Quality evaluation for compressed medical images: Diagnostic Accuracy

Pamela Cosman, Robert Gray, Richard Olshen

1 Introduction

We examined in the previous chapter several common computable measures of image quality, as well as subjective quality ratings. While these quality measures are useful in many ways, for medical images one wishes a quality measure to take proper account of the diagnostic purpose of the image. The ROC methodology discussed in the previous chapter is one approach to this. In this chapter, we present several studies which attempt to evaluate diagnostic utility of the images more directly. The radiologists were not specially trained or calibrated in any way for these judging tasks, as the goal of these studies was specifically to evaluate compression performance in the context of radiologists carrying out tasks that resembled their everyday work. No constraints were placed on the viewing time, the viewing distance, or the lighting conditions. The judges were encouraged to simulate the conditions they would use in everyday work. The tasks were detection of lung nodules and mediastinal adenopathy in CT images, measurement of blood vessels in MR chest scans, and detection and management tasks in mammography. As we shall see, the results indicate that when these images are used in situations resembling everyday work, substantial compression can be applied without affecting the interpretation of the radiologist.

2 CT study: example of detection accuracy

The detection of abnormal lymphoid tissue is an important aspect of chest imaging. This is true especially for the mediastinum, the central portion of the chest that contains the heart, major blood vessels, and other structures. Abnormally enlarged lymph nodes, or lymphadenopathy, in the mediastinum can be caused by primary malignancy such as lymphoma, metastatic disease that results from the spread of breast or lung cancer through the lymphatics, tuberculosis, or non-infectious inflammatory diseases such as sarcoidosis. Typically radiologists can easily locate lymph nodes in a chest scan. The detection task is therefore to determine which of the located lymph nodes are enlarged.

The detection of lung nodules is another major objective in diagnostic imaging of the chest. A common cause of these nodules is malignancy, primary or metastatic. The latter, which spreads through the blood stream from a primary cancer almost anywhere in the body, can cause multiple nodules in one or both lungs. Other causes include fungal and bacterial infections, and noninfectious inflammatory conditions. Nodules range in size from undetectably small to large enough to fill an entire segment of the lung.

The compressed and original images were viewed by 3 radiologists. For each of the 30 images in a study, each radiologist viewed the original and 5 of the 6 compressed levels, and thus 360 images were seen by each judge. The judges were blinded in that no information concerning patient study or compression level was indicated on the film. Images were viewed on hardcopy film on a lightbox, the usual way in which radiologists view images. The "windows and levels" adjustment to the dynamic range of the image was applied to each image before filming. This simple contrast adjustment technique sets maximum and minimum intensities for the image. All intensities above the maximum are thresholded to equal that maximum value. All intensities below the minimum are thresholded to equal that minimum value. This minimum value will be displayed on the output device as black, and the maximum value will be displayed as white. All intensity values lying in between the minimum and maximum are linearly rescaled to lie between black and white. This process allows for more of the dynamic range of the display device (in this case, the film) to be used for the features of interest. A radiologist who was not involved in the judging applied standard settings for windows and levels for the mediastinal images, and different standard settings for the lung nodule images. The compressed and original images were filmed in standard 12-on-1 format on 14"$\times$ 17" film using the scanner that produced the original images.

The viewings were divided into 3 sessions during which the judges independently viewed 10 pages, each with 6 lung nodule images and 6 mediastinal images. The judges marked abnormalities directly on the films with a grease pencil, although mediastinal lymph nodes were not marked unless their smallest cross-sectional diameter measured 10 mm or greater. All judges were provided with their own copy of the films for marking. No constraints were placed on the viewing time, the viewing distance, or the lighting conditions; the judges were encouraged to simulate the conditions they would use in everyday work. They were, however, constrained to view the 10 pages in the predetermined order, and could not go back to review earlier pages. At each session, each judge saw each image at 2 of the 7 levels of compression (7 levels includes the original). The two levels never appeared on the same film, and the ordering of the pages ensured that they never appeared with fewer than 3 pages separating them. This was intended to reduce learning effects. Learning effects will be discussed in the next chapter. A given image at a given level was never seen more than once by any one judge, and so intra-observer variability was not explicitly measured. Of the 6 images in one study on any one page, only one image was shown as the original, and exactly 5 of the 6 compressed levels were represented. The original versions of the images are denoted 'g'. The compressed versions are 'a' through 'f'. The randomization follows what is known as a "Latin square" arrangement.

The consensus gold standard for the lung determined that there were, respectively, 4 images with 0 nodules, 9 with 1, 4 with 2, 5 with 3, and 2 with 4 among those images retained. For the mediastinum, there were 3 images with 0 abnormal nodes, 17 with 1, 2 with 2, and 2 with 3.

Once a gold standard is established, a value can be assigned to the sensitivity and the predictive value positive (PVP). The sensitivity and PVP results are shown graphically using scatter plots, spline fits, and associated confidence regions. The spline fits are quadratic splines with a single knot at 1.5 bits per pixel (bpp), as given in the previous chapter. The underlying probability model that governs the 450 observed values of y (= sensitivity or PVP) is taken to be 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. What is observed as the bit rate varies is the value for the chosen 5 compression levels plus independent mean 0 noise. 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 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 (RMS), 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 [20], which is valid under certain Gaussian assumptions that do not hold for our data. Therefore we use the statistical technique called "the bootstrap" [12, 4, 10, 11], specifically a variation of the "correlation model" [13] that is related to the bootstrap-based prediction regions of Olshen *et al.* [22]. We denote the estimate of PVP for the lung study at a bit rate *bpp* by $\hat{E}(y(bpp))$.

