Richard M. Slone, MD
David H. Foos, MS
Bruce R. Whiting, PhD
Edward Muka, MSE
David A. Rubin, MD
Thomas K. Pilgram, PhD
Kevin S. Kohm, MSEE
Susan S. Young, MS
Paul Ho, MD
Dan D. Hendrickson, MS

Index terms:
Computers, diagnostic aid
Data compression
Images, artifact, **.93, **.99[2]
Images, display, **.1215, **.99[2]
Images, processing, **.99[2]
Picture archiving and communication system (PACS)

Radiology 2000; 215:543–553

Abbreviations:
ICW = image compression workstation
JPEG = Joint Photographic Experts Group
PACS = picture archiving and communication system
WTCQ = wavelet-based trellis-coded quantization

[1] From the Mallinckrodt Institute of Radiology, Washington University School of Medicine, Barnes-Jewish Hospital, Box 8131, 510 S Kingshighway Blvd, St Louis, MO 63110. From the 1998 RSNA scientific assembly. Received December 30, 1998; revision requested February 23, 1999; revision received July 16; accepted August 18. **Address correspondence to** R.M.S. (e-mail: *sloner@mir.wustl.edu*).

[2] **. Multiple body systems

Author contributions:
Guarantor of integrity of entire study, R.M.S.; study concepts, B.R.W., E.M.; study design, E.M., D.H.F., S.S.Y., K.S.K.; definition of intellectual content, E.M., B.R.W., D.H.F.; literature research, R.M.S.; experimental studies, D.H.F., K.S.K., S.S.Y.; data acquisition, R.M.S., E.M., P.H., D.A.R., B.R.W.; data analysis, R.M.S., T.K.P.; statistical analysis, T.K.P.; manuscript preparation, R.M.S.; manuscript editing, D.D.H., R.M.S., B.R.W., E.M.; manuscript review, D.A.R., D.H.F.

Assessment of Visually Lossless Irreversible Image Compression: Comparison of Three Methods by Using an Image-Comparison Workstation[1]

PURPOSE: To determine the degree of irreversible image compression detectable in conservative viewing conditions.

MATERIALS AND METHODS: An image-comparison workstation, which alternately displayed two registered and magnified versions of an image, was used to study observer detection of image degradation introduced by irreversible compression. Five observers evaluated 20 16-bit posteroanterior digital chest radiographs compressed with Joint Photographic Experts Group (JPEG) or wavelet-based trellis-coded quantization (WTCQ) algorithms at compression ratios of 8:1–128:1 and ×2 magnification by using (*a*) traditional two-alternative forced choice; (*b*) original-revealed two-alternative forced choice, in which the noncompressed image is identified to the observer; and (*c*) a resolution-metric method of matching test images to degraded reference images.

RESULTS: The visually lossless threshold was between 8:1 and 16:1 for four observers. JPEG compression resulted in performance as good as that with WTCQ compression at these ratios. The original-revealed forced-choice method was faster and as sensitive as the two-alternative forced-choice method. The resolution-metric results were robust and provided information on performance above visually lossless levels.

CONCLUSION: The image-comparison workstation is a versatile tool for comparative assessment of image quality. At ×2 magnification, images compressed with either JPEG or WTCQ algorithms were indistinguishable from unaltered original images for most observers at compression ratios between 8:1 and 16:1, indicating that 10:1 compression is acceptable for primary image interpretation.

For the past 2 decades, many researchers in radiology have predicted the replacement of hard copy–based, manually administered, diagnostic imaging operations by electronic picture archiving and communication systems (PACS). Re-engineering of radiology operations with PACS can dramatically improve health care delivery by enabling rapid distribution of images and information, improving resource utilization, and providing better service to caregivers (1,2). Although advances in technology have allowed successful demonstration of this concept, the high cost of systems (3) and the low comfort level of health care delivered with new and unfamiliar techniques has impeded widespread deployment. An important factor in the cost of PACS is the large amount of information contained in radiologic images, which results in terabytes of data that must be managed and distributed (1,2). The requirements for storage devices and networks thus constitute a substantial investment and ongoing costs to achieve the benefits of PACS.

Data compression, a technology that reduces the size of image files, provides immediate

and substantial reduction in the cost of PACS deployment. Image compression techniques are designed to reduce data redundancy by means of special image coding and, as a result, can greatly reduce the effective amount of image data and, therefore, the volume of storage or transmission time required per image. Mathematically lossless compression techniques (compression and reconstruction with no loss of original data) result in compression factors on the order of 2:1 to 3:1 for radiologic images, which are insufficient to produce adequate reductions in transmission time or storage costs. To achieve these goals, compression factors on the order of 10:1 or higher are required, which implies that irreversible or lossy compression must be used; that is, some information must be lost in the compression and reconstruction process. However, some loss of image data can be tolerated without affecting the visual interpretation of an image (4).

A major challenge in the adoption of lossy image compression in the medical community is to develop a body of research that supports the use of fewer data than are included in the full image for primary diagnostic interpretation. The characteristics of the human visual system are such that an image reconstructed after irreversible compression may appear indistinguishable from the original, and, thus, the compression is ''visually lossless'' (5). We believe that under these circumstances, the image is therefore diagnostically lossless; that is, image compression will have no effect on diagnostic interpretation. The purpose of our investigation was to compare three methods for the evaluation of compression artifacts by using a workstation designed to increase observer sensitivity to subtle differences, thereby arriving at a conservative and, we hope, widely accepted estimate of the visually lossless threshold. Image quality assessed by evaluating observers' perception of degradation as a function of compression ratio was the primary focus of our investigation.

MATERIALS AND METHODS

Image Comparison Workstation

The image comparison workstation (ICW) was developed as a collaborative project between the Electronic Radiology Laboratory at our institution and the Health Imaging Research Laboratory of Eastman Kodak (Rochester, NY). The goal was to construct a workstation and software that would allow rapid processing and presentation of images on high-resolution (2,000 × 2,500-pixel) monitors. The ICW was designed specifically for the study of performance with lossy image compression techniques (6).

The workstation consisted of a personal computer (Kayak XU6/300; Hewlett Packard, Palo Alto, Calif) equipped with dual 300-MHz Pentium II processors (Intel, Santa Clara, Calif), a 9-Gbyte RAID (redundant array of inexpensive disks) disk array, and 512 Mbytes of random access memory), with model P1540 display cards (Metheus, Beaverton, Ore) driving a 21-inch-diagonal (53.3-cm-diagonal), 2,048 × 2,560–pixel, low-spatial-noise phosphor (P45), 71-Hz monitor (model DR 110; Data Ray, Westminister, Colo) with a maximum luminance of 220 candelas per square meter (luminance dynamic range of 650:1). The software application, which is implemented for the Windows NT operating system (Microsoft, Redmond, Wash), was developed by the Health Imaging Research Laboratory (Eastman Kodak). The user interface is shown in Figure 1.

