Subjective Quality Assessment of Computed Radiography Hand Images

Cynthia A. Britton, Orlando F. Gabriele, Thomas S. Chang, Jeffrey D. Towers, David A. Rubin, Walter F. Good, and David Gur

To evaluate the sensitivity of a non-receiver-operating characteristic (ROC) study in assessing small differences of perceived image quality of hand images acquired by computed radiography (CR) and conventional screen-film systems, hand images were acquired on 12 patients with both conventional screen-film and CR. Each CR image was then processed with three different edge-enhancement algorithms. One conventional film and four CR images were then viewed side by side by five radiologists. Observers rated perceived image quality of each radiograph using a 10-category discrete scale. The study was repeated after 6 weeks using a different block randomization scheme. Despite the small sample size, significant differences ($P < .05$) in assigned image quality were detected among CR images acquired at low, medium, and high resolutions. Image processing routines did not fully compensate for differences in quality between conventional film and CR-acquired images. The quality rating of the reference conventional image was found to be dependent on the quality of images with which it was compared. Small, highly sensitive study designs can be used to identify radiologists' perceived differences in image quality. "Reference" or "gold standard" quality are important in such studies. Edge-enhancement schemes cannot fully compensate for perceived image quality degradations because of reduced image resolution.

KEY WORDS: imaging, image quality, subjective assessment, computed radiography (CR).

SUBJECTIVE assessments of image quality have long been used in diagnostic imaging to establish observer preferences.[1] These highly sensitive, albeit subjective and qualitative, techniques are quite important because many times they correlate with actual observer performance, and, perhaps as important, user acceptance of an imaging modality is an important aspect of the clinical practice.[2] Whether side-by-side reviews or independent observations are used for this purpose, the observers' reference point or "gold standard" is an important parameter that cannot be ignored.[2,3] This is particularly true when small differences in quality or observer performance may exist among the compared imaging modalities.[4] If indeed a strong correlation exists between actual performance as measured by receiver-operating characteristic (ROC) studies and subjective assessments of quality, the latter approach is not only a "beauty" contest, but could potentially be used to optimize evaluation protocols.

Although used clinically for extremity imaging in various environments, computer radiography (CR) has not been universally accepted for this purpose.[5-7] An important question related to the assessment of image quality is whether or not image processing (eg, edge enhancement) can compensate (either partially or fully) for degradation of image resolution in images that potentially contain important diagnostic high-frequency information (eg, extremities). Image quality degradations through resolution reduction may be the method of choice in many procedures when dose reduction is desired.

Before performing an objective observer performance study and to assess radiologists' subjective assessment of and preference for specific types of processed CR images, we performed the following multiimage side-by-side reviews.

MATERIALS AND METHODS

During a 1-month period, 12 patients undergoing a hand examination (R/O fracture, foreign body) in our Emergency Department were asked to participate in this study under an Institutional Review Board-approved protocol. Conventional exposures in this setting are routinely performed using a LANEX Regular Screen with TMG film (Eastman Kodak, Rochester, NY). After a 400-speed conventional image was acquired on each patient, a CR image was also acquired within 2 minutes, using a cassette that contained either a high-, medium-, or low-resolution plate. The resolution of the screen to be used on a particular patient was randomly predetermined. CR exposures were performed using the same view and exposure parameters (ie, skin to detector distance, kVp) as the conventional image, with the exception of speed (mAs), which was

From the Department of Radiology, University of Pittsburgh, Pittsburgh, PA.

Supported in part by Grant No. CA66594 from the National Cancer Institute, National Institutes of Health.

Address reprint requests to Cynthia A. Britton, MD, A449 Scaife Hall, Department of Radiology, University of Pittsburgh, Pittsburgh, PA 15261-0001.