1. A quadratic spline equation can be written as

$$\hat{E}(y(bpp)) = a_0 + a_1 x + a_2 x^2 + b_2 (max(0, x - x_0))^2,$$

   where x_0 is the "knot" (in this study, x = bit rate and x_0 = 1.5 bpp). This equation comes from the linear model

$$\mathbf{Y} = \mathbf{D}\beta + \mathbf{e},$$

   with one entry of **Y** (and corresponding row of **D**) per observation. **D** is the "design matrix" of size 450 × 4. It has four columns, the first having the multiple of a_0 (always 1), the second the multiple of a_1 (that is the bit rate), and so on. We use $\hat{E}(\underline{a})$ to denote the 4-dimensional vector of estimated least squares coefficients:

$$\hat{E}(\underline{a}) = (\hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{b}_2)^t.$$

2. For a given bit rate b, write the row vector dictated by the spline as $\mathbf{d}^t = \mathbf{d}^t(b)$. Thus $\hat{E}(y(bpp)) = \mathbf{d}^t \hat{E}(\underline{a})$.

3. The confidence region will be of the form

$$\mathbf{d}^t\hat{E}(a) - S\sqrt{F}\,\sqrt{\mathbf{d}^t(\mathbf{D}^t\mathbf{D})^{-1}\mathbf{d}} \le y \le \mathbf{d}^t\hat{E}(a) + S\sqrt{F}\,\sqrt{\mathbf{d}^t(\mathbf{D}^t\mathbf{D})^{-1}\mathbf{d}},$$

where S is the square root of the residual mean square from an analysis of variance of the data. So, if $\mathbf{Y}$ is $n \times 1$ and β is $k \times 1$, then

$$S = \sqrt{\frac{1}{n-k}\,\|\mathbf{Y} - \mathbf{D}\hat{E}(a)\|^2}.$$

The region will be truncated, if necessary, so that always $0 \le y \le 1$.

4. The bootstrapping is conducted by first drawing a sample of size 3 with replacement from our group of 3 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 in the bootstrap sample all 5 of the observed values of y. The motivation for this bootstrap sampling is simple: the bootstrap sample bears the same relationship to the original sample that the original sample bears to "nature." We do not know the real relationship between the true data and nature; if we did, we would use it in judging coverage probabilities in Steps 7 and 8. However, we do know the data themselves, and so we can imitate the relationship between nature and the data by examining the observed relationship between the data and a sample from them [12, 10].

5. Each bootstrap sample $\mathbf{Y}^*$ entails a bootstrap design matrix $\mathbf{D}^*$, as well as corresponding $\hat{E}^*(a)$ and S^*. This bootstrap process will be carried out $n_b = 1000$ times.

6. For the j^{th} bootstrap sample compute the four new bootstrap quantities as in 5.

7. Compute for each $\sqrt{F}$

$$\hat{G}_B(\sqrt{F}) = (n_b)^{-1}\{\#j : \mathbf{d}^t\hat{E}^*(a) - S^*\sqrt{F}\sqrt{\mathbf{d}^t(\mathbf{D}^{*\mathbf{t}}\mathbf{D}^*)^{-1}\mathbf{d}} \le \mathbf{d}^t\hat{E}(a) \le$$

$$\mathbf{d}^t\hat{E}^*(a) + S^*\sqrt{F}\sqrt{\mathbf{d}^t(\mathbf{D}^{*\mathbf{t}}\mathbf{D}^*)^{-1}\mathbf{d}}\;\forall\mathbf{d}\} =$$

$$(n_b)^{-1}\{\#j : (\hat{E}^*(a) - \hat{E}(a))^t(\mathbf{D}^{*\mathbf{t}}\mathbf{D}^*)(\hat{E}^*(a) - \hat{E}(a)) \le F(S^*)^2\}$$

Note that the latter expression is what is used in the computation. This is the standard Scheffé method, as described in [20].

8. For a 100p% confidence region compute $(\sqrt{F})_p = \min\{\sqrt{F} : \hat{G}_B(\sqrt{F}) \ge p\}$ and use that value in the equation in step 4. In our case, we are interested in obtaining a 95% confidence region, so $\sqrt{F}$ is chosen so that for 95% of the bootstrap samples

$$(\hat{E}^*(\underline{a}) - \hat{E}(\underline{a}))^t(\underline{\mathbf{D}}^{*t}\underline{\mathbf{D}}^*)(\hat{E}^*(\underline{a}) - \hat{E}(\underline{a})) \le F(S^*)^2.$$

In this model, the bit rate is treated as a nonrandom predictor that we control, and the judges and images are "random effects" because our 3 judges and 30 images have been sampled from arbitrarily large numbers of possibilities.