The approach used in this study was to compare two versions of an image on a single monitor by using an interactive soft-copy feature. Inherent in the design was the use of ''flicker,'' which is defined as sequential display of two registered images on the same monitor. This method was used to exploit the observer's temporal sensitivity to differences in the image, because the human visual system is naturally drawn to changes in structure or brightness. This technique allows detection of subtle differences and provides a mechanism for comparing image quality loss caused by different kinds of distortion. The observer has direct control of flicker and can set it to automatically change images at up to five times per second; alternatively, the user can use a manual mode, which allows the observer to selectively toggle between the two images, as desired.

The current software allowed ×1 magnification to simulate clinical application, while ×2 and ×4 magnifications were available to help improve detection of subtle differences. A small representation of the entire image with the area chosen for magnification was displayed in the lower left-hand corner of the monitor. The portion of the image displayed for evaluation occupied the upper 80% of the screen, as shown in Figure 1, and measured 30 × 30 cm. The observer could change the region of the image displayed by panning with the mouse on the image representation in the lower left-hand corner. Thus when ×2 or ×4 magnification was used, the observer was free to study any desired segment of the full image. A wide assortment of information could be recorded automatically by the computer as an observer worked through an experiment to compare a series of images and to make choices.

Images

The image set used in this study consisted of 20 digital posteroanterior chest radiographs obtained from the outpatient admitting area at our institution. A commercial selenium detector system (Thoravision; Philips Medical Systems, Shelton, Conn) was used to obtain the images. Images normally have an addressable area of 2,048 × 2,560 pixels with a pixel size of 0.2 mm. The use of digital chest images in the context of primary interpretation is well supported by research results (7–9) on radiologists' preference and performance as assessed with receiver operating characteristic analysis for comparison with state-of-the-art, wide-latitude, dual screen-film images. The images selected for this investigation included studies in men and women with pneumonia, pulmonary nodules, interstitial lung disease, mediastinal masses, catheters, or implanted hardware.

Image Compression

There are many compression techniques available and much research on algorithm improvement (10,11); however, interoperability is crucial to a successful PACS, so we have focused our efforts on existing standards defined by the Joint Photographic Experts Group (JPEG) (12). JPEG baseline is the most widely available block discrete cosine transform algorithm. Advantages include wide availability, interoperability with other JPEG-compliant encoding and decoding software, reasonably fast ''run times,'' and widespread vendor support. It is the only compression algorithm sufficiently documented to be proposed by the National Electrical Manufacturers Association (13) as a standard for Digital Imaging and Communications in Medicine, or DICOM.

Other important methods to investigate are wavelet-compression techniques, because these techniques have special features and functionality (14–16). Driven by broad interest, the JPEG 2000 Committee was established to formulate a new standard, ostensibly to be based on wave-

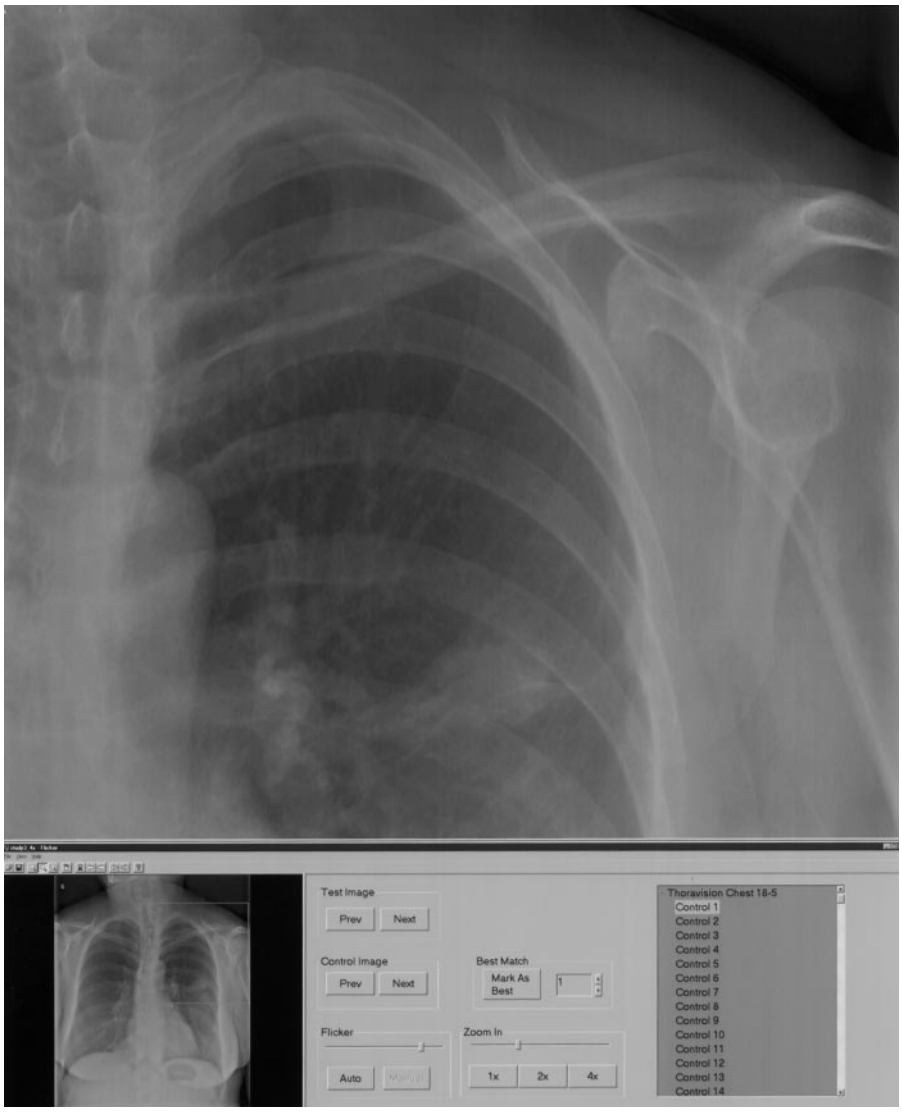

Figure 1. Direct screen capture shows the ICW graphical user interface. A small replication of the entire image is shown in the lower left-hand corner, where the white box defines the portion of the image displayed in the top 80% (2,048 × 2,048 pixels) of the monitor. The set of resolution images for comparison is listed in the box on the right for test case 18–5. The ''Prev'' and ''Next'' buttons in the ''Test Image'' area change the test image being compared. The ''Prev'' and ''Next'' buttons in the ''Control Image'' area change the resolution reference image. This can also be changed by using the wheel on the mouse or the up and down arrows in the ''Best Match'' area. The ''Flicker'' area offers selection of ''Auto'' or ''Manual'' for automatic or manual control of flicker, respectively, and a slider to set the flicker rate in the automatic control mode. The ''Zoom In'' area offers a choice of ×1, ×2, or ×4 magnification. The ×2 magnification is displayed in this example. The ''Mark as Best'' button records the control image selected as the best match to the test case. A warning is displayed if the user tries to move to the next case without recording a choice.

let compression. We thus included the wavelet-based trellis-coded quantization (WTCQ) algorithm developed at the University of Arizona as a representative example of this class of algorithm (17).

