped anatomy. For both examples, the technologists rejected the images, seemingly based on technical quality considerations. The radiologist reader group, on average, accepted both of these images, seemingly basing their decisions on the diagnostic sufficiency of the images for the intended purpose

for category 3 images (the QA-rejected images that were also algorithm-rejected), that is, these images were almost always consensus rejects as judged by the independent peer group. Of greater interest is the algorithm precision performance for the category 2 imagery (the QA-accepted images that were algorithm-rejected). The values shown in Table 12 for these images represent the fraction of the total number of algorithm-rejects that were also consensus rejects. For this image category, precision provides an indication of the potential QA benefit from introducing the algorithms into a clinical imaging scenario. For instance, suppose the algorithms were run on all chest images collected at an imaging center, and algorithm-detected rejects were automatically routed to a QA technologist for independent assessment. Under this scenario, the precision performance suggests that the QA technologist would find that 20 % of the PA chest cases that were accepted by the technologist that captured the image should have been rejected. This information could potentially be used as feedback to the technologist shortly afterwards and allow for the image to be repeated, or alternatively, to be compiled as part of a centralized reject-tracking database and used for retrospective QA.

Table 13 Interpretation of Kappa

Kappa value	Interpretation of kappa
<0	Less than chance agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

When the algorithm-detection rate (Table 4) is multiplied with the precision value, the result is an estimate of the fraction of images acquired at the site over the 14-month time period that would have been rejected had they been reviewed by a technologist other than the technologist that originally captured the image. These fractional values, calculated for all the view positions and then converted into percentages, were 4.06 % for PA chests, 1.25 % for AP portables, and 0.68 % for lateral chests. The values, although relatively small, are still potentially significant because of the large volume of chest X-ray imaging that is typically performed at a large hospital.

A limitation for this study was that algorithms were developed and evaluated for only a subset of reject types and for a subset of chest exams. Given this limitation, the results still provide an indication of how this type of technology can be leveraged as part of an overall QA program.

Conclusions

Radiographic technologists agreed only moderately in their assessments of image quality deficiencies. This leads to an intrinsic variability in reject rates among technologists and, further, leads to variability in the quality of images delivered to the PACS. When compared against each other, radiologist and technologist groups were found to have less agreement than the inter-reader agreement within each group. Radiologists were found to be more accepting of limited quality studies than technologists. Evidence from this study suggests that technologists weigh their reject decisions more heavily on objective technical attributes, while radiologists weigh their decisions more heavily on diagnostic interpretability relative to the image indication. Objective technical criteria tend to be more stringent to satisfy, which explains, in part, why the technologist reject rates were found to be consistently higher than that of the radiologists. Having the reject-detection algorithm results available to the technologist did not improve inter-reader agreement in terms of the technologist's decisions about whether to accept or reject. However, if the algorithms were optimized based on the opinion of the radiologists, the technologist might be able to better utilize the software to improve consistency, and they could potentially reduce repeats by not accepting cases that were rejected by the algorithm and by having the option to reject an image that is accepted by the algorithm. Over time, the algorithms could be refined with information learned from radiologists' review, and when the algorithms were optimized sufficiently to be in high correlation with the radiologists' opinions, the software could be introduced into the operational environment. The algorithms were shown to detect a small percentage of QA-accepted images that should have been rejected, and thus, the algorithms do provide information that could be captured within a reject-tracking database and leveraged as part of a site-wide QA program.