Figure 1 displays all data for lung sensitivity and lung PVP (calculated relative to the consensus gold standard) for all 24 images, judges, and compressed levels for which there was a consensus gold standard. There are 360 x's: 360 = 3 judges × 24 images × 5 compressed levels seen for each image. Figure 2 is the corresponding figure for the mediastinum relative to the personal gold standard. The o's mark the average of the x's for each bit rate. The values of the sensitivity and PVP are simple fractions such as 1/2 and 2/3 because there are at most a few abnormalities in each image. The curves are least squares quadratic spline fits to the data with a single knot at 1.5 bpp, together with the two-sided 95% confidence regions. Since the sensitivity and PVP cannot exceed 1, the upper confidence curve was thresholded at 1. The residual root mean square (RMS) is the square root of the residual mean square from an analysis

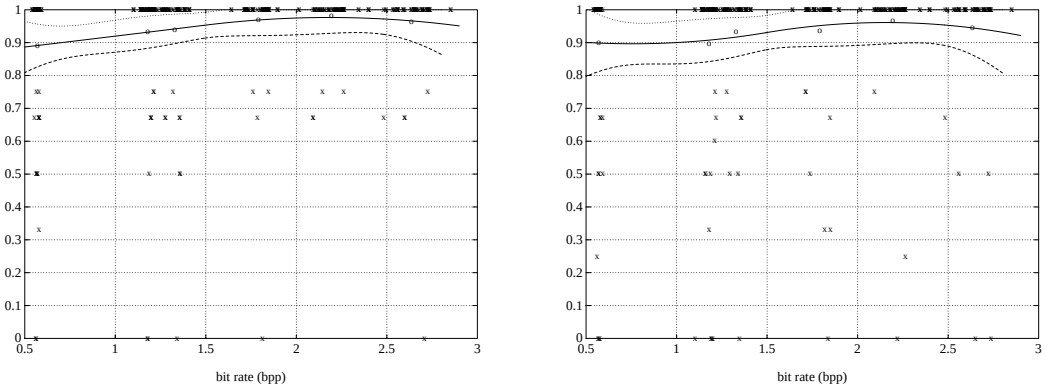

Figure 1: Relative to the consensus gold standard: (a) Lung Sensitivity (RMS=.177), (b) Lung PVP (RMS=.215)

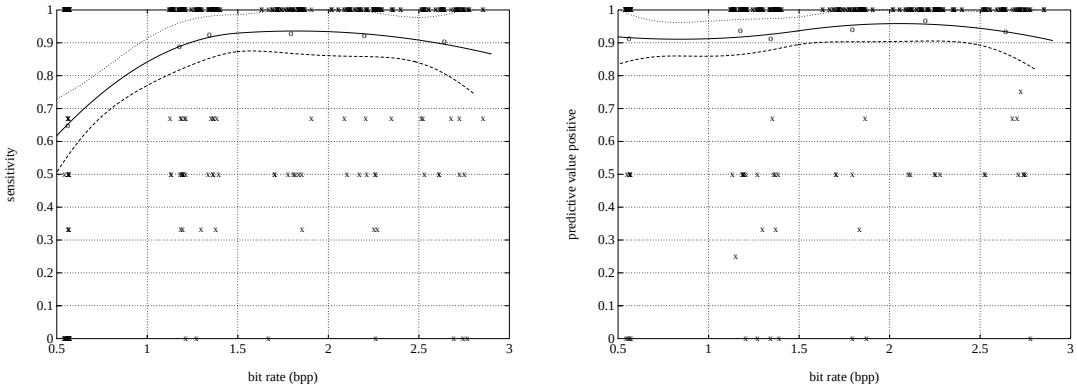

Figure 2: Relative to the personal gold standard: (a) Mediastinum Sensitivity (RMS=.243), (b) Mediastinum PVP (RMS=.245)

Type	Judge	Number of abnormalities								
		0	1	2	3	4	5	6	7	8
Lung	1	3	11	7	6	2		1		
Lung	2	4	9	10	4	2		1		
Lung	3	3	8	8	5	2	2	1		1
Lung	All	4	9	4	5	2				
Mdst	1	3	14	7	6					
Mdst	2	2	22	2	4					
Mdst	3	3	22	4	1					
Mdst	All	3	17	2	2					

Table 1: Number of test images which contain the listed number of abnormalities (Mdst = mediastinum)

of variance of the spline fits. Sensitivity for the lung seems to be nearly as good at low rates of compression as at high rates, but sensitivity for the mediastinum drops off at the lower bit rates, driven primarily by the results for one judge. PVP for the lung is roughly constant across the bit rates, and the same for the mediastinum.

Table 1 shows the numbers of original test images (out of 30 total) 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 (out of 24 total) which contain the listed number of abnormalities according to the consensus gold standard. We examine this table to determine whether or not it is valid to pool the sensitivity and PVP results across judges. Simple chi-square tests for homogeneity show that for both the lung and the mediastinum judges do 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, then the chi-square statistic is 3.16 (on 8 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 chi-square value (on 6 degrees of freedom) is 8.83. However, Table 1 does not fully indicate the variability among the judges. For example, the table shows that each judge found 6 lung nodules in an original test image only once. However, it was not the same test image for all three for which this occurred.