The original radiograph (postprocessed, relative log luminance data) was retrieved from an optical disk and linearly transformed from a 0–30,000 scale to a 0–4,095 scale to match the 12-bit requirements of the compression algorithms. The image compression rates were selected on the basis of pilot data and corresponded to 2.00, 1.50, 1.00, 0.75, 0.50, 0.25, and 0.125 bits per pixel. Because the original image was created with 2 bytes per pixel, as is typical for most commercial digital radiographic systems wherein ''byte-packing'' is not used, we calculated, for the convenience of the reader, a compression ratio defined as 16 bits divided by the number of compressed bits per pixel. Representative examples are shown in Figure 2.

Observers

The five independent observers included two imaging scientists (B.R.W., E.M.) with extensive experience in image processing and display and three board-certified radiologists, including specialists in chest (R.M.S.), musculoskeletal (D.A.R.), and general (P.H.) radiology. The introductory training session for each observer included a discussion of the purpose and objectives of the evaluation, a description of the protocol, and an online walk-through of the evaluation procedure and operation of the ICW.

Image Evaluation Methods

Three methods were chosen for image comparison and evaluation. All three were conducted by using the ICW at ×2 magnification, which allowed 25% of the total image area to be viewed at a time. Although it is possible to pan (roam) and view the entire image, observers limited their observations to the upper right-hand quadrant of each image for this study.

Twenty test ''folders'' (computer directories) were prepared and randomized for each reader. Each folder used a different image and contained 16 randomized image replicates, including two unaltered original radiographs and images compressed with each of the seven ratios for the two compression algorithms. The order in which the five observers performed the three experiments was random. The amount of time needed for each reading session, reader confidence for each decision, and representative reading distances were manually recorded.

Two-alternative forced choice.—The 16 test images in each folder were paired with a control image (unaltered original) and were presented sequentially without identification (one pair at a time), with the observer toggling between them in a rapid fashion. The observer was asked to choose the image with ''better quality'' and was forced to choose even if the observer perceived no difference between the images. The observer was asked to record both the image selected and the decision confidence by using a three point scale: score of 0, uncertain or guessing; score of 1, confident; score of 2, very confident. No reader feedback was provided.

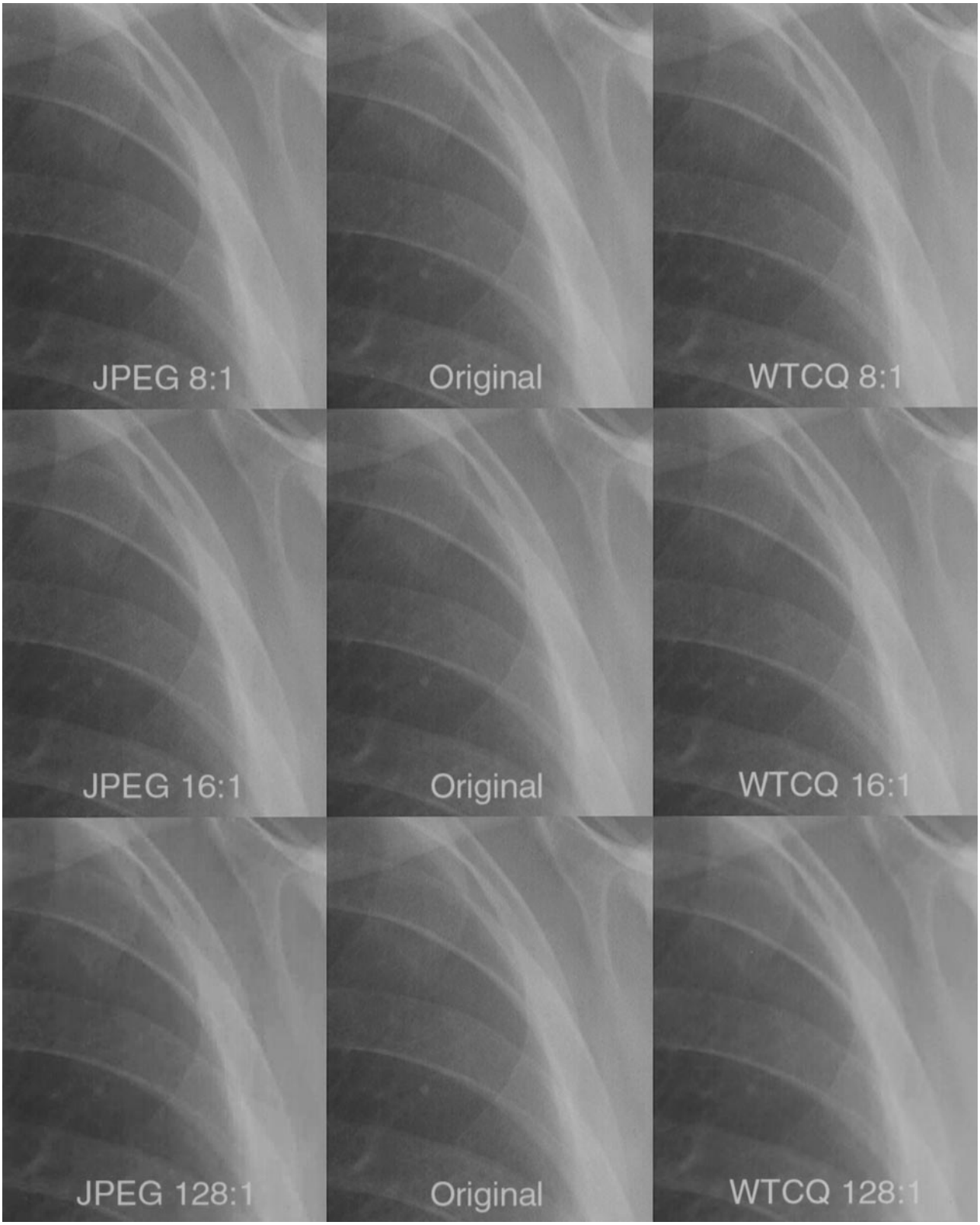

Figure 2. Compression artifacts. Compressed (JPEG, left column; WTCQ, right column) and noncompressed digital (middle column) posteroanterior radiographs of a selected region of interest show the effects of compression at ratios of 8:1 (top row), 16:1 (middle row), and 128:1 (bottom row). With both JPEG and WTCQ algorithms, the image compressed at 8:1 is indistinguishable from the original. At a compression ratio of 128:1, the manifestation of ''tiling'' or ''blocking'' artifacts on the JPEG image and blurring on the WTCQ image are readily apparent.

With this method, a visually lossless level was indicated when responses were evenly divided between test and control images or matched the score distribution for the original-original image pair.

Original-revealed forced choice.—The same 20 test folders were used to conduct a modified two-alternative forced-choice experiment. As in the traditional forced-choice method, each test image was paired with an unaltered original and was presented sequentially (one pair at a time) on the ICW, with the observer toggling between them. The difference between this task and two-alternative forced choice was that the original image was identified for the reader. The observer was asked to compare the test image to the original and decide if the images were equivalent or if there was visible artifact, loss of fidelity, or degradation in the quality of the test image. The observer knew that sometimes two originals would be shown. We refer to this method as the original-revealed forced-choice method. In an investigation of the human visual system, Gur et al (18) advocate comparison of test images with a known original image to maximize sensitivity to distortion. With this method, a visually lossless level was indicated when the percentage of test images rated as equivalent approximated 100% or matched the score distribution for the original-original image pairs.