0897-1889/96/0901-0002$3.00/0

Table 1. Example of Individual Ordinal Ratings of Quality by One of the Raters

Case No.	Conventional Film Screen	CR Images				
				Edge Enhancement		
		Resolution	Original	Low	Medium	High
01	4	High	8	7	4	7
02	10	High	6	8	6	4
03	8	Low	5	7	3	3
04	10	Low	7	5	3	4
05	8	Medium	4	7	3	4
06	10	Low	3	4	4	5
07	7	Medium	6	7	7	6
08	10	Medium	8	4	6	4
09	7	High	10	7	7	4
10	6	Low	3	3	3	5
11	3	Medium	9	5	6	6
12	4	High	5	7	4	6

Data taken from the first experiment.

fluence-adjusted to compensate for the relative speed differences of the plates. Hence, the CR scanner's signal level was comparable for all plates. Relative exposures were set at 1:1:2:3 for the conventional and the low-, medium-, and high-resolution CR images, respectively. The study group of 12 patients thus yielded four CR images at each of three resolutions. These were then compared with conventional images during the subjective side-by-side quality ratings' portion of the study. Had we acquired all images at the same or comparable radiation dose, the noise in the high-resolution CR images would be quite high, to a point where radiologists would likely object to using such "noisy" images clinically. In addition, electronic systems' noise was virtually eliminated as a dependent variable in our experiment.

All CR images were acquired on 18 × 24 cm^2 plates and were scanned with a CR scanner (KODAK EKTASCAN Storage Phosphor Reader; Eastman Kodak) with a matrix of 1,792 × 2,392. Images were then processed with an unsharp-masking routine using a low enhancement factor (0.2 to 0.4) and three different kernel sizes of 75, 37, and 17 pixels. This process resulted in minimal to moderate edge enhancement of the CR images that increased with decreased kernel size. After lookup table adjustments, the original (nonunsharp masked) image and the three edge-enhanced images were laser-printed onto film using a laser printer (KODAK EKTASCAN Laser Printer; Eastman Kodak). Image sizes were the same as on the conventional films. System modulation transfer function of the CR images was measured at 0.27, 0.40, and 0.49 at a frequency of 2.0 lp/mm for the low-, medium-, and high-resolution plates, respectively. At 3.5 lp/mm, it was 0.11, 0.18, and 0.25 for the low-, medium-, and high-resolution plates, respectively. As a result of our acquisition and processing protocol for each case, 5 films were generated (1 conventional, 1 original [nonunsharp masked CR image], and 3 unsharp masked CR images). The 60 films (5 images/case × 12 cases) were then assigned random serial numbers for identification purposes during the subjective quality assessment experiment.

Two subjective rating experiments were performed during this project. In each, five experienced radiologists were presented with 12 sets of five images that were displayed side by side on view boxes. The following are excerpts from the "Instructions to Observers" that were provided to them:

> "You are presented with sets of images to be subjectively rated on a relative and an absolute scale as to their image quality. Images in each set were acquired at different resolutions and exposure factors and were processed using a variety of parameters. Therefore, all we want you to do is to assess their "quality for primary diagnosis." For each image, you should rate "image quality" on a scale of 0-10 (10 = superb quality for primary diagnosis; 5 = acceptable quality [barely]; and 0 = totally unacceptable quality). Remember that these sets are presented side by side for reference only. Several of the images in one set could be rated at the same quality level. Before rating each set, please shuffle the images around on different viewing boxes."

Five board-certified radiologists who routinely diagnose bone images participated in this study, and their image quality ratings were entered into a computer data base designed specifically for this purpose. All of them had used CR images to varying extents in general (eg, chest), and in particular for CR images of extremities. In the first experiment, one image from each acquisition and processing mode (but not of the same patient) was included in each set. Block randomization in this intercase study assured that each image was seen only once and that each set contained all resolutions and processing modes possible. During the second experiment, which was performed 6 to 8 weeks after the completion of the first, observers were presented side by side with both acquisition options (ie, conventional and CR) and all processing options of the same patient. In the former intercase study an optimal randomization was achieved, but no direct side-by-side comparison of different images from the same patient was possible. In the latter intracase study, specific features could be directly compared by the rater.

