Pamela C. Cosman, PhD • H. Christian Davidson, MD • Colleen J. Bergin, MD
Chien-Wen Tseng, MS • Lincoln E. Moses, PhD • Eve A. Riskin, PhD
Richard A. Olshen, PhD • Robert M. Gray, PhD

Thoracic CT Images: Effect of Lossy Image Compression on Diagnostic Accuracy[1]

PURPOSE: **To evaluate the effects of lossy image (noninvertible) compression on diagnostic accuracy of thoracic computed tomographic images.**

MATERIALS AND METHODS: **Sixty images from patients with mediastinal adenopathy and pulmonary nodules were compressed to six different levels with tree-structured vector quantization. Three radiologists then used the original and compressed images for diagnosis. Unlike many previous receiver operating characteristic–based studies that used confidence rankings and binary detection tasks, this study examined the sensitivity and predictive value positive scores from nonbinary detection tasks.**

RESULTS: **At the 5% significance level, there was no statistically significant difference in diagnostic accuracy of image assessment at compression rates of up to 9:1.**

CONCLUSION: **The techniques presented for evaluation of image quality do not depend on the specific compression algorithm and provide a useful approach to evaluation of the benefits of any lossy image processing technique.**

Index terms: Computed tomography (CT), image processing • Data compression • Images, interpretation • Thorax, CT, 60.1211

Radiology 1994; 190:517–524

THE need for data compression in medical imaging is increasing with the growing use of digital imaging modalities. A tertiary care hospital of average size, for example, with a mostly digital diagnostic radiology department, must maintain over 1 terabyte (1 trillion bytes) of image data that it produces each year (1). Storage and transmission of such a large number of digital images pose a substantial computational challenge despite technologic advances. Although digital images can be stored in a hard-copy format, long-term storage on film is not completely satisfactory, as film occupies valuable space and is often lost, misfiled, damaged, or needed in more than one location at a time. In addition, an image reduced to analog form is no longer available for digital processing such as window and level manipulation, cine display, or transmission for off-site or computer-aided diagnoses.

Image compression can be divided into two general categories: lossless and lossy. Lossless algorithms allow perfect reconstruction of the original image after compression and typically provide compression ratios of between 2:1 and 4:1 on medical images, depending on the particular imaging modality (2–5). Lossy compression schemes provide higher ratios but only at the cost of irrecoverable data loss (6–11). In this study, we evaluate vector quantization, a compression technique that has worked well on nonmedical images and has several useful properties: It is simple and fast to implement in software, requires no real-time adaptation or special purpose hardware, and has a natural use in progressive transmission.

Because a lossy compressed image differs from the uncompressed original, methods for evaluating medical image quality have been examined in many recent studies (9–17). In general, image quality is quantified with signal-to-noise ratios or with statistical analyses of viewer quality ratings. Most such studies used receiver operating characteristic (ROC) curves, which have several drawbacks that are discussed later. This study examined the effects of lossy compression on the diagnostic quality of computed tomographic (CT) images of patients with two common abnormalities: mediastinal adenopathy and pulmonary nodules. Images could contain multiple abnormalities, and the evaluation task was structured to simulate the actual clinical tasks of radiologists. This approach necessitates a departure from most traditional evaluation tasks and statistical analysis methods. However, the statistical methods developed in this study are applicable to the evaluation of any compression technique.

MATERIALS AND METHODS

Tree-structured Vector Quantization

Vector quantization is based on the concept of dividing an image into small blocks of pixels. Each block from an original image is compressed or encoded by selecting a good approximation from a relatively small collection of possible blocks called code words. The collection of code words is called a code book. When a code word is chosen as a good match to the original

[1] From the Information Systems Laboratory, Department of Electrical Engineering, Stanford University, 127 Durand Bldg, Stanford, CA 94305-4055 (P.C.C., C.W.T., R.M.G.); the Departments of Radiology (H.C.D.) and Statistics, Division of Biostatistics (R.A.O., L.E.M.), Stanford University School of Medicine, Calif; the Department of Radiology, University of California, San Diego, La Jolla (C.J.B.); and the Department of Electrical Engineering, University of Washington, Seattle (E.A.R.).
From the 1991 RSNA scientific assembly. Received June 1, 1993; revision requested June 25; revision received August 5; accepted September 7. Supported by the National Institutes of Health grants CA49697-02 and 5 ROI CA55325 and by the National Science Foundation grant DMS-9101528. **Address reprint requests to** R.M.G.

Abbreviations: bpp = bits per pixel, PTSVQ = pruned tree-structured vector quantization, PVP = predictive value positive, ROC = receiver operating characteristic.

pixel block, its index in binary form provides a digital representation for the original pixel block. These indexes represent the image in compressed form. When the image is to be viewed, decompression is a memory lookup to retrieve the code word corresponding to each index. If the input pixel block has N pixels (typically $2 \times 2 = 4$) and the binary index consists of r binary numbers, then the bit rate of the compression algorithm is $R = r/N$ bits per pixel (bpp). If the original image has 12 bpp (each pixel requires 12 bits for representation) and the compression algorithm produces an image of R bits per pixel, the compression ratio is given as 12:R.

There are two key issues in the design and application of a vector quantizer: How does one choose a good code book, and how does the encoder find a good code word from the code book? The code book can be generated from training images with statistical clustering techniques. Training images are divided into small blocks (eg, 2×2 pixels) called training vectors. Vectors that are most representative of pixel blocks found in that type of image are chosen with the recursive Lloyd clustering algorithm (18,19). Once the code book has been created, new images are encoded according to a nearest neighbor or minimum distortion rule. The encoder searches the entire code book to find the best possible match to the original pixel block, where "best" means the code word yielding the smallest mean-squared error distortion. This approach is referred to as full-search quantization. A faster and more flexible method is to structure the code book as a binary tree. A binary tree begins with a root node that has two branches (daughter nodes). Each of these nodes either has two further daughter nodes or is a terminal node (leaf). A pixel block (code word) is associated with each node of the tree. An original pixel block is encoded into a leaf code word with a series of comparisons. Beginning with the root node, the original pixel block is compared with the two vectors associated with the daughter nodes. The encoder chooses the best match, proceeds to that node, and compares it against the next pair of vectors. It continues in this fashion until a leaf is reached. The path traversed through the tree from the root node to the leaf is the binary index of the selected code word. The tree-structured encoder is far less complex than the full-search encoder because it makes R binary comparisons instead of 2^R to select a good code word.