Spatial resolution metric.—In this experiment, a test image that had been prepared by applying the compression algorithm was compared to a set of reference images that had been prepared by degrading the image in a controlled manner. A set of 15 reference images was prepared for each test image by using ''blur'' as the metric. Blur was introduced by degrading the spatial resolution of the image while maintaining image size. This bandwidth reduction was performed by transforming the image into the frequency domain, applying a set of 14 power-law filters, then ''back-transforming'' the filtered data into the spatial domain. Power-law filters were selected to provide rectangular bandpass characteristics in the frequency domain while producing minimal distortion in the spatial domain (6).

Because the image is two-dimensional while the filters are separable (one-dimensional), the square of the bandwidth reduction can be used as a measure of two-dimensional spatial resolution. Figure 3 shows the full range of grades of spatial filtration applied to a representative image. For example, a bandwidth reduction factor of 1.25 (grade 4), which corresponded to a 20% reduction in bandwidth, can be thought of as displaying 80% of the original pixels, which were then magnified to the full display size, resulting in a spatial resolution of 64%. This would be analogous to displaying a 2,048 × 2,048-pixel image on a 1,600 × 1,600-pixel monitor. Similarly, a bandwidth reduction factor of 2 (grade 7) corresponded to displaying half the pixels, or 25% resolution, and was analogous to displaying a 2,048 × 2,048-pixel image on a 1,024 × 1,024-pixel monitor.

The test and reference images were presented sequentially for comparison at the ICW, with the observer toggling between them in a rapid fashion. The observer changed the comparison image by using the wheel on the mouse to select from the set of 15 reference images. The observer selected the reference image that most closely matched the test image in terms of clinical utility.

Data Analysis

Data took the form of reader decisions: For two-alternative forced-choice experiments, observers decided which of the two images appeared to be superior; for original-revealed forced-choice experiments, observers decided if an image was equivalent to the original or degraded; for resolution-metric tests, observers decided which level of blurring most closely matched, with respect to clinical utility, the level of compression. Analysis took the form of calculation of the proportion of decisions in a given category, with 95% confidence limits (19). Time needed to complete image comparison and reading distance were evaluated by comparing means.

RESULTS

When making comparisons, observers reported that their focus of attention included structural detail, particularly bone edges and trabecular patterns, and areas of uniform opacity such as the soft tissues of the chest wall. Observers differed in their ability to detect degraded images, but when results from all observers were combined, a fairly clear pattern was found. The mean results for two-alternative forced-choice, original-revealed forced-choice, and spatial resolution–metric experiments for all cases and observers are shown in Table 1. Individual results are shown in Table 2.

Two-Alternative Forced Choice

The overall results for the two-alternative forced-choice experiments are presented in Figure 4. When presented with images that were indistinguishable, observers guessed, which resulted in a chance, or approximately 50:50, distribution. As part of the two-alternative forced-choice experiment, observers were presented with 40 pairs of original images, one denoted as the ''test'' image and one as the ''control'' image. The overall response rate for selecting the test or control image as the better image (when in fact both were identical) was 47% or 53%, respectively.

The mean responses for images compressed at 8:1 and 11:1 with both algorithms fall within the 95% CIs of the responses for the original images. The 8:1 compressed images were selected as the better image slightly more often than the original for both JPEG (57%) and WTCQ (65%) images. The responses for both JPEG and WTCQ images compressed at 11:1 indicate a slight tendency for observers to select the original image as the better image. The responses for images at 16:1 compression indicate that observers differentiated WTCQ images from original images more frequently than they differentiated JPEG images and that the visually lossless compression threshold was crossed for both. At compression ratios of 21:1 and higher, observers consistently (more than 95% of the time) chose the original as the better image, which indicates that the compression artifact was clearly evident.

These results are supported by the mean confidence scores for two-alternative forced-choice experiments (Table 1). Observer confidence was low (mean score of 0.2) when the choice was between images in an original-original pair and remained low with 8:1 and 11:1 compressed images. Observer confidence increased (mean score > 1.0) when evaluating images compressed at 16:1 and became substantially higher (mean score > 1.7) at compression ratios of 21:1 and higher.

Individual observers trends varied, as shown in Table 2. With JPEG images, the 8:1 compression images were indistinguishable from the original images for four observers, and the 11:1 compression images were indistinguishable for three observers. Observer C, who also had the closest mean viewing distance (as close as 8 cm), noted degradation in all compressed images except for a few compressed at 8:1. Although observer C noted no structural degradation, he noted a change in some individual pixels. With WTCQ compression, the pattern for all five observers was similar and suggested the presence of a visually lossless thresh-