Table 2. Average Quality Ratings (and standard deviations) by Resolution, Acquisition Type, and Filtering for One Reader in the First Experiment

Conventional Film Screen	CR Images				
			Edge Enhancement		
	Resolution	Original	Low	Medium	High
6.25 (2.87)	High (n = 4)	7.25 (2.22)	7.25 (0.50)	5.25 (1.50)	5.25 (1.50)
7.00 (2.94)	Medium (n = 4)	6.75 (2.22)	5.75 (1.50)	5.50 (1.73)	5.00 (1.15)
8.50 (1.91)	Low (n = 4)	4.50 (1.91)	4.75 (1.71)	3.25 (0.50)	4.25 (0.96)

Standard deviation appears in parentheses.

Table 3. Average Scores for All Readers in the First Experiment

Conventional Film Screen	CR Images				
			Edge Enhancement		
	Resolution	Original	Low	Medium	High
6.65 (1.49)	High (n = 4)	6.80 (1.19)	7.05 (0.74)	6.80 (1.08)	6.50 (1.43)
6.70 (1.45)	Medium (n = 4)	6.20 (1.36)	6.55 (0.91)	6.15 (1.08)	6.35 (1.28)
7.50 (1.36)	Low (n = 4)	5.15 (1.14)	5.60 (0.98)	4.60 (1.47)	5.60 (1.43)

Standard deviation appears in parentheses.

RESULTS

Table 1 is an example of the individual ratings for one reader in the first intercase comparison experiment. This reader's summary of average ratings by mode (processing routines) and acquisition resolution is shown in Table 2. The intrareader variability of individual scores within and between cases is noted. From this table, it can be seen that on the average this radiologist rated the low resolution CR images lower than the medium and high resolution images for all CR processing modes. In addition, in general this reader did not particularly like edge enhancement, and his/her quality ratings of the "reference" conventional film-screen images were affected by the comparison images. Namely, the lower the quality of the CR images that were viewed side by side, the higher the assigned or perceived quality of the conventional images.

Table 3 summarizes the average scores over all readers and images for the first intercase experiment, classified by modality (ie, acquisition type), degree of image processing, and resolution. Note that although the actual resolution for the conventional film-screen images is the same for all 12 cases, the ratings are affected by the actual quality of the CR images being viewed side by side. Each average is obtained from the 20 scores in the same group (four images rated by five readers). Page's test[8] for an increasing trend from low to high resolution results in a statistically significant trend at $P < .001$. Furthermore, when this same test is applied to average scores for each individual reader, the trend is significant ($P < .05$) for three readers and is suggestive of a trend ($.05 < P < .10$) for a fourth reader. The average assigned scores for the conventional film-screen images, when compared with CR images with different resolutions, decreases as the quality of the CR images increases. This opposite trend, which is summarized in Table 4, has a two-sided significance level of $P < .08$.

Table 5 is a summary of the average scores over all readers and images classified by type of image, resolution, and degree of image processing in the second intracase experiment. In this experiment, the reference quality (ie, conventional image) and all observations regarding the low-resolution CR images during the first experiment were virtually the same. CR quality scores for medium- and high-resolution images were comparable, largely because of two readers whose ratings of the medium-resolution CR images were equal to or higher than those of the high-resolution images. Nevertheless, the trend of higher scores with higher resolutions persisted ($P < .05$).

DISCUSSION

Several observations can be made from the results of this study. As a group, the unprocessed high-resolution CR images acquired at higher radiation exposure (three times that of the conventional film) were rated to be of comparable image quality to those acquired with 400-speed conventional double screen-film technology.