A tree-structured code book can be designed by combining ideas from classification trees (20) with the Lloyd clustering algorithm (19,21,22). A binary tree associates the centroid of all vectors with the root node and creates new code words by successively splitting the leaves of the tree. Splitting a leaf replaces one code word with a choice of two possible code words for representing a vector. Average distortion decreases, but the number of bits required to index the tree increases, that is, a larger tree produces a lower distortion but a higher bit rate. Without consideration of the effects on possible future splits, the growing algorithm finds the leaf that, when split, yields the largest decrease in distortion per increase in average length. This greediness is offset in part by growing the tree beyond the target rate and then optimally pruning it back to maximize the compression-to-distortion ratio. The result is an unbalanced tree that implements a variable rate code. The average rate in bits per pixel is the average number of binary decisions needed to reach a leaf of the tree. This algorithm is known as pruned tree-structured vector quantization (PTSVQ).

Vector quantizer performance can be enhanced by incorporating prediction into the compression. Instead of encoding the pixel blocks directly, the blocks are first expressed as predictions that are based on neighboring pixels (19,23–25). The difference between the predicted and the actual vectors is called the residual vector, and it is this residual vector that is quantized rather than the pixel block itself. To reconstruct the image for viewing, the decoder retrieves the code word for the residual block from memory and adds that code word representation to the block prediction that is based on the vectors encoded up to that point. The overall system is referred to as predictive PTSVQ. The code was implemented in C on a SPARC 1+ workstation (Sun Microsystems, Mountain View, Calif) (19,21,22).

Image Selection and Evaluation

For each disease category (mediastinal adenopathy and pulmonary nodules), 20 training images and 30 test images were selected. As discussed earlier, the training images are used only for developing the code book. The judging radiologists saw only the test images, and, to avoid giving the code book an unfair advantage, patient studies represented in the training set were not used as test images. Images consisted of isolated 5- and 10-mm sections from clinical thoracic studies performed with a CT scanner (9800; GE Medical Systems, Milwaukee, Wis) with the following parameters: 120 kV; 140 mA; scan time, 2 seconds per section; bore size, 38 cm; field of view, 32–34 cm; resolution, 512×512 pixels; and pixel depth, 12 bits. Although no formal research was undertaken to accurately determine what constitutes "representative" CT images, two radiologists (H.C.D., C.J.B.) were consulted concerning the typical range of appearance of adenopathy and nodules that occurs in daily clinical practice. The training and test images were chosen to be approximately representative of this range and included images of both normal and abnormal chests. The lung nodules ranged in size from 0.4 to 3.0 cm, with almost all nodules between 0.4 and 1.5 cm, and the abnormal lymph nodes were between 0.6 and 3.5 cm. The study had a lower percentage of normal chest images than would be encountered in daily practice.

The training images were used to design a compression system for each disease category. In each case, a binary tree of the residual training vectors was grown to a depth of 2.8 bpp and then pruned back to six different target bit rates, ranging from 2.65 to 0.55 bpp. The digital data for the test images from the adenopathy series ($N = 30$) and lung series ($N = 30$) were sent over the network from the scanner to the workstation, where each image was compressed and reconstructed at six different compression levels. The data for the compressed versions were sent back to the GE 9800 scanner to be displayed with a standard 12-on-1 format on 14×17-inch film. Each film had six images from the lymph node series displayed with soft-tissue window settings and six images from the nodule series displayed with lung window settings.

The images were evaluated independently by three radiologists with three separate sets of films. The judges (including C.J.B.) were blinded in that no information concerning patient study or compression level was indicated on the film. For each test image, each radiologist viewed the uncompressed original and five of the six compressed versions of each section. One hundred twenty images were judged per session in three sessions that were at least 2 weeks apart. At each session, each judge saw each section at two of the seven levels of compression and the two images never appeared with fewer than three pages separating them. Abnormalities were marked directly on the hardcopy films with a grease pencil, although mediastinal lymph nodes were not marked unless their smallest cross-sectional diameter measured 10 mm or greater. The viewing time and distance and the lighting conditions were not constrained, but the judges were asked to view the pages in the given randomized order and not to return to previously viewed pages. These design efforts were undertaken both to simulate ordinary diagnostic tasks and to minimize learning about particular sections. The radiologists were not specially trained for this judging task, nor were their responses calibrated to any scale, as the goal of this study was specifically to evaluate compression performance in the context of radiologists performing tasks that resembled their everyday work.

Statistical Analysis

To quantify the accuracy of diagnosis for subsequent statistical analysis, it was necessary to establish a standard of reference for each image that could be used to compare readings from the compressed versions. Such a standard of reference is often referred to in the statistical literature as a "gold" standard. One solution is to use each judge's readings of an original (uncompressed) image as the standard for the readings of that same judge from the compressed versions of that same image. We refer to this as a personal standard of

reference. Another choice is to use the consensus of the three judges on the 60 original images as they had been evaluated during the judging sessions. Use of the consensus standard of reference entails having to eliminate sections with irreconcilable differences. Among the 60 images in the study, there were 12 for which the judges could not agree on which structures were abnormal on the original, even when they discussed the images together. The personal standard of reference allows all of the images to be used, but, as discussed later, the disadvantage is that it defines the original images as having perfect diagnoses, and therefore precludes comparison of the compressed images with the uncompressed originals.

For each compressed image, one can count for each judge the number N_{TP} of abnormalities that matched those identified on the standard of reference (true-positive findings), the number of marks N_{FP} indicating anomalies that were considered normal tissue on the standard of reference (false-positive findings), and the number N_{FN} of anomalies indicated on the standard of reference that were not marked. The standard of reference provides the total number $N_T = N_{TP} + N_{FN}$ of anomalies on that image for that judge. The sensitivity or true-positive rate is defined by N_{TP}/N_T and has the interpretation of being the probability that an abnormality is detected if it is present. The predictive value positive (PVP) is defined by $N_{TP}/(N_{TP} + N_{FP})$ and is the probability that something is actually an abnormality if it has been marked as one (26).