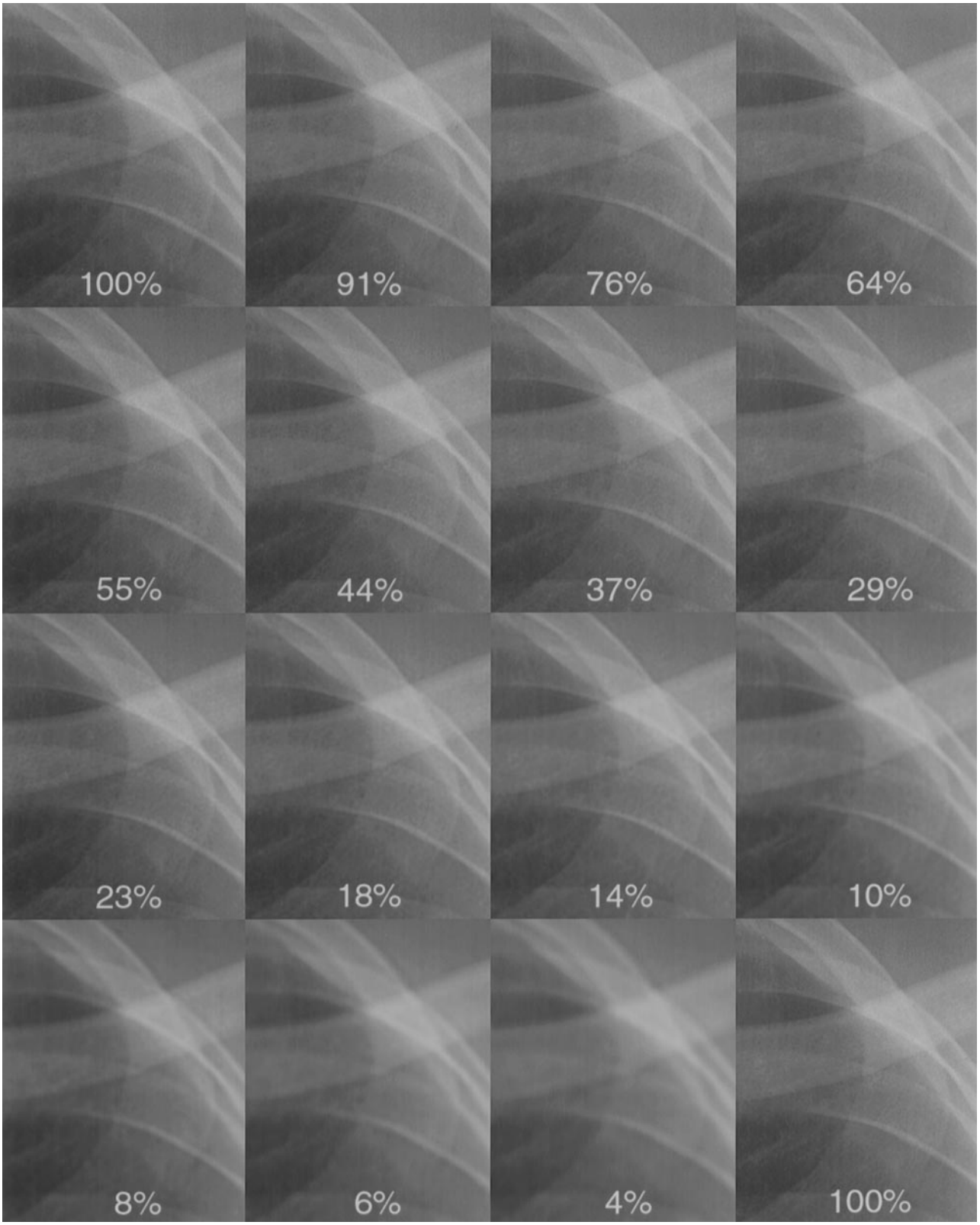

Figure 3. Resolution metric. Composite image shows the spatial-resolution scale applied to a representative portion of a radiograph. Reduction in resolution progresses from grade 1 (top left and bottom right: unaltered original, *100%*) to grade 15 (bottom row, second from right: *4%*). The percentages are a measure of two-dimensional spatial resolution, which was determined on the basis of the square of the bandwidth reduction implemented with power-law filters and can be thought of as the percentage of pixels displayed.

old at ×2 magnification for images compressed with a ratio higher than 8:1 for all observers and 11:1 to 16:1 for two of the observers.

Original-revealed Forced Choice

In this experimental design, the observer was presented with an original image and was asked if the test image was indistinguishable or if there was any visible artifact, degradation, or loss of fidelity. Observers were told that there were several unaltered original images included with the test images. When presented with a pair of original images, the observer declared the images to be indistinguishable 95% of the time. The composite results for all observers are shown in Figure 5. As with the two-alternative forced-choice results, the greatest change in performance was found between 11:1 and 16:1 compression for both JPEG and WTCQ algorithms.

Review of the individual results showed a tendency for two observers to detect degradation in about half of the WTCQ images compressed at 11:1 and a clear trend for four observers to detect degradation in such images compressed at a ratio higher than 16:1. For JPEG images, a larger percentage of images were considered to be degraded at a given compression ratio overall, but this was due primarily to the results of observer C, who detected a difference in all compressed images except for a few compressed at 8:1. Of the other four observers, one detected degradation in about half the JPEG images compressed at 11:1, but none detected degradation in images compressed at 8:1. The mean percentages for these four observers were 94% for JPEG images compressed at 8:1 and 90% for images compressed at 11:1—in both cases, close to the 95% rate found for original-original pairs. At compression ratios of 21:1 and higher with either JPEG or WTCQ, the degradation was detected over half the time by four of the observers.

Spatial Resolution Metric

Fifteen reference images with controlled degradation were available for comparison with each test image, including grade 1 degradation (ie, an unaltered original image). Although observers were able to score intermediate grades, this seldom occurred, which suggests that the existing choices offered were sufficient. Observers reported the ability to consistently select a reference image that matched the degradation in the compressed image, although this was easier with WTCQ images than with JPEG images at high compression ratios.

The original images were matched with grade 1 images (unaltered original) on the resolution scale 96% of the time; with grade 2, 3% of the time; and with grade 3, 1% of the time. At a compression ratio of 8:1, JPEG images were matched with grade 1 images 94% of the time; with grade 2 images, 5% of the time; and with grade 3, 1% of the time. At a ratio of 11:1, JPEG images were matched with grade 1 images 71% of the time; with grade 2 images, 15% of the time; with grade 3 images, 11% of the time; and with grade 4–6 images, 1% of the time each. At a compression ratio of 8:1, WTCQ images were matched with grade 1 images 96% of the time; with grade 2 images, 3% of the time; and with grade 4 images, 1% of the time. At a compression ratio of 11:1, WTCQ images were matched with grade 1 images 67% of the time; with grade 2

TABLE 1
Combined Results for All Observers and Methods

Compression Factor*	No. of Observations	Two-Alternative Forced Choice†		Original-revealed Forced Choice‡		Spatial Resolution Metric§	
		JPEG	WTCQ	JPEG (%)	WTCQ (%)	JPEG (%)	WTCQ (%)
16 (1:1)	200	47 (0.2)	47 (0.2)	95	95	99 ± 3	99 ± 3
2 (8:1)	100	57 (0.4)	65 (0.2)	79	92	99 ± 4	98 ± 8
1.5 (11:1)	100	41 (0.6)	36 (0.4)	70	76	94 ± 11	94 ± 13
1 (16:1)	100	22 (1.3)	14 (1.2)	29	30	76 ± 17	72 ± 21
0.75 (21:1)	100	5 (1.7)	4 (1.8)	7	11	65 ± 17	57 ± 18
0.5 (32:1)	100	1 (2.0)	2 (1.9)	2	5	37 ± 11	36 ± 11
0.25 (64:1)	100	3 (1.9)	0 (2.0)	5	3	17 ± 8	21 ± 5
0.125 (128:1)	100	0 (2.0)	1 (2.0)	2	3	8 ± 4	10 ± 3

* Value is the number of bits per pixel. Numbers in parentheses are the compression ratio.
† Data are percentage of test images selected as having higher quality than the original image. Number in parentheses is the confidence score.
‡ Data are percentage of test images judged to be equivalent to the original image.
§ Values are the mean percentage two-dimensional spatial resolution plus or minus the SD.

TABLE 2
Visually Lossless Threshold Estimates for Individual Observers

Observer	Mean Viewing Distance ± SD (cm)	Two-Alternative Forced Choice			Original-revealed Forced-Choice			Spatial Resolution Metric†		
		JPEG*	WTCQ*	Time (sec)	JPEG*	WTCQ*	Time (sec)	JPEG‡	WTCQ‡	Time (sec)
A	28 ± 4.2	11–16:1	8–11:1	23	11–16:1	11–16:1	15	13:1 (0.91)	11:1 (0.91)	37
B	30 ± 12.0	11–16:1	8–11:1	19	8–11:1	8–11:1	16	11:1 (0.86)	6:1 (0.90)	68
C	13 ± 2.9	<8:1	8–11:1	34	<8:1	8–11:1	29	11:1 (0.86)	12:1 (0.90)	47
D	25 ± 2.7	16–21:1	11–16:1	32	16–21:1	16–21:1	25	16:1 (0.93)	15:1 (0.92)	53
E	38 ± 2.9	11–16:1	11–16:1	30	11–16:1	11–16:1	21	12:1 (0.90)	10:1 (0.93)	30
Average	26‡	11–16:1§	8–11:1§	28‡	11–16:1§	11–16:1§	21‡	12:1‡	11:1‡	47‡

* Data are the mean highest compression ratio for images judged to be indistinguishable from the original image (score similar to that for original-original image pairs).
† Visually lossless threshold was based on the x intercept of the plot of log compression ratio versus log spatial resolution. Value in parentheses is the r^2 value from the linear regression equation used to calculate the compression level.
‡ Value is the overall mean.
§ Value is the median.