In the first experiment, the group as a whole and several individual readers clearly selected ($P < .05$) the high-resolution images as having better image quality than medium-resolution

Table 4. Average Score by Resolution and Reader for Conventional Film-Screen Images When Compared Side by Side With CR Images at a Given Resolution

Resolution of Comparison CR Images	Average Conventional Film-Screen Score by Reader Number				
	1	2	3	4	5
High	6.25	5.0	5.75	8.75	7.5
Medium	7.0	5.25	5.75	9.0	6.5
Low	8.5	6.5	6.0	9.25	7.25

Table 5. Average Scores Over All Readers and Images for the Second Experiment

Conventional Film Screen	CR Images				
			Edge Enhancement		
	Resolution	Original	Low	Medium	High
7.05 (1.54)	High (n = 4)	7.00 (0.95)	6.70 (0.57)	6.85 (1.58)	7.05 (1.36)
7.10 (1.61)	Medium (n = 4)	6.85 (0.98)	7.05 (0.27)	6.80 (1.80)	7.45 (1.39)
7.90 (1.21)	Low (n = 4)	6.00 (0.77)	6.30 (0.67)	5.25 (0.98)	6.30 (0.89)

Standard deviation appears in parenthesis.

images, and the latter group to be of better quality than the low-resolution images. The sensitivity of detecting these quality differences is high enough to enable statistically significant determinations with a small number of cases (12) and readers (5). In the second experiment, when side-by-side comparisons of the same case were made, the low-resolution images were rated lower ($P < .05$) than either the high- or medium-resolution images.

There was no significant or consistent trend in perceived quality among the different types of CR image processing. In other words, the processing modes used in this study did not compensate for differences in perceived image quality between conventional and CR images or differences in CR resolutions during acquisition.

Perceived quality of the reference conventional image was affected by the quality of the image with which it was compared in side-by-side reviews. As CR quality decreased, the perceived quality of the conventional image ("gold standard") increased, while, in reality, it was constant. The issue of using an appropriate gold standard (reference point) in this type of study has been discussed previously,[3] and it may be significantly underestimated as to its effect on the subjective assessment of quality. Had we used the difference in "quality" between the gold standard (ie, the conventional images) used in the side-by-side review and the "quality" of CR images in each group, the observations made in this study are even stronger. The latter index, namely the difference, is perhaps a more appropriate measure of relative quality for such studies.

In light of the highly consistent results of the first experiment (intercase comparison), we were somewhat surprised with the differences in the second experiment (intracase comparison), where some readers rated the medium-resolution images to be of comparable quality to high-resolution images when rating images of the same patient. To date, we have no conclusive explanation for these differences.

We wish to emphasize that despite the many "statistically significant observations" that could be made from this study, the most important aspect of this effort is the ability to derive such observations from images whose visual differences are very small using an extremely small number of cases and readers. This approach, which proved to be highly sensitive in this and similar studies,[2] should be further explored as a potential pilot effort for determining whether to perform an ROC study and what type of images should be used for this purpose.

REFERENCES

1. Fuhrman CR, Gur D, Good BC, et al: The diagnostic quality of storage phosphor radiographs compared to conventional films: Interpreters' perception. Am J Roentgenol 150:1011-1014, 1988

2. Good WF, Gur D, Feist JH, et al: Subjective and objective assessment of image quality—A comparison. J Digit Imaging 7:77-78, 1994

3. Gur D, Fuhrman CR, Feist JH, et al: Spontaneous migration to a higher dose in computed radiography (CR) imaging. Proc SPIE 2436:70-73, 1994

4. Gur D: Operating at the diagnostic margins: Image quality considerations. Am J Roentgenol 160:1341-1342, 1993

5. Wilson AJ, Mann FA, Murphy WA Jr, et al: Photostimulable phosphor digital radiography of the extremities: Diagnostic accuracy compared with conventional radiography. Am J Roentgenol 157:533-538, 1991

6. Jonsson A, Borg A, Hannesson P, et al: Film-screen vs digital radiography in rheumatoid arthritis of the hand: An ROC analysis. Acta Radiol 35:311-318, 1994

7. Murphey MD: Digital skeletal radiography: Spatial resolution requirements for detection of subperiosteal resorption. Am J Roentgenol 152:541-546, 1989

8. Page EB: Ordered hypothesis for multiple treatments: A significance test for linear ranks. J Am Statist Assoc 58:216-230, 1963