The performance of the compression algorithm is indicated with scatterplots, quadratic spline fits, and associated confidence regions for sensitivity and PVP versus bit rate data. Regression splines are simple and flexible models for tracking data; the fitting is by least squares analysis (27). The fitting tends to be "local" in that the fitted average value at a particular bit rate is influenced primarily by observed data at nearby bit rates. Ours are quadratic splines with single knots at 1.5 bpp. The curve is continuous and has a continuous derivative. It has four unknown parameters and can be expressed as $y = a_0 + a_1x + a_2x^2 + b_2\,[max(0, x - 1.5)]^2$, where x is bit rate and the *max* function selects the larger of its two arguments. The underlying probability model that governs the 450 observed values of y (= sensitivity or PVP) is as follows. The random vector of quadratic spline coefficients (a_0, a_1, a_2, b_2) has a single realization for each (judge, image) pair. As the bit rate varies, the values for the chosen five compression levels plus independent mean 0 noise are observed. The expected value of y is $E(y) = E(a_0) + E(a_1)x + E(a_2)x^2 + E(b_2)[max(0, x - 1.5)]^2$, where the expectation (E) is with respect to the unconditional distribution of the random vector (a_0, a_1, a_2, b_2). Associated with each spline fit is the residual root mean square, an estimate of the standard deviation of the individual measurements from an analysis of variance of the spline fits.

The standard method for computing simultaneous confidence regions for such curves is the S or Scheffé method (28), which is valid under Gaussian assumptions that are not true for our data. Therefore, we use the statistical technique called the "bootstrap" (29,30), specifically a variation of the "correlation model" (31) that is related to the bootstrap-based prediction regions of Olshen et al (32). We denote the estimate of PVP for the lung study at a bit rate of *bpp* by $\hat{E}[y(bpp)]$ and denote the four-dimensional vector of estimated least squares coefficients $(\hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{b}_2)^t$ by $\hat{E}(a)$. The residual mean square that results from the fit is written S^2.

The bootstrapping is conducted by first drawing a sample of size three, with replacement from our group of three judges. This bootstrap sample may include one, two, or all three of the judges. For each chosen judge (including multiplicities, if any), we draw a sample of size 30, with replacement from the set of 30 original images. It can be shown that typically about $63\% = [100(1 - e^{-1})]\%$ of the images will appear at least once in each bootstrap sample of images. For each chosen judge and original image, we include all five of the observed values of y in the bootstrap sample. The intuitive justification for the bootstrap is that the bootstrap sample of 450 (bpp, y) pairs bears the same relationship to the real sample of 450 as the real sample does to "nature."

In this model, the bit rate is treated as a nonrandom predictor that we control by choosing a pruned subtree, and the judges and images are "random effects" because our three judges and 30 images have been sampled from arbitrarily large numbers of possibilities. For each bootstrap sample, we computed the bootstrap design matrix D^* of size 450×4 and the bootstrap estimates $\hat{E}^*[Y(bpp)]$, $\hat{E}^*(a)$, and S^{2*}. This process of sampling and computing estimates was repeated 1,000 times, and the values were saved. For a particular bit rate, the confidence region is in the form $\hat{a}_0 + \hat{a}_1(bpp) + \hat{a}_2(bpp)^2 + \hat{b}_2\,[max(0, bpp - 1.5)]^2 \pm \sqrt{F}\sqrt{S^2}\sqrt{d^t(D^tD)^{-1}d}$, where $d^t = \{1, bpp, [bpp]^2, [max(0, bpp - 1.5)]^2\}$, D is the 450×4 design matrix corresponding to the real observed values, and $\sqrt{F}$ is chosen so that for 95% of the bootstrap samples $[\hat{E}^*(a) - \hat{E}(a)]^t [D^{*t}D^*] [\hat{E}^*(a) - \hat{E}(a)] \leq FS^{2*}$.

Although the scatterplots, spline fits, and confidence regions provide excellent summaries of the effect of the algorithm on the judge's performances, the Behrens-Fisher t test (33) is used to quantify statistically significant differences in performance at various bit rates. A variation of the two-sample t test, the Behrens-Fisher test, accounts for inequality of variances in the different test images; the test is quite robust when data are not Gaussian, as our data clearly are not. The use of this statistic is illustrated with the following example. Suppose judge 1 has judged N lung images at both levels A and B. These images can be divided into nine groups, according to whether the original image for that judge contained zero, one, . . . or eight abnormalities. Let N_i be the number of images in the ith group. Let Δ_{ij} represent the difference in sensitivities (or PVP) for the jth image in the ith group seen at levels A and B. Let $\bar{\Delta}_i$ be the average difference as in the following equation:

$$\bar{\Delta}_i = \frac{1}{N_i}\sum_j \Delta_{ij}.$$

We define the following:

$$S_i^2 = \frac{1}{N_i - 1}\sum_j (\Delta_{ij} - \bar{\Delta}_i)^2,$$

and then the Behrens-Fisher t statistic is given by this equation:

$$t_{BF} = \frac{\sum_i \bar{\Delta}_i}{\sqrt{\sum_i \frac{S_i^2}{N_i}}}.$$

Our Δ_{ij} are fractions with denominators of not more than eight; thus, they are utterly non-Gaussian. Therefore, computations of attained statistical significance (P values) are based on the restricted permutation distribution of t_{BF}. For each of the N images, we can permute the results from the two levels (A → B and B → A) or not. There are 2^N points possible in the full permutation distribution, and we calculate t_{BF} for each one. If there were no difference between the bit rates, it should not matter whether we use level A minus level B or vice versa to compute the differences Δ_{ij}, and we would not expect the "real" t_{BF} to be an extreme value among the 2^N values of t_{BF} that correspond to the full permutation distribution. If k is the number of permuted t_{BF} values that exceed the real one, then $(k + 1)/2^N$ is the attained one-sided statistical significance level for the test of the null hypothesis that the lower bit rate performs at least as well as the higher one. The one-sided test of statistical significance is chosen to be conservative and to argue most strongly against compression.

Because the radiologists would see an image at six different compression levels during the course of the study, we needed to ascertain whether learning effects were statistically significant. Learning and fatigue are processes that might change the score of an image, depending on when it was seen. In each session, each image was seen at two levels, and the ordering of the pages ensured that they never appeared with fewer than three pages separating them. To examine the possibility of intrasession learning and fatigue, we examined the paired data in which the first occurrence of a given image in a session was paired with the second occurrence of that same image (at a different compression level) in the same session. Each image in the pair was either "perfect" (sensitivity = 1, PVP = 1) or not. There were thus four types of pairs: those with both members perfect, those with only the first occurrence perfect, those with only the second occurrence perfect, and those with neither one perfect. In the McNemar analysis (34), we concern ourselves with two of the four types: those pairs in which

the members differ. If it did not matter whether an image was seen first or second—then conditional on the numbers of the other two types—each would have a binomial distribution with a parameter of ½.