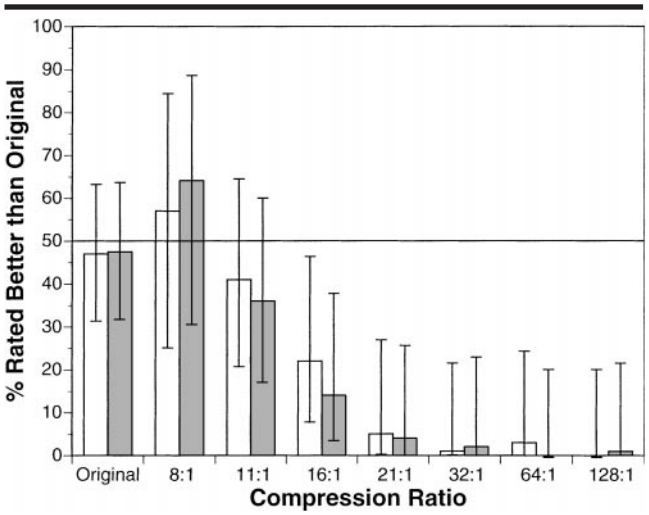

Figure 4. Bar graph shows the combined results for five observers in the two-alternative forced-choice experiments, with the percentage of test images judged to be superior to the original image. The horizontal line at 50% represents the expected result for a purely random selection. Error bars = 95% CIs, gray bars = WTCQ images, white bars = JPEG images.

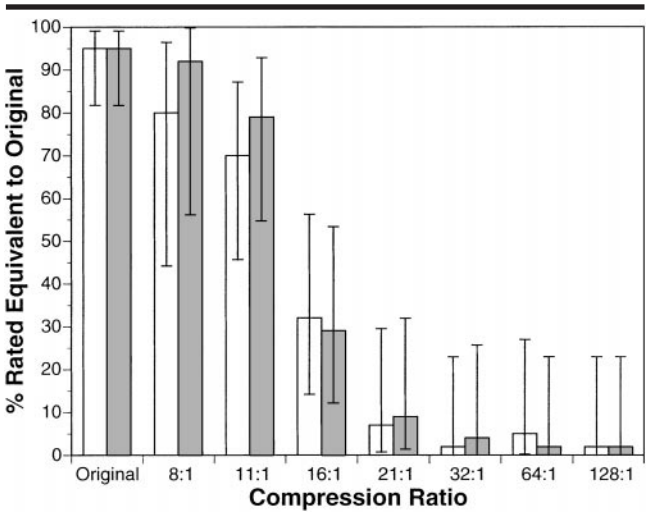

Figure 5. Bar graph shows the combined results for five observers in the original-revealed forced-choice experiments, with the percentage of test images classified as equivalent to an original image. Error bars = 95% CIs, gray bars = WTCQ images, white bars = JPEG images.

images, 8% of the time; with grade 3 images, 13% of the time; and with grades 4–6 images, 3% of the time each.

The mean percentage resolution matched with each level of compression for the two algorithms is shown in Table 1. There was a progressive decrease in resolution as the compression level increased. As with the two-alternative forced-choice and original-revealed forced-choice experiments, the greatest initial change for both algorithms occurred between compression ratios of 11:1 and 16:1. Although the data were not highly precise, at low compression levels JPEG images were matched with a higher resolution than were WTCQ images, whereas the opposite occurred for the highest compression ratios.

The relationship between compression and resolution was further assessed by plotting the data for JPEG images (Fig 6a) and WTCQ images (Fig 6b). Transformation of the data onto a log scale provided a distribution suitable for linear regression. The log of the compression ratio was plotted on the x axis. In this format, 1.0 corresponded to a compression ratio of 10:1; 1.2, to a compression ratio of 16:1; and so forth. The log of the reciprocal of the percentage spatial resolution was plotted on the y axis and ranged from 0 for the unaltered original to 1.4 for the bandwidth reduction factor of 5.0 (4% resolution). Linear regression was used to calculate the line of best fit. The x intercept was a predictor of the visually lossless threshold. This method was used to calculate the expected visually lossless threshold for each observer by using data for compression ratios higher than the visually lossless level, namely, 16:1 and higher. The calculated threshold and r^2 value for the linear regression fit to the data are shown in Table 2 for each reader and both algorithms.

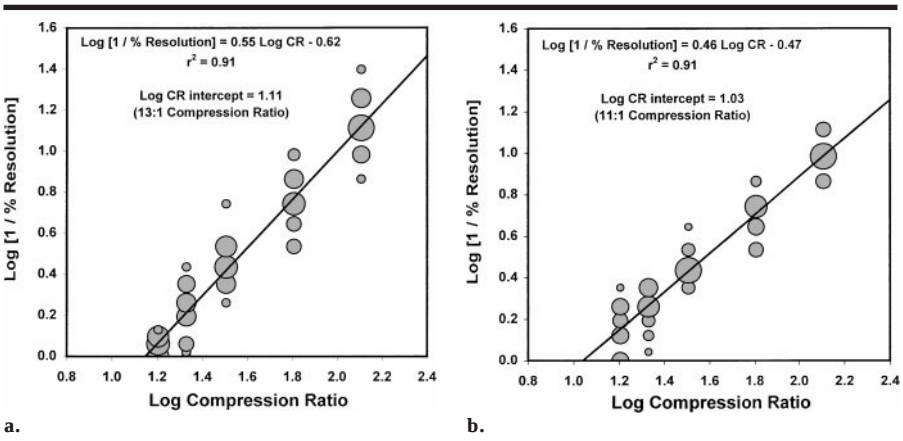

Figure 6. Resolution-metric matching with **(a)** JPEG images and **(b)** WTCQ images. Scatterplots show the relationship between compression ratio (*CR*) and percentage resolution for observer A. These data were obtained by matching the blur image set with 20 test images at each of five compression ratios (16:1, 21:1, 32:1, 64:1, and 128:1). The area of each dot is proportional to the number of superimposed data points. The straight lines are the regression lines. **(a)** The x intercept for JPEG images, determined by means of linear regression on the 100 data points, indicates that the estimated visually lossless threshold is 13:1. **(b)** The x-intercept for WTCQ images, determined by means of linear regression on the 100 data points, indicates that the visually lossless threshold is 11:1.

Linear regression also was performed for each observer with systematic elimination of one compression level and again with only the 16:1, 32:1, and 64:1 data. These data manipulations had almost no effect on the x intercept. The mean visually lossless compression ratio was 12:1 for JPEG images and 11:1 for WTCQ images, in all cases.