Behrens-Fisher-Welch t-statistic The comparison of sensitivity and PVP at different bit rates was carried out using a permutation distribution of a two-sample t-test that is sometimes called the Behrens-Fisher-Welch test [3, 18]. The statistic takes account of the fact that the within group variances are different. In the standard paired t-test where we have n pairs of observations, let μ_D denote the true, and unknown, average difference between the members of a pair. If we denote the sample mean difference between the members of the pairs by $\bar{D}$, and the estimate of standard deviation of these differences by $s_{\bar{D}}$, then the quantity

$$t = \frac{\bar{D} - \mu_D}{s_{\bar{D}}}$$

follows (under certain normality assumptions) Student's t distribution with $(n-1)$ degrees of freedom, and this may be used to test the null hypothesis that $\mu_D = 0$, that is, that there is no difference between the members of a pair [27]. Now, with our sensitivity and PVP data, there is no single estimate $s_{\bar{D}}$ of the standard deviation that can be made. For an image I_1 which has only one abnormality according to the consensus gold standard, the judges can have sensitivity equal to either 0 or 1, but for an image I_2 with three abnormalities the sensitivity can equal 0, 0.33, 0.67, or 1. So, in comparing bit rates b_1 and b_2, when we form a pair out of image I_1 seen at bit rates b_1 and b_2, and we form another pair out of image I_2 seen at bit rates b_1 and b_2, we see that the variance associated with some pairs is larger than that associated with other pairs. The Behrens-Fisher-Welch test takes account of this inequality of variances. The test is exact and does not rely on Gaussian assumptions that would be patently false for this data set. The use of this statistic is illustrated by the following example. Suppose Judge 1 has judged N lung images at both levels A and B. These images can be divided into 5 groups, according to whether the consensus gold standard for the image contained 0, 1, 2, 3, or 4 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 level A and at level B. Let $\bar{\Delta}_i$ be the average difference:

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

We define

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

and then the Behrens-Fisher-Welch t statistic is given by

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

In the consensus gold standard, there were never more than 4 abnormalities found. So the Δ_{ij} are fractions with denominators not more than 4, and are utterly nonGaussian. (For the personal gold standard, the denominator could be as large as 8.) Therefore, computations of attained significance (p-values) are based on the restricted permutation distribution of t_{BFW}. For each of the N images, we can permute the results from the two levels [A $\rightarrow$ B & B $\rightarrow$ A] or not. There are 2^N points possible in the full permutation distribution, and we calculate t_{BFW} for each one. The motivation for the permutation distribution is that if there were no difference between the bit rates, then in computing the differences Δ_{ij}, it should not matter whether we compute Level A $-$ Level B or vice versa, and we would not expect the "real" t_{BFW} to be an extreme value among the 2^N values. If k is the number of permuted t_{BFW} values that exceed the "real" one, then $(k+1)/2^N$ is the attained one-sided significance level for the test of the null hypothesis that the lower bit rate performs at least as well as the higher one. As discussed later, the one-sided test of significance is chosen to be conservative and to argue most strongly against compression.

When the judges were evaluated separately, level A (the lowest bit rate) was found to be significantly different at the 5% level against most of the other levels for two of the judges, for both lung and mediastinum sensitivity. No differences were found among levels B through G. There were no significant differences found between any pair of levels for PVP. When judges were pooled, more significant differences were found. Level A was generally inferior to the other levels for both lung and mediastinal sensitivity. Also levels B and C differed from level G for lung sensitivity (p = 0.016 for both) and levels B and C differed from level G for mediastinal sensitivity (p = 0.008 and 0.016, respectively). For PVP, no differences were found against level A with the exception of A vs. E and F for the lungs (p = 0.039 and 0.012, respectively), but B was somewhat different from C for the lungs (p = 0.031), and C was different from E, F, and G for the mediastinum (p = 0.016, 0.048, and 0.027, respectively).

Using the consensus gold standard, the results indicate that level A (0.56 bpp) is unacceptable for diagnostic use. Since the blocking and prediction artifacts became quite noticeable at level A, the judges tended not to attempt to mark any abnormality unless they were quite sure it was there. This explains the initially surprising result that level A did well for PVP, but very poorly for sensitivity. Since no differences were found among levels D (1.8 bpp), E (2.2 bpp), F (2.64 bpp), and G (original images at 12 bpp), despite the biases against compression contained in our analysis methods, these 3 compressed levels are clearly acceptable for diagnostic use in our applications. The decision concerning levels B (1.18 bpp) and C (1.34 bpp) is less clear, and would require further tests involving a larger number of detection tasks, more judges, or use of a different gold standard that in principle could remove at least one of the biases against compression that are present in this study.

Since the personal gold standard has the advantage of using all the images in the study, and the consensus gold standard has the advantage of having little bias between original and compressed images, we can capitalize on both sets of advantages with a 2-step comparison. Sensitivity and PVP values relative to the consensus gold standard show there to be no significant differences between the slightly compressed images (levels D, E, and F) and the originals. This is true for both disease categories, for judges evaluated separately and pooled, and using both the Behrens-Fisher-Welch test to examine the sensitivity and PVP separately and using the McNemar test (discussed in the next chapter) to examine them in combination. With this assurance, the personal gold standard can then be used to look for differences between the more compressed levels (A, B, C) and the less compressed ones (D, E, F). The most compressed level A (0.56 bpp, 21:1 compression ratio) is unacceptable as observations made on these images were significantly different from those on less compressed images for two judges. Level B (1.18 bpp) is also unacceptable, although barely so, because the only significant difference was between the sensitivities at levels B and F for a single disease category and a single judge. No differences were found between level C and the less compressed levels, nor were there any significant differences between levels D, E, and F.