As an example of the calculation, judge 1, in evaluating lung nodules over the course of three sessions, saw 89 pairs of images, in which an image seen at one compression level in a given session was paired with the same image seen at a different level in the same session. Of the 89 pairs, both images in the pair were judged perfectly 55 times; both images were judged incorrectly 11 times. We concern ourselves with the other 23 pairs: 13 times the image seen first was incorrect, whereas the second one was correct; 10 times the image seen second was incorrect, when the first one was correct. The probability that a fair coin flipped 23 times will produce a heads/tails split at least as great as 13 to 10 is 0.68; thus, this result is not statistically significant. In addition to its use in analyzing learning effects, the McNemar test was also used to compare the compression levels in a paired fashion, as a secondary analysis to the Behrens-Fisher statistic. It differs from the Behrens-Fisher statistic in that it combines sensitivity and PVP scores and it dichotomizes the results into perfect and not perfect (35,36).

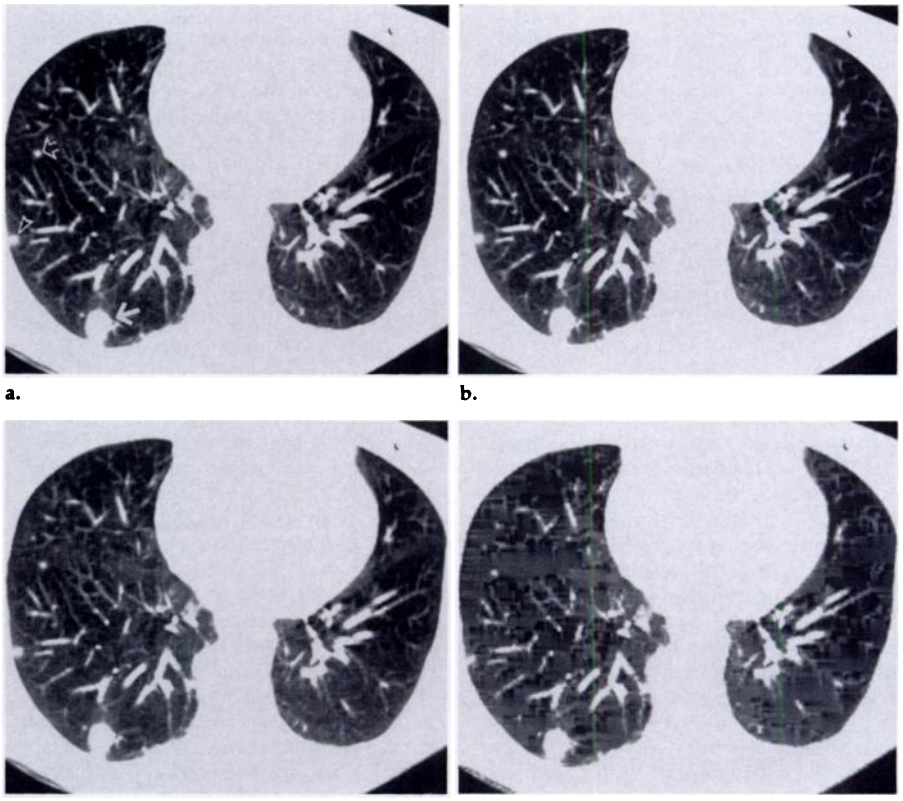

Figure 1. **(a)** Original (uncompressed) chest CT image demonstrates three nodules (arrows and arrowhead) according to all three personal standards. One of the nodules (open arrow) was overlooked by two judges at level A; another (arrowhead) was missed by one judge at level A. **(b)** Level E image has 2.19 bpp and 5.5:1 compression. **(c)** Level C image has 1.33 bpp and 9:1 compression. **(d)** Level A image has 0.57 bpp and 21:1 compression.

RESULTS

The six bit rates achieved for compressed images of both lung and mediastinal images were nearly identical: 0.57, 1.18, 1.33, 1.79, 2.19, 2.63 bpp in the lung and 0.56, 1.18, 1.34, 1.79, 2.20, 2.64 bpp in the mediastinum. These levels are subsequently referred to as A, B, C, D, E, and F, respectively, and correspond approximately to compression ratios of 21:1, 10:1, 9:1, 7:1, 5.5:1, and 4.5:1. Because PTSVQ produces variable rate compression, the 30 images in a given disease category will compress to a set of bit rates clustered around each target bit rate. The achieved rates represent an average across the 30 images for a given target bit rate. Examples of images from both disease categories at various compression rates are shown in Figures 1 and 2.

The Table shows the numbers of original test images (total, 30) that contain the listed number of abnormalities for each disease type according to each judge. Also, the rows marked "All" show the number of original test images (total, 24) that contain the listed number of abnormalities according to the consensus standard. Simple χ^2 tests for homogeneity show that for both the lung and the mediastinum judges did not differ beyond chance from equality in the numbers of abnormalities they found. In particular, if for the lung we categorize abnormalities found as 0, 1, 2, 3, or at least 4, the χ^2 statistic is 3.16 (on eight degrees of freedom). Six cells have expectations below 5—a traditional concern—but an exact test would not have a different conclusion. Similar comments apply to the mediastinum, where the χ^2 value (on six degrees of freedom) is 8.83. However, the Table does not fully indicate the variability among the judges. For example, the Table shows that each judge found six lung nodules on an original test image only once. However, this did not occur with the same test image for all three judges.

With sensitivity and PVP values defined relative to the consensus standard of reference, no differences were found between the original images and the three least compressed levels (levels D, E, and F) at the 5% significance level, whether the findings of the judges were pooled or evaluated separately. With the personal standard of reference, as discussed later, it is not appropriate to compare the compressed levels with the original images. When used to compare the compressed levels among themselves, the sensitivity defined relative to the personal standard of reference showed that level A was statistically significantly different from most of the other levels for two judges in evaluating the mediastinum ($P < .001$), and level B was different from level F for one judge in evaluating the lung ($P = .031$). There were no statistically significant differences in PVP between any levels.