Viewing Distance and Reading Time

Viewing distance and reading time varied among observers (Table 2). The resolution-metric task required the greatest time commitment, taking about twice as long

to complete as the two-alternative forced-choice task. The original-revealed forced-choice task was the quickest method, taking, on average, 25% less time than the two-alternative forced-choice task.

Four observers wore prescription eyeglasses, and one (observer D) wore nonprescription reading glasses. Observer C positioned himself much closer to the images than did the other four observers. This observer was very nearsighted and, with prescription lenses, was able to focus much closer than the typical viewer. We have already noted the differences in some of the results of observer C.

DISCUSSION

Reversible, or mathematically lossless, image compression provides an inadequate reduction in the amount of data to provide substantial engineering advantages or cost reductions for image transmission or storage. Irreversible, or lossy, image compression is needed to achieve these goals. In certain circumstances, visually lossy images may be diagnostically lossless; although image compression artifacts are detectable, their presence does not affect diagnostic performance (20). Several studies in which receiver operating characteristic analysis was used (21) have shown this to be true (22–24).

Although diagnostically lossless criteria would likely allow relatively high degrees of compression and substantial cost savings, the presence of perceivable artifact reduces acceptance among skeptical radiologists, and the validity of diagnoses based on lossy images in medicolegal proceedings is yet to be determined. In addition, receiver operating characteristic analyses are time-consuming and expensive and can usually be used to address only narrowly defined tasks. Multiple receiver operating characteristic studies would be required to provide results that cover a broad range of potential abnormalities with confidence. In light of this, we chose to concentrate on a more conservative criterion—namely, visually lossless compression for the full range of texture and density gradients in the image—in the belief that compressed images that are indistinguishable in any way from an original image are diagnostically lossless and would be readily acceptable even by skeptical radiologists.

The ICW proved to be a powerful tool for conducting observer studies. A large number of image sets could be evaluated efficiently. The alternating presentation of registered images at the ICW was intended to maximize viewer sensitivity to subtle image compression artifacts. Similarly, viewing distance was unconstrained, and images were magnified by a factor of two. Both of these latter factors should increase the conspicuity of subtle detail not visible when the region of interest occupied a smaller viewing angle. Results of previous studies (25,26) have shown decreases in detection of degradation due to image compression as viewing distance increases.

Display of images magnified by a factor of two also has clinical relevance, because most PACS stations have a ''magnifying glass'' tool. Radiologists may use this tool to evaluate isolated areas of the image at $\times 2$ magnification to detect subtle disease, particularly pneumothorax, fracture, and interstitial lung disease. We believe, therefore, that the use of magnification, close viewing distance, and flicker to exploit an observer's temporal sensitivity between image differences should result in a conservative and, we hope, widely accepted estimate of the visually lossless threshold.

We compared three methods for studying observer detection of image degradation. The two-alternative forced-choice and original-revealed forced-choice methods were similar. The difference was subtle but important. In the former method, when confronted with two original images or a visually lossless image, the observer was forced to guess, thus selecting the image that was compressed but indistinguishable from the original image 50% of the time. As the degradation became more apparent, selection of the compressed image decreased toward 0%. In the original-revealed forced-choice method, the observer would be expected to declare two original or visually lossless images to be equivalent 100% of the time. However, sometimes an original was judged to be degraded because the observer was intent on detecting the subtlest differences and may have ''over read'' the image. This was observed in our original-revealed forced-choice results, where observers declared the unaltered test image to be degraded 5% of the time and were thus operating at a low threshold for reporting degradation.

With the original-revealed forced-choice method, as the degradation became more apparent, judgment of the compressed image as equivalent to the original decreased toward 0%. Thus, the expected results ranged from 0% to 50% with the two-alternative forced-choice method and from 0% to 100% with the original-revealed forced-choice method. This has certain advantages for the display and analysis of the data. The patterns shown in the data suggest that observers detected degradation in the compressed images as frequently when using the original-revealed forced-choice method as they did with the two-alternative forced-choice method, which implies that the former technique is as sensitive as the latter. The original-revealed forced-choice experiment was conducted in less time by the observers and thus may have provided results more efficiently.

In contrast, an advantage of the two-alternative forced-choice method is that it can reveal trends in preference rather than just provide information about detection of differences. Of particular interest are the results for comparisons with images compressed at 8:1. For both JPEG and WTCQ, the 8:1 images were considered to be better than the original image 55%–65% of the time. Although this was a slight deviation from the expected chance result of 50% for visually lossless images, the trend was the opposite of that with all other levels of compression, where the compressed image was judged to be better less than 50% of the time. We hypothesize that low compression levels have the same effect as a low-pass filter, because the image is smoothed and the conspicuity of image noise is effectively decreased. Thus, the original-revealed forced-choice method was sensitive for detecting differences, but only with the two-alternative forced-choice method did the results reflect a preference for a test image over an original image.

Despite expectations for improved performance with wavelet-based algorithms, we found that the JPEG baseline algorithm resulted in performance that was as good as, if not better than, performance with the WTCQ algorithm implemented at low compression ratios. This is critical, because JPEG is a current standard that permits interoperability in a PACS environment. The data did suggest that WTCQ was better at very high compression ratios. These results are probably related to the way these two algorithms handle the data and the manner in which the artifact is manifested (ie, primarily as ''tiling'' or ''blocking'' with JPEG and as blurring with WTCQ).

Although radiographs of different body parts are likely to emphasize different compression artifacts, the chest radiograph demonstrates a broad range of structural and tonal characteristics that provide the opportunity to note degradation in soft tissues of uniform opacity and intricate trabecular and pulmonary parenchymal detail. Thus, although our study

was limited to posteroanterior chest radiographs, we believe the results should be representative of other projection radiographs.

The resolution-metric method was intended to project the image characteristics of the compression algorithm onto a quantifiable dimension, thus providing a basis for comparison of various compression techniques. Observers reported being able to consistently select a reference image that matched the degradation in the compressed image, although at high compression ratios this was easier with WTCQ images than with JPEG images. We believe this is because the artifact introduced by the WTCQ algorithm is similar to the degradation introduced by blurring. At high compression ratios, it was more difficult to equate the tiling or blocking artifacts introduced by JPEG with blur artifacts introduced by WTCQ.

The use of spatial blurring by means of bandwidth limitation in the frequency domain proved to be a reasonable choice for a matching metric. Further exploration of degradation mechanisms for reference images might include different frequency filtering, variation of quantization levels, addition of random noise (both white noise and quantum mottle), or a combined blur-noise operation that follows a relationship similar to that of conventional x-ray detectors (ie, a speed-sharpness trade-off similar to that of screen-film or storage phosphor images).

This technique required more time than did the forced-choice methods but provided more information, particularly about the relative "performance" of algorithms at compression ratios above the visually lossless threshold. Such a comparison would be critical in a comparison of a new algorithm with an accepted one in the visually lossy range. In addition, we found that data from the resolution matching method could be used to estimate the visually lossless threshold by using linear regression. The results were nearly identical to those obtained with the two-alternative forced-choice and original-revealed forced-choice methods, which showed that the visually lossless threshold for the conservative viewing situation (superimposed images, ×2 magnification, close viewing conditions) was greater than 10:1 for both algorithms, with JPEG images resulting in slightly better performance than WTCQ images. We also found this technique to be robust, with the number of data points and specific compression levels used having little effect on the predicted value.