In summary, using the consensus gold standard alone, the results indicate that levels D, E, and F are clearly acceptable for diagnostic use, level A is clearly unacceptable, and levels B and C are marginally unacceptable. Using the

personal and consensus gold standard data jointly, the results indicate that levels C, D, E, and F are clearly acceptable for diagnostic use, level A is clearly unacceptable, and level B is marginally unacceptable.

We would like to conclude that there are some compression schemes whose implementation would not degrade clinical practice. To make this point, we must either use tests that are unbiased, or, acting as our own devil's advocates, use tests that are biased against compression. This criterion is met by the fact that the statistical approach described here contains 4 identifiable biases, none of which favors compression. The biases are as follows.

1. As discussed in the previous chapter, the gold standard confers an advantage upon the original images relative to the compressed levels. This bias is mild in the case of the consensus gold standard, but severe in the case of the personal gold standard.

2. There is a bias introduced by multiple comparisons [20]. Since (for each gold standard) we perform comparisons for all possible pairs out of the 7 levels, for both sensitivity and PVP, for both lung and mediastinal images, and both for 3 judges separately and for judges pooled, we are reporting on $21 \times 2 \times 2 \times 4 = 336$ tests for each gold standard. One would expect that, even if there were no effect of compression upon diagnosis, 5% of these comparisons would show significant differences at the 5% significance level.

3. A third element which argues against compression is the use of a 1-sided test instead of a 2-sided test. In most contexts, for example when a new and old treatment are being compared and subjects on the new treatment do better than those on the old, we do a 2-sided test of significance. Such 2-sided tests implicitly account for both possibilities: that new interventions may make for better or worse outcomes than standard ones. For us, a 2-sided test would implicitly recognize the possibility that compression improves, not degrades, clinical practice. In fact, we believe this can happen, but to incorporate such beliefs in our formulation of a test would make us less our own devil's advocates than would our use of a 1-sided test. Our task is to find when compression might be used with clinical impunity, not when it might enhance images.

4. The fourth bias stems from the fact that the summands in the numerator of t_{BFW} may well be positively correlated (in the statistical sense), though we have no way to estimate this positive dependence from our data. If we did, the denominator of t_{BFW} would typically be smaller, and such incorporation would make finding "significant" differences between compression levels more difficult.

For all of these reasons, we believe that the stated conclusions are conservative.

3 MR study: Example of measurement accuracy

Previous studies of the effects of lossy compression on diagnostic accuracy have focused on the detection of structures [5, 7, 8, 15, 19, 26]. However, measurement tasks also play a crucial role in diagnostic radiology. Measurements on structures such as blood vessels, other organs, and tumors take a central role in the identification of abnormalities and in following therapeutic response. Radiologists routinely measure almost everything they detect. For example, while diagnosing a fractured bone, they might measure the displacement between the two pieces of bone, or when reporting the presence of metastatic lesions in the liver, they might measure the size of the largest one. Often such measurements are not of great importance to clinical decision-making; but in some examples, they are of extreme significance. In vascular surgery, for example, precise measurements taken on angiograms of the distance from an area of stenosis to the nearest bifurcation in the vascular structure are needed to reduce surgical exposure. In the evaluation of aneurysms, size is an important prognostic feature in any presurgical assessment. Precise measurements on images are increasingly important in those areas where 3-D stereotactic localization can lead to less invasive surgical biopsy methods. For example, in mammography, fine needle biopsy techniques require careful distance measurements in order to place the needle correctly. For our study of the effects of compression on measurement accuracy, we chose to look at measurements of aortic aneurysms, one of the most common areas where size measurements radically affect clinical decision making.

Abdominal aortic aneurysms are usually evaluated with ultrasound, and thoracic aortic aneurysms are evaluated by CT or MRI. In the latter case, the aortic diameter is usually measured manually with calipers. If the aorta exceeds 4 cm in diameter, an aneurysm is diagnosed. A larger aneurysm carries a greater risk of rupture, with approximately 10% risk of rupture for aneurysms between 5 and 10 cm in diameter, and about 50% for aneurysms greater than 10 cm [17]. Rupture is invariably fatal, and so when the aorta measures more than about 5 or 6 cm in diameter, operative

repair is usually recommended [28, 6]. The clinical decision depends not only on the size of the aneurysm but also on the clinical status of the patient (issues of pain and hemodynamic instability). Dilation less than 5 cm in diameter may be followed conservatively by serial MR imaging studies at 6-month intervals. Observing an increase in the aortic diameter of 0.5 cm over the course of a 6 month interval would be indication for surgical repair.

The study described here had as its goal to quantify the effects of lossy compression on measurement accuracy through experiments that follow closely the clinical tasks of radiologists evaluating aortic aneurysms [25, 24]. We wished to examine whether compression maintains the information required for accurate measurements, or whether it leads to inaccuracies by blurring edges or distorting structures. If compression at a certain bit rate caused a 0.5 cm error in the aortic measurement, that would impact on the clinical decision, and the compression would be unacceptable. Although we focused on the medical problem of thoracic aortic aneurysms as seen on MR scans, the methodology developed in this research is applicable to any medical task requiring the measurement of structures.