Figure 3 shows the sensitivity of detecting pulmonary nodules and mediastinal lymph nodes as a function of bit rate for images pooled across judges, sessions, and compression levels ($n = 450 = 30$ images $\times$ 3 judges $\times$ 5 compressed versions seen for each image). The PVP for both nodules and lymph nodes is shown in Figure 4. Values of sensitivity and PVP are simple fractions such as ½ and ⅔ because there are at most eight abnormalities found in each image and generally fewer than that. Average sensitivity is above 0.85 (out of a perfect score of 1.0) for all bit rates except level A at 0.56 bpp, whereas average PVP is above 0.85 at all bit rates. Figures 5 and 6 show the sensitivity and PVP outcomes for the three judges considered separately. Figures

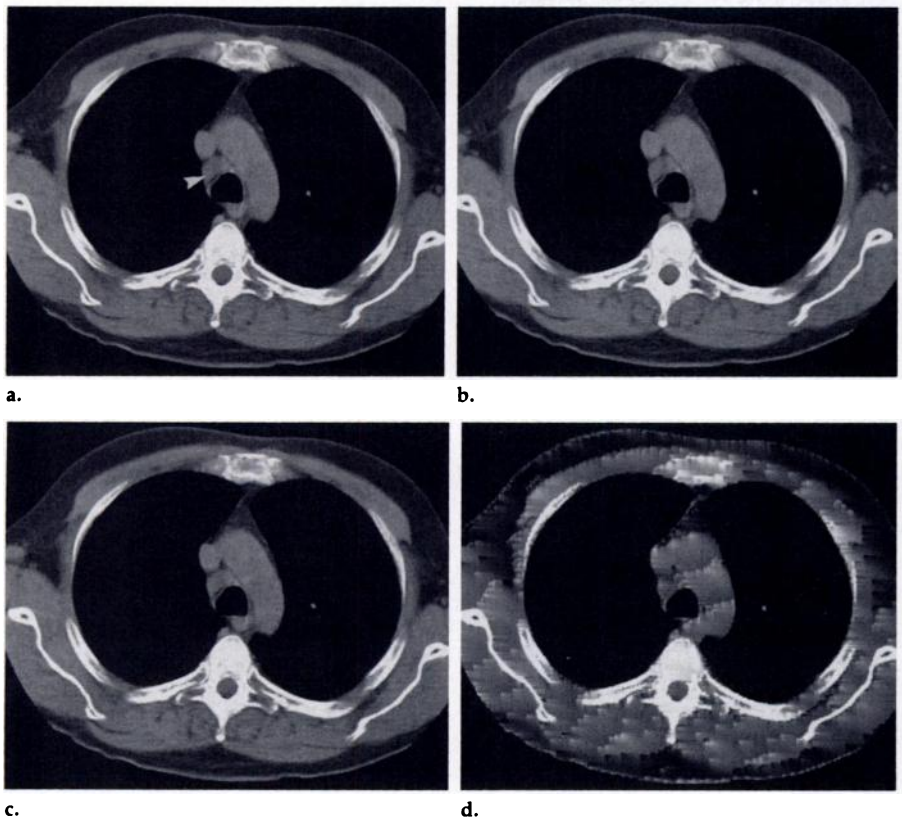

Figure 2. **(a)** Original (uncompressed) chest CT image demonstrates one mediastinal lymph node (arrowhead), which was correctly identified by all judges at all levels of compression. **(b)** Level E image has 2.20 bpp and 5.5:1 compression. **(c)** Level C image has 1.34 bpp and 9:1 compression. **(d)** Level A image has 0.56 bpp and 21:1 compression.

Number of Test Images That Contain the Listed Number of Abnormalities

Type	Judge	No. of Abnormalities								
		0	1	2	3	4	5	6	7	8
Lung	1	3	11	7	6	2	0	1	0	0
Lung	2	4	9	10	4	2	0	1	0	0
Lung	3	3	8	8	5	2	2	1	0	1
Lung	All	4	9	4	5	2	0	0	0	0
Mediastinum	1	3	14	7	6	0	0	0	0	0
Mediastinum	2	2	22	2	4	0	0	0	0	0
Mediastinum	3	3	22	4	1	0	0	0	0	0
Mediastinum	All	3	17	2	2	0	0	0	0	0

3–6 display sensitivity and PVP defined relative to the personal standards.

DISCUSSION

This study departs from the majority of similar studies in both the choice of compression algorithm and in the statistical methods used to evaluate the compressed images. The statistical techniques reported here can be applied to evaluate any compression technique. Virtually all reported studies of lossy compression of medical images use some form of transform coding based on the discrete cosine transform. This technique has been made into an international standard (37), and there are many software and hardware implementations. There are many variations and combinations of other techniques, including full-frame or large-block discrete cosine transforms combined with adaptive scalar quantization of transform coefficients and lossless coding such as the Huffman, Lempel-Ziv, and arithmetic codes (18,19,38).

Tree-structured vector quantization can be tailored to specific types of data to a greater or lesser degree by the choice of training images. For example, use of training images from a single scanner that correspond to a single part of the body would cause the tree-structured vector quantization to perform best for other images from the same scanner showing the same body features. With a more restrictive set of training images, the code book can be tailored to specific diseases or even to specific patients. With a less restrictive set, it can be made appropriate for different scanners and for many anatomic locations. Training images can in fact be dispensed with altogether, as the clustering algorithm can be run on mathematic models of various types of data. In this study, we used restrictive sets of training images and therefore produced task-specific code books. Task specificity precludes the use of the results of this study to conclude with certainty that these code books will yield similar diagnostic accuracy for all possible abnormalities of the imaged organs.

This study had the modest goal of developing methods for examining diagnostic accuracy for specific and well-defined diagnostic tasks. This is the first step toward demonstration that a compression technique causes no loss of accuracy for a variety of diagnostic tasks on a common image, which is a necessary goal if compressed images are to be used for initial diagnosis and not just for recall, archival, and educational purposes. Such an extension of our methods is straightforward, with suitable expansion of the training set, use of universal or adaptive vector quantizers (19), and modification of the clinical simulations to detect multiple or unspecified abnormalities.

No clinical experiment, however, can guarantee that there is no loss of accuracy for all conceivable abnormalities, including those not yet discovered. This shortcoming is shared by all compression schemes, including JPEG (Joint Photographic Experts Group)– and other discrete cosine transform–based coding schemes, that allow the user to specify quantization tables and therefore can also be tailored to be more or less task specific. It is also implicit in traditional ROC- and analysis of variance–based analyses of diagnostic accuracy for specific binary detection tasks.