The ICW is a versatile and powerful tool for comparative assessment of image quality. Sequential registered display of magnified images should help optimize observer sensitivity to differences and improve detection of subtle degradation, resulting in conservative estimates. The ICW monitor is a component of currently available commercial PACS systems, which makes translation of the results to a true clinical environment practical. These studies were conducted in the context of primary, rather than secondary, interpretation, and the methods were robust in that they could be used with various equipment and acquisition, presentation, display, and viewing tasks.

The objective of our research was to establish a basis for acceptance of irreversible image compression for primary interpretation of diagnostic images. We believe that diagnostic loss can be avoided by using visually lossless levels of compression and that these compression levels are high enough to provide important time and cost savings and thus serve as a critical component for improved health care delivery in the PACS environment. Our results suggest that the JPEG baseline algorithm results in performance that is as good as that which results from a more complex wavelet-compression algorithm and that 10:1 image compression is visually lossless for most observers and is, therefore, acceptable for primary image interpretation without risk of affecting diagnosis. The composite results shown in Figures 4 and 5 demonstrate that most observers consistently detect image degradation at compression ratios of 21:1 and higher. This is supported by other researchers (27) who have suggested that 10:1 compression does not influence detection of subtle interstitial abnormalities but that important information may be lost at a ratio of 20:1, particularly when the images are interpreted by experienced thoracic radiologists.

Our research design, which used magnification, unrestricted viewing distance, and superimposition of registered images, was intended to produce a conservative estimate of acceptable image compression levels. Our hope is that comfort with visually lossless but mathematically degraded data sets will open the path for studies in a routine clinical environment, where less conservative constraints would support the use of more aggressive compression with its accompanying benefits. Future work is needed to address less stringent visually lossy but diagnostically lossless levels of compression by assessing diagnostic performance outcomes.

Acknowledgments: We are grateful to G. James Blaine, DSc, Jerome Cox, PhD, and R. Gilbert Jost, MD, at the Mallinckrodt Institute of Radiology, Washington University School of Medicine (St Louis, Mo), for their critical insight, thoughtful suggestions, and support in facilitating this collaborative investigation.

References

1. Blaine GJ, Jost RG, Martin L, Weiss DA. Information and image integration: Project Spectrum. Proc SPIE 1998; 3339: 430–439.
2. Kahn MG. Enterprise-wide clinical data integration. In: Brennan P, Schneider SJ, Tornquist E, eds. Information networks for community health. New York, NY: Springer-Verlag 1997; 41–45.
3. Beard D, Parrish D, Stevenson D. A cost analysis of film image management and four PACS based on different network protocols. J Digit Imaging 1990; 3:108–118.
4. Cox GG, Cook LT, Insana MF, et al. The effects of lossy compression on the detection of subtle pulmonary nodules. Med Phys 1996; 23:127–132.
5. Daly S. Application of a noise-adaptive contrast sensitivity function to image data compression. Opt Eng 1990; 29:977–987.
6. Foos DH, Slone RM, Whiting B, et al. Dynamic viewing protocols for diagnostic image comparison. Proc SPIE 1999; 3663: 108–120.
7. Woodard PK, Slone RM, Gierada DS, Reiker GR, Pilgram TK, Jost RG. Chest radiography: depiction of normal anatomy and pathologic structures with selenium-based digital radiography versus screen-film radiography. Radiology 1997; 203:197–201.
8. Woodard PK, Slone RM, Sagel SS, et al. Detection of CT-proved pulmonary nodules: comparison of selenium-based digital and conventional screen-film chest radiographs. Radiology 1998; 209:705–709.
9. Floyd CE, Baker JA, Chatos HG, Delong DM, Ravin CE. Selenium-based digital radiography of the chest: radiologists preference compared with film-screen radiographs. AJR Am J Roentgenol 1995; 165: 1353–1358.
10. Jones PW, Daly S, Gaborski RS, Rabbani M. Comparative study of wavelet and DC decompositions with equivalent quantizations and encoding strategies for medical images. Proc SPIE 1995; 2431:571–582.
11. Eckert MP. Lossy compression using wavelets, block DCT, and lapped orthogonal transforms optimized with a perceptual model. Proc SPIE 1997; 3031:339–350.
12. Pennebaker WB, Mitchell JL. JPEG still image data compression standard. New York, NY: Van Nostrand Reinhold, 1993.
13. Digital Imaging and Communications in Medicine (DICOM). V. Data structures and encoding. NEMA standards publication no. PS 3.5-1996. Rosland, Va: National Electrical Manufacturers Association, 1996.
14. Saipetch P, Ho B, Ma M, Chuang K, Wei J. Radiological image compression using wavelet transform with arithmetic coding. Proc SPIE 1994; 2164:449–459.
15. DeVore R, Jawerth B, Lucier B. Image compression through wavelet transform coding. IEEE Trans Info Theory 1992; 38:719–746.

16. Goldberg M, Pivovarov M, Mayo-Smith WW, et al. Application of wavelet compression to digitized radiographs. AJR Am J Roentgenol 1994;163:463–468.
17. Sriram P, Marcellin MW. Image coding using wavelet transforms and entropy-constrained trellis coded quantization. IEEE Trans Image Proc 1995; 4:725–733.
18. Gur D, Rubin D, Kart B, et al. Forced choice and ordinal discrete rating assessment of image quality: a comparison. J Digit Imaging 1997; 10:103–107.
19. Fleiss JL. Statistical methods for rates and proportions. New York, NY: Wiley, 1981.
20. Mori T, Nakata H. Irreversible data compression in chest imaging using computed radiography: an evaluation. J Thorac Imaging 1994; 9:23–30.
21. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986; 21:720–733.
22. MacMahon H, Doi K, Sanada S, et al. Data compression: effect on diagnostic accuracy in digital chest radiography. Radiology 1991; 178:175–179.
23. Aberle DR, Gleeson F, Sayre JW, et al. The effect of irreversible image compression on diagnostic accuracy in thoracic imaging. Invest Radiol 1993; 28:398–403.
24. Savcenko V, Erickson BJ, Palisson PM, et al. Detection of subtle abnormalities on chest radiographs after irreversible compression. Radiology 1998; 206:609–616.
25. Cox JE, Muka E, Wang X, Blaine GJ. Factors affecting the selection of compression algorithms for projection radiography. Proc SPIE 1997; 3031:256–264.
26. Pilgram TK, Slone RM, Muka E, Cox JR, Blaine GJ. Perceived fidelity of compressed and reconstructed radiological images: a preliminary exploration of compression, luminance, and viewing distance. J Digit Imaging 1998; 11:168–175.
27. Kido S, Ikezoe J, Kondoh H, et al. Detection of subtle interstitial abnormalities of the lungs on digitized chest radiographs: acceptable data compression ratios. AJR Am J Roentgenol 1996; 167:111–115.