The task studied was the measurement of four primary blood vessels in the mediastinum: the ascending aorta, descending aorta, right pulmonary artery (RPA), and superior vena cava (SVC). A set of 9-bit original MR chest images containing aneurysms and normal vessels was compressed to five bit rates between 0.36 and 1.7 bpp. Radiologists measured the four vessels on each image.

In our statistical analyses, we set two gold standards, "personal" and "independent". As discussed in the previous chapter, these represent two methods of establishing the correct size of each blood vessel, that is, the underlying diagnostic "truth" of each image. For each of these gold standards, we quantify the accuracy of the measurements at each compression level by taking the percent measurement error for each image, defined to be the difference between a radiologist's measurement and the gold standard, scaled by the gold standard measurement. This error is examined as a function of bit rate by using the t-test and a nonparametric competitor, the Wilcoxon signed rank test.

3.1 Study Design and Statistical Analysis

To simulate normal clinical practice, test images were selected from 30 sequential thoracic MR examinations of diagnostic quality obtained after February 1, 1991. The patients studied included 16 females and 14 males, with ages ranging from 1 to 93 years and an average age of 48.0 $\pm$ 24.7 years (mean $\pm$ s.d.). Clinical diagnoses included aortic aneurysm (n=11), thoracic tumors (n=11), pre- or post-lung transplant (n=5), constrictive pericarditis (n=1), and subclavian artery rupture (n=1). From each examination, one image which best demonstrated all four major vessels of interest was selected. The training images were selected similarly from different examinations. All analyses were based solely on measurements made on the test images.

The 30 test scans compressed to 5 bit rates plus the originals give rise to a total of 180 images. These images were arranged in a randomized sequence and presented on separate hardcopy films to three radiologists. The viewing protocol consisted of 3 sessions held at least 2 weeks apart. Each session included 10 films viewed in a pre-determined order with six scans on each film. The three radiologists began viewing films at a different starting point in the randomized sequence. To minimize the probability of remembering measurements from past images, a radiologist saw only 2 of the 6 levels of each image in each session, with the second occurrence of each image spaced at least 4 films after the first occurrence of that image.

Following standard clinical methods for detecting aneurysms, the radiologists used calipers and a mm scale available on each image to measure the four blood vessels appearing on each scan. Although the use of digital calipers might have allowed more accurate measurements, this would have violated one of our principal goals, namely to follow as closely as possible actual clinical practice. It is the standard practice of almost all radiologists to measure with manual calipers. This is largely because they lack the equipment, or they would prefer not to take the time to bring up the relevant image on the terminal and then perform the measurements with electronic calipers. We asked radiologists to make all measurements between the outer walls of the vessels along the axis of maximum diameter. It is this maximum diameter measurement that is used to make clinical decisions. Both the measurements and axes were marked on the film with a grease pencil.

The independent gold standard was set by having two radiologists come to an agreement on vessel sizes on the original scans. They first independently measured the vessels on each scan and then remeasured only those vessels on which they initially differed until an exact agreement on the number of millimeters was reached. These two radiologists are different from the three radiologists whose judgments are used to determine diagnostic accuracy. A personal standard was also derived for each of the three judging radiologists by taking their own measurements on the original images.

Once the gold standard measurement for each vessel in each image was assigned, measurement error can be quantified in a variety of ways. If z is the radiologist's measurement and g represents the gold standard measurement, then some potential summary statistics are

$$(z-g), \quad \log(\frac{z}{g}), \quad \frac{z-g}{g}, \quad |\frac{z-g}{g}|.$$

These statistics have invariance properties that bear upon understanding the data. For example, $z-g$ is invariant to the same additive constant (that is, to a change in origin), $\log(z/g)$ is invariant to the same multiplicative constant (that is, to a change in scale), and $(z-g)/g$ is invariant to the same multiplicative constant and to the same sign changes. For simplicity and appropriateness in the statistical tests carried out, the error parameters chosen for this study are percent measurement error (pme)

$$\text{pme} = \frac{z-g}{g} \times 100\%$$

and absolute percent measurement error (apme)

$$\text{apme} = \frac{|z-g|}{g} \times 100\%,$$

both of which scale the error by the gold standard measurement to give a concept of error relative to the size of the vessel being measured.

The differences in error achieved at each bit rate can be quantified as statistically significant by many tests. Each should respect the pairing of the measurements being compared and the multiplicity of comparisons being made. In order to ensure that our conclusions are not governed by the test being used, we chose to use two of the most common, the t and Wilcoxon tests. We also employed statistical techniques that account for this multiplicity of tests. The measurements are considered paired in a comparison of two bit rates since the same vessel in the same image is measured by the same radiologist at both bit rates. For instance, let x_1 be the measurement of a vessel at bit rate 1, x_2 be its measurement at bit rate 2, and g be the vessel's gold standard measurement. Then the pme at bit rates 1 and 2 are

$$\text{pme}_1 = \frac{x_1-g}{g} \times 100\% \quad \text{and} \quad \text{pme}_2 = \frac{x_2-g}{g} \times 100\%,$$

and their difference is

$$\text{pme}_D = \frac{x_1-x_2}{g} \times 100\%.$$

In such a two-level comparison, pme more accurately preserves the difference between two errors than does apme. A vessel that is over-measured by α % (positive) on bit rate 1 and under-measured by α (negative) % on bit rate 2 will have an error distance of 2α % if pme is used but a distance of zero if apme is used. Therefore both the t-test and the Wilcoxon signed rank test are carried out using only pme. Apme is used later to present a more accurate picture of error when we plot an average of apme across the 30 test images versus bit rate.