To evaluate the quality of lossy compressed images in the context of a diagnostic application, it is necessary to quantify the consistency of abnormalities observed on compressed images and that observed on the original uncompressed images. What is important is not whether compressed images allow for completely accurate diagnoses, but rather whether they

allow for accurate diagnoses at least as often as do original images. The personal standard of reference defines a judge's reading on an original image to be perfect and uses that reading as the basis of comparison for the compressed versions of that image. With any random noise in the judging process, compressed images cannot be as accurate as originals. The presence of a substantial noise component is suggested by the fact that there were several images on which a judge changed his or her diagnosis back and forth, marking, for example, one lesion on the original image as well as on levels B and E, and marking two lesions on levels A, D, and F. When such cases are compared with a consensus standard of reference, some of the times the consensus will determine that there is only one lesion and the original image scores perfectly.

For other images exhibiting the same kind of fluctuation, the consensus standard may cause a decision of two lesions and the original image will have a sensitivity of only 0.5. Thus, what appears to be noise in the diagnosis will tend not to favor the original images or any particular compression level on the average. With the personal standard, however, the decision based on the original image is always correct, so that the random noise is interpreted as incorrect diagnoses on the compressed versions. Because the compressed levels have this severe disadvantage, the personal standard of reference is primarily useful for comparing the six compressed levels among themselves.

The consensus standard has some biases also, but these are small in comparison. Because the consensus was achieved for only 48 original images of the 60 studied, there is an unknown bias that results from having 12 test images eliminated from the study. Although these 12 images were clearly more controversial and difficult to diagnose than the others, it cannot be said whether the removal of diagnostically controversial images from the study biases the results in favor of compression or against it. Their failure to have a standard defined was based only on the uncompressed versions, and it cannot be said a priori that the compression algorithm would present more difficulty in compressing such images. The consensus, when achieved, could be attained either by initial concordance among the readings of the three radiologists or by subsequent discussion of the readings, during which one or more of the judges might change his or her decision. The consensus was clearly more likely to be attained for those original images for which the judges were in perfect agreement initially and thus for which the original images would have perfect diagnostic accuracy relative to that standard. Therefore, this standard has a slight bias favoring the original images also, which is thought to help make the study safely conservative and not unduly promotional of our compression techniques.

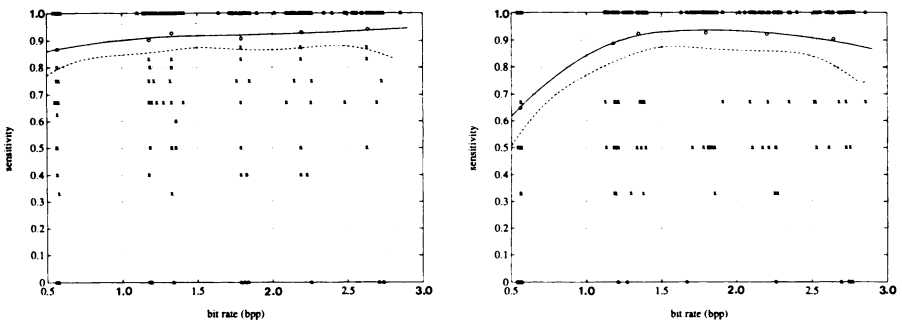

Figure 3. Sensitivity for detecting abnormalities on 450 images as a function of bit rate. Graphs show sensitivity for detecting pulmonary nodules with a root mean square (estimate of standard deviation) of 0.216 **(a)** and for lymph nodes with a root mean square of 0.262 **(b).** x = data points (sensitivity, bit rate) for entire data set (all images, judges, and compression levels), o = the average of the x's for each target bit rate, solid curves = quadratic splines fit to the data with a single knot at 1.5 bpp, dashed and dotted curves = two-sided 95% confidence regions.

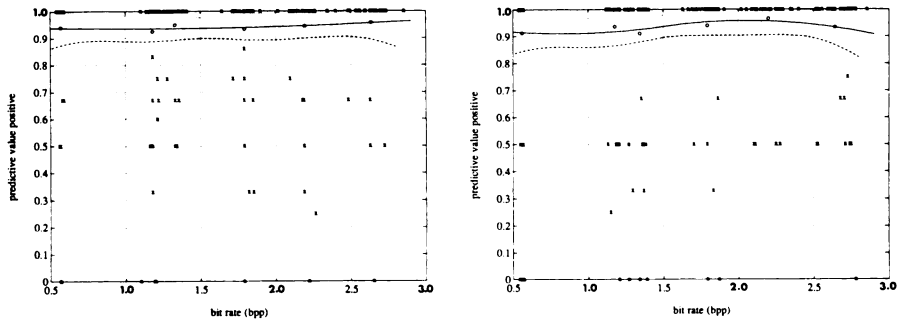

Figure 4. PVP for detecting abnormalities on 450 images as a function of bit rate. Graphs show PVP for detecting pulmonary nodules with a root mean square (estimate of standard deviation) of 0.181 **(a)** and for lymph nodes with a root mean square of 0.203 **(b).** See Figure 3 for explanations.

Because the personal standard has the advantage of using all of the images in the study and the consensus standard has the advantage of having little bias between original and compressed images, we can capitalize on both sets of advantages with a two-step comparison. Sensitivity and PVP values relative to the consensus standard show that there are no substantial differences between the slightly compressed images (levels D–F) and the original images. This is true for both disease categories, judges evaluated separately and pooled, and use of both the Behrens-Fisher test to examine the sensitivity and PVP separately and the McNemar test to combine them. With this assurance, the personal standard can then be used to look for differences between the more compressed levels (A–C) and the less compressed levels (D–F). The most compressed level (A: 0.56 bpp, 21:1 compression ratio) is unacceptable, as observations made with these images were statistically significantly different from those with less compressed images for two judges ($P < .001$). Level B (1.18 bpp) is also unacceptable, although barely, because the only statistically significant difference was between the sensitivities at levels B and F for a single disease category and a single judge ($P = .031$). No differences were found between level C and the less compressed levels, and there were no statistically significant differences at the 5% significance level between levels D, E, and F.