References

1. IMV Medical Information Division: 2010 X-ray/DR/CR market outlook report. November 2010
2. Foos DH, Sehnert WJ, Reiner B, Siegel EL, Segal A, Waldman DL: Digital radiography reject analysis: data collection methodology, results, and recommendations from an in-depth investigation at two hospitals. J Digit Imaging 22:89–98, 2009
3. Jones AK, Polman R, Willis CE, Shepard SJ: One year's results from a server-based system for performing reject analysis and exposure analysis in computed radiography. J Digit Imaging 24(2):243–255, 2011
4. Tzeng W, Kuo K, Liu C, Yao H, Chen C, Lin H: Managing repeated digital radiography images—a systematic approach and improvement. J Med Syst 36(4):2697–2704, 2012. doi:10.1007/s10916-011-9744-8
5. Reiner B: Automating quality assurance for digital radiography. J Am Coll Radiol 6:486–490, 2009
6. Reiner B, Siegel E, Sehnert WJ: Beauty in the Eyes of the Beholder: Image Quality Perceptions for Digital Radiography. Presented at: Society of Imaging Informatics in Medicine, Seattle, May 17, 2008
7. Shet N, Chen J, Siegel E: Continuing challenges in defining image quality. Pediatr Radiol 41(5):582–587, 2011
8. Reiner BI, Siegel E, Foos D: Quantitative analysis of digital radiography QA deficiencies. In: Abstracts of the Annual Meeting of the Society for Imaging Informatics in Medicine. Providence, 2007
9. Waaler D, Hofmann B: Image rejects/retakes—radiographic challenges. Radiat Prot Dosim 139(1–3):375–379, 2010
10. Nagy PG, Pierce B, Otto M, Safdar NM: Quality control management and communication between radiologists and technologists. J Am Coll Radiol 5(6):759–765, 2008
11. Minnigh TR, Gallet J: Maintaining quality control using a radiological digital X-ray dashboard. J Digit Imaging 22(1):84–88, 2009
12. Prieto C, Vano E, Ten JI, Fernandez JM, Iñiguez AI, Arevalo N, Litcheva A, Crespo E, Floriano A, Martinez D: Image retake analysis in digital radiography using DICOM header information. J Digit Imaging 22(4):393–399, 2009
13. Meyer-Base A: Pattern Recognition for Medical Imaging: Application of Statistical Classification Methods in Biomedical Imaging. Elsevier Academic, San Diego, 2004
14. Zhang P, Varma B, Kumar K: Neural vs. statistical classifier in conjunction with genetic algorithm feature selection in digital mammography. The 2003 Congress on Evolutionary Computation 2:1206–1213, doi:10.1109/CEC.2003.1299806, December 8–12, 2003
15. Castro A, Boveda C, Arcay B: Comparison of various fuzzy clustering algorithms in the detection of ROI in lung CT and a modified kernelized-spatial fuzzy c-means algorithm. 2010 10th IEEE International Conference on Information Technology and Applications in Biomedicine, 1–4, doi:10.1109/ITAB.2010.5687726, November 3–5, 2010
16. Rale AP, Gharpure DC, Ravindran VR: Comparison of different ANN techniques for automatic defect detection in X-ray images. Int Conf Emerg Trends Electron Photon Devices Syst 2009:193–197, 2009
17. Shiraishi J, Li Q, Appelbaum D: Computer-aided diagnosis and artificial intelligence in clinical imaging. Semin Nucl Med 41 (6):449–462, 2011
18. Dayhoff JE, DeLeo JM: Artificial neural networks: opening the black box. Cancer 91:1615–1635, 2001
19. Luo H, Sehnert WJ, Ellinwood JS: Method for Detecting Anatomical Motion Blur in Diagnostic Images. Patent US7899229B2, March 1, 2011
20. Luo H: Method for Detecting Clipped Anatomy in Medical Images. Patent US7912263B2, March 22, 2011
21. Hamel L: Knowledge Discovery with Support Vector Machines. Wiley, Hoboken, 2009

22. Gallet J: The Concept of Exposure Index for Carestream Direct-view Systems. Carestream Technical Brief Series. CAT No. 120 7091, 2010
23. Radiological Society of North America: RadLex: A Lexicon for Uniform Indexing and Retrieval of Radiology Information Resources. http://www.rsna.org/radlex/. Accessed 09 April 2010
24. Cohen J: A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46, 1960
25. Birger H: The foundation of the concept of relevance. J Am Soc Inf Sci 61(2):217–237, 2010
26. Landis JR, Koch GG: The measurement of observer agreement for categorical data. Biometrics 33:159–174, 1977