The t-statistic quantifies the statistical significance of the observed difference between two data sets in which the data can be paired. Unlike the CT study, in which the Behrens-Fisher-Welch t-test was used because of the obviously different variances present for different images, here the ordinary t-test was applicable. The difference in error for two bit rates is calculated for all the vessels measured at both bit rates. If the radiologists made greater errors at bit rate 1 than at bit rate 2, the average difference in error over all the data will be positive. If bit rate 1 is no more or less likely to cause error than bit rate 2, the average difference in error is zero. The t-test assumes that the sample average difference in error between two bit rates varies in a Gaussian manner about the real average difference [27]. If the data are Gaussian, which they clearly cannot exactly be in our application, the paired t-test is an exact test. Quantile-Quantile plots of pme differences for comparing levels vary from linear to S-shaped; in general, the Q-Q plots indicate a moderate fit to the Gaussian model. The size of our data set (4 vessels $\times$ 30 images $\times$ 6 levels $\times$ 3 judges = 2160 data points) makes a formal test for normality nearly irrelevant. The large number of data points serves to guarantee failure of even fairly Gaussian data at conventional levels of significance. (That is, the generating distribution is likely not to be exactly Gaussian, and with enough data, even a tiny discrepancy from Gaussian will be apparent.) Even if the

data are non-Gaussian, however, the central limit theorem renders the t-test approximately valid. With the Wilcoxon signed rank test [27] the significance of the difference between the bit rates is obtained by comparing a standardized value of the Wilcoxon statistic against the normal standard deviate at the 95% 2-tail confidence level. The distribution of this standardized Wilcoxon is nearly exactly Gaussian if the null hypothesis is true for samples as small as 20.

Results using the independent gold standard: Plots of trends in measurement error as a function of bit rate are presented in Figures 3–6. In all cases, the general trend of the data is indicated by fitting the data points with a quadratic spline having one knot at 1.0 bpp. Figure 3 gives average pme against the mean bit rate for all radiologists pooled (i.e., the data for all radiologists, images, levels, and structures, with each radiologist's measurements compared to the independent gold standard) and for each of the three radiologists separately. In Figure 4, the pme versus actual achieved bit rate is plotted for all data points. The relatively flat curve begins to increase slightly at the lowest bit rates, levels 1 and 2 (0.36, 0.55 bpp). It is apparent from an initial observation of these plots that except for measurement at the lowest bit rates, accuracy does not vary greatly with lossy compression. Possibly significant increases in error appear only at the lowest bit rates, whereas at the remaining bit rates measurement accuracy is similar to that obtained with the originals. The average performance on images compressed to level 5 (1.7 bpp) is actually better than performance on originals.

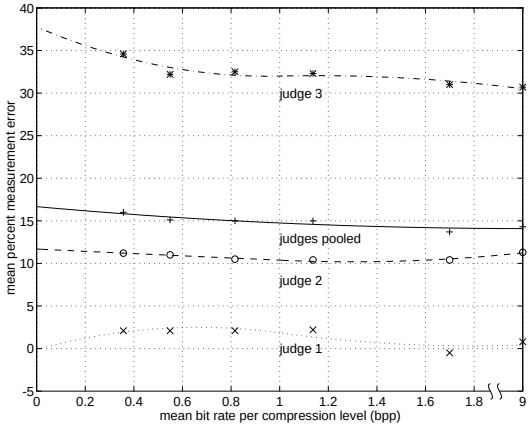

Figure 3: Mean pme vs. mean bit rate using the independent gold standard. The dotted, dashed, and dash-dot curves are quadratic splines fit to the data points for Judges 1, 2, and 3, respectively. The solid curve is a quadratic spline fit to the data points for all judges pooled. The splines have a single knot at 1.0 bpp.

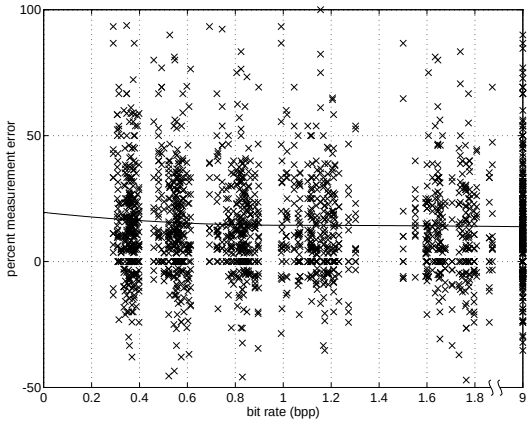

Figure 4: Percent measurement error vs. actual bit rate using the independent gold standard. The x's indicate data points for all images, pooled across judges and compression levels. The solid curve is a quadratic spline fit to the data with a single knot at 1.0 bpp.

While the trends in pme vs. bit rate are useful, over-measurement (positive error) can cancel under-measurement

(negative error) when these errors are being averaged or fitted with a spline. For this reason, we turn to apme which measures the error made by a radiologist regardless of whether it originated from over-measurement or under-measurement. Figure 5 plots average apme versus average bit rate for each radiologist and for all radiologists pooled. Figure 6 shows actual apme versus actual bit rate achieved. These plots show trends similar to those observed before. The original level contains more or less the same apme as compression levels 3, 4, and 5 (0.82, 1.14, 1.7 bpp). Levels 1 and 2 (0.36, 0.55 bpp) show slightly higher error. These plots provide only approximate visual trends in data.