One concern of this study is the question of whether a judge could remember and benefit from what was seen on the earlier of two images dur-

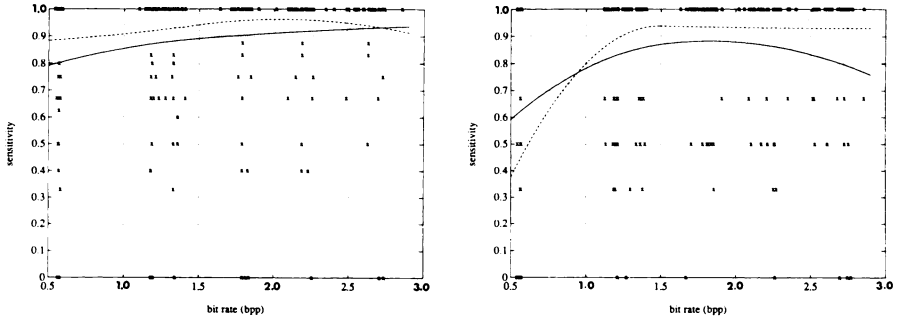

Figure 5. Sensitivity for detecting abnormalities as a function of bit rate. Graphs show sensitivity for detecting pulmonary nodules **(a)** and mediastinal lymph nodes **(b).** x = data points (sensitivity, bit rate) for entire data set (all images, judges, and compression levels). The three curves are quadratic splines fit to the data separately for the three judges.

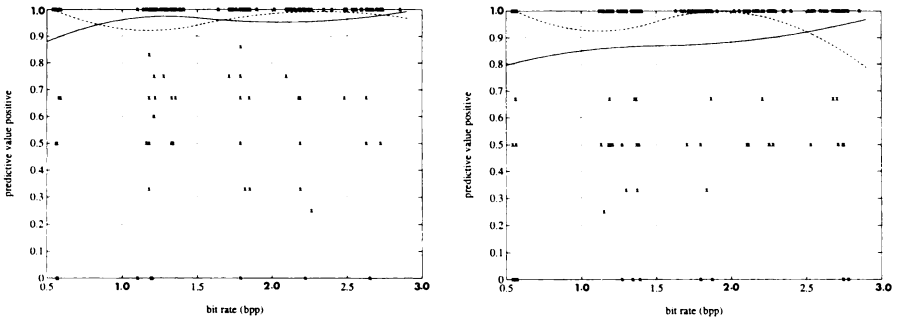

Figure 6. PVP for detecting abnormalities as a function of bit rate. Graphs show PVP for detecting pulmonary nodules **(a)** and mediastinal lymph nodes **(b).** See Figure 5 for explanation of x. The three curves are quadratic splines fit to the data separately for the three judges.

ing the same session. We examined the possibility of intrasession learning with the McNemar test, as described previously. No differences at the 5% statistical significance level were found between images seen first and those seen second, with the judges considered separately or pooled, for either disease category, and with either the personal or consensus standard (P values ranged from .06 to 1.0). A regression analysis with the actual sensitivity and PVP observations similarly indicated that page order and session order had no statistically significant effect at the 5% significance level on the diagnostic result.

Many studies of diagnostic accuracy of lossy image compression have argued that compression ratios from 5:1 to 29:1 can be obtained with no loss in diagnostic accuracy (9–11,13–18). Virtually all of these studies have been based on ROC analysis (39,40). ROC curves are plots of sensitivity (true-positive rate) against specificity (true-negative rate). ROC analysis has three drawbacks. It requires confidence rankings, which are a departure from normal diagnostic practice. In a typical ROC study, judges are asked to decide whether an abnormality is present on an image and to assign an integer rating (one to five) to their level of confidence in that decision. Second, ROC analysis is not location specific. A case in which an observer overlooks the lesion that is present on an image but mistakenly identifies some noise feature as a lesion on that image would be scored as a true-positive event. The third and most serious drawback is that traditional ROC analysis is not appropriate for nonbinary detection tasks. There is no difficulty assigning a numeric value to sensitivity, since the number of abnormalities present can be ascertained. Measuring the specificity, however, requires one to know the number of abnormalities absent on an image. In truly binary diagnostic tasks such as pneumothorax detection, if the image is normal, exactly one abnormality is absent and specificity is measurable. For nonbinary tasks in which there can be many abnormalities, however, specificity does not make sense because it is not possible to say how many abnormalities are absent. It does not seem realistic to segment the images with the proviso that each segment has to contain either zero abnormalities or one. For nonbinary tasks then, PVP can be used to fill the role of specificity in measuring false-positive reporting. A judge who is too aggressive in finding an abnormality could have high sensitivity at the expense of low PVP, whereas a judge who is too stringent about what defines an abnormality could have a high PVP at the expense of low sensitivity. Just as an image is judged perfectly if both sensitivity and specificity (when possible to measure both) are one, an image is also judged perfectly if both sensitivity and PVP are one.

The free-response ROC observer performance experiment, a version of the standard ROC that has not yet been widely used (41–44), allows an arbitrary number of abnormalities per image, and the observer indicates his or her perceived locations and a confidence rating for each. Although free-response ROC resolves the binary task limitations and location insensitivity of traditional ROC, free-response ROC does retain the constrained five-point integer rating system for observer confidence and makes certain normality assumptions about the resultant data. In our study, use of the Behrens-Fisher t statistic on the sensitivity and PVP scores allowed the radiologists to make diagnoses in their usual fashion of locating and marking all visible lesions, with no need for associated confidence rankings and no normality assumptions about the resultant data.

In summary, we have introduced the appropriate statistical tools for a novel approach to testing diagnostic accuracy. The approach is valid for nonbinary detection tasks and permits a realistic simulation of ordinary diagnostic tasks. We have used this approach to demonstrate that predictive PTSVQ provides lossy compression of images to bit rates as low as 1.3 bpp (compression ratios of up to 9:1) while maintaining diagnostic accuracy of pulmonary nodules and mediastinal lymph nodes on thoracic CT images. We believe that our approach represents a useful and appropriate way of analyzing this type of nonbinary detection task and that the compression technique considered shows promise as a useful tool for image management in digital medical imaging systems. ■

Acknowledgments: The authors gratefully acknowledge the assistance of D. Brown, MD, and D. Lentz, MD, in judging the images. The authors thank the anonymous reviewers for their comments and suggestions.

References

1. Frost MM Jr, Honeyman JC, Staab EV. Image archival technologies. RadioGraphics 1992; 12:339–343.
2. Boncelet C, Cobbs J, Moser A. Error free compression of medical x-ray images. Proceedings of the Conference on Visual Communications and Image Processing. Proc SPIE 1988; 1001:269–276.
3. Rhodes M, Quinn J, Silvester J. Locally optimal run-length compression applied to CT images. IEEE Trans Med Imaging 1985; 4:84–90.
4. Redfern A, Hines E. Medical image data compression techniques. Proceedings of the Third International Conference on Image Processing and Its Applications. Vol 307. Coventry, England: IEE, Michael Faraday, 1989; 558–562.
5. Bramble J. Comparison of information-preserving and information-losing data-compression algorithms for CT images. Radiology 1989; 170:453–455.
6. Ishigaki T, Sakuma S, Ikeda M, Itoh Y, Suzuki M, Iwai S. Clinical evaluation of irreversible image compression: analysis of chest imaging with computed radiography. Radiology 1990; 175:739–743.
7. Chan K, Lou S, Huang H. Full-frame transform compression of CT and MR images. Radiology 1989; 171:847–851.
8. Lo S, Huang H. Radiological image compression: full-frame bit-allocation technique. Radiology 1985; 155:811–817.
9. Sayre J, Aberle DR, Boechat MI, et al. Effect of data compression on diagnostic accuracy in digital hand and chest radiography. Proceedings of Medical Imaging VI: Image Capture, Formatting, and Display. Vol 1653. Newport Beach, Calif.: Society of Photo-optical Engineers, 1992; 232–240.
10. MacMahon H, Doi K, Sanada S, et al. Data compression: effect on diagnostic accuracy in digital chest radiographs. Radiology 1991; 178:175–179.
11. Wilhelm P, Haynor DR, Kim Y, Riskin EA. Lossy image compression for digital medical imaging system. Opt Eng 1991; 30:1479–1485.
12. Chen J, Flynn M, Gross B, Spizarny D. Observer detection of image degradation caused by irreversible data compression processes. In: Proceedings of Medical Imaging V: Image Capture, Formatting, and Display. Vol 1444. San Jose, Calif: Society of Photo-optical Engineers, 1991; 256–264.
13. Fiete R, Barrett H, Cargill E, Myers KJ, Smith WE. Psychophysical validation of the Hotelling trace criterion as a metric for system performance. Proc SPIE Med Imaging 1987; 767:298–305.
14. Fiete R, Barrett HH, Smith WE, Meyers KJ. The Hotelling trace criterion and its correlation with human observer performance. J Opt Soc Am [A] 1987; 4:945–953.
15. Lee H, Rowberg AH, Frank MS, Choi HS, Kim Y. Subjective evaluation of compressed image quality. In: Proceedings of Medical Imaging VI: Image Capture, Formatting, and Display. Vol 1653. Newport Beach, Calif: Society of Photo-optical Engineers, 1992; 241–251.
16. Barrett HH, Gooley T, Girodias K, Rolland J, White T, Yao J. Linear discriminants and image quality. In: Proceedings of the 1991 International Conference on Information Processing in Medical Imaging. Wye, United Kingdom: Springer-Verlag, 1991; 458–473.
17. Bramble J, Cook L, Murphey M, Martin N, Anderson W, Hensley K. Image data compression in magnification hand radiographs. Radiology 1989; 170:133–136.
18. Rabbani M, Jones PW. Digital image compression techniques. Vol TT7. Tutorial texts in optical engineering. Bellingham, Wash: SPIE Optical Engineering, 1991; 102–128.
19. Gersho A, Gray RM. Vector quantization and signal compression. Boston, Mass: Kluwer Academic, 1992.
20. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Belmont, Calif: Wadsworth; 1984.
21. Chou PA, Lookabaugh T, Gray RM. Optimal pruning with applications to tree-structured source coding and modeling. IEEE Trans Inform Theory 1989; 35:299–315.
22. Riskin EA, Gray RM. A greedy tree growing algorithm for the design of variable rate vector quantizers. IEEE Trans Signal Process 1991; 39:2500–2507.
23. Lookabaugh T, Riskin EA, Chou PA, Gray RM. Variable rate vector quantization for speech, image, and video compression. IEEE Trans Comm 1993; 41:186–199.
24. Riskin EA, Lookabaugh T, Chou PA, Gray RM. Variable rate vector quantization for medical image compression. IEEE Trans Med Imaging 1990; 9:290–298.
25. Riskin EA. Variable rate vector quantization of images. PhD dissertation, Stanford University, Stanford, Calif, 1990.
26. Weinstein M, Fineberg H. Clinical decision analysis. Philadelphia, Pa: Saunders, 1980.
27. Powell M. Approximation theory and methods. Cambridge, England: Cambridge University Press, 1981.
28. Miller RG Jr. Simultaneous statistical inference. 2nd ed. New York, NY: Springer-Verlag, 1981.
29. Battilana C, Zhang H, Olshen R, Wexler L, Myers B. PAH extraction and the estimation of plasma flow in diseased human kidneys. Am J Physiol 1991; 30:F726–F733.
30. Efron B. The jackknife, the bootstrap, and other resampling plans. Philadelphia, Pa: Society for Industrial and Applied Mathematics, 1982.
31. Freedman D. Bootstrapping regression models. Ann Stat 1981; 9:1218–1228.
32. Olshen R, Biden E, Wyatt M, Sutherland D. Gait analysis and the bootstrap. Ann Stat 1989; 17:1419–1440.
33. Armitage P. Statistical methods in medical research. Oxford, England: Blackwell Scientific, 1971.
34. McNemar I. Note on the sampling errors of the differences between correlated proportions of percentages. Psychometrika 1947; 12:153–157.
35. Olshen RA, Cosman PC, Tseng C, et al. Evaluating compressed medical images. In: Proceedings of COMCON III, Victoria, British Columbia, Canada: UNLV Publications, 1991; 830–840.
36. Cosman P, Tseng C, Gray R, et al. Tree-structured vector quantization of CT chest scans: image quality and diagnostic accuracy. IEEE Trans Med Imaging 1993 (in press).
37. Wallace G. The JPEG still picture compression standard. Commun ACM 1991; 34:30–44.
38. Netravali AN, Haskell BG. Digital pictures: representation and compression. New York, NY: Plenum Press, 1988.
39. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978; 8:282–298.
40. Swets JA. ROC analysis applied to the evaluation of medical imaging techniques. Invest Radiol 1979; 14:109–121.
41. Egan J, Greenberg G, Schulman A. Operating characteristics, signal detectability, and the method of free response. J Acoust Soc Am 1961; 33:993–1007.
42. Bunch P, Hamilton J, Sanderson G, Simmons A. A free-response approach to the measurement and characterization of radiographic observer performance. J Appl Photogr Eng 1978; 4:166–171.
43. Chakraborty D. Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data. Med Phys 1989; 16:561–568.
44. Chakraborty D, Winter L. Free-response methodology: alternate analysis and a new observer-performance experiment. Radiology 1990; 174:873–881.