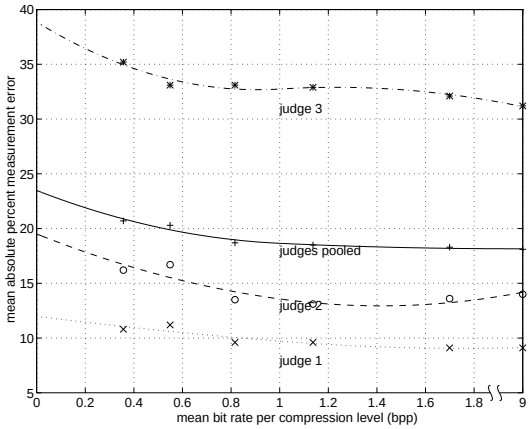

Figure 5: Mean apme vs. mean bit rate using the independent gold standard. The dotted, dashed, and dash-dot curves are quadratic splines fit to the data points for Judges 1, 2, and 3, respectively. The solid curve is a quadratic spline fit to the data points for all judges pooled.

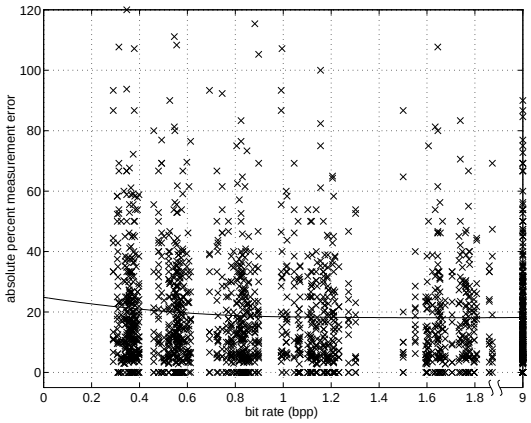

Figure 6: Apme vs. actual bit rate using the independent gold standard. The x's indicate data points for all images, pooled across judges and compresssion levels. The solid curve is a quadratic spline fit to the data.

The t-test was used to test the null-hypothesis that the "true" pme between two bit rates is zero. The standardized average difference is compared with the "null" value of zero by comparing with standard normal tables. None of the compressed images down to the lowest bit rate of 0.36 bpp was found to have a significantly higher pme when compared to the error made on the originals. Among the compressed levels however, level 1 (0.36 bpp) was found to be significantly different from level 5 (1.7 bpp). As was mentioned, the performance on level 5 was better than that on all levels, including the uncompressed level.

When using the Wilcoxon signed rank test to compare compressed images against the originals, only level 1 (0.36 bpp) differed significantly in the distribution of pme. Within the levels representing the compressed images, levels 1, 3, and 4 (0.36, 0.82, 1.14 bpp) had significantly different pme than those at level 5 (1.7 bpp). Since measurement accuracy is determined from the differences with respect to the originals only, a conservative view of the results of the analyses using the independent gold standard is that accuracy is retained down to 0.55 bpp (level 2).

Results using the personal gold standard:

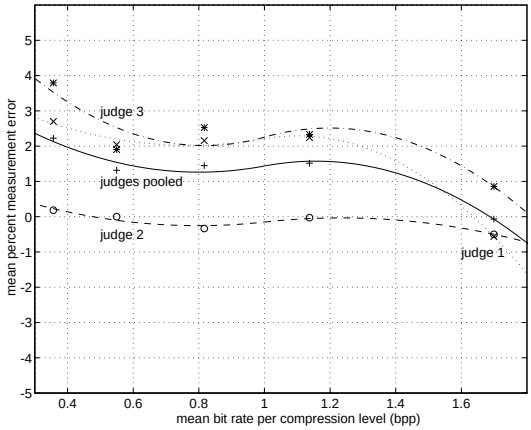

Figure 7: Mean pme vs. mean bit rate using the personal gold standard. The dotted, dashed, and dash-dot curves are quadratic splines fit to the data points for Judges 1, 2, and 3, respectively. The solid curve is a quadratic spline fit to the data points for all judges pooled.

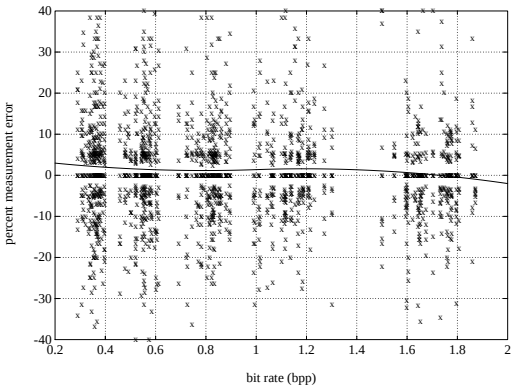

Figure 8: Pme vs. actual bit rate using the personal gold standard. The x's indicate data points for all images, pooled across judges and compression levels. The solid curve is a quadratic spline fit to the data.

As was discussed previously, the personal gold standard was set by taking a radiologist's recorded vessel size on the uncompressed image to be the correct measurement for judging her or his performance on the compressed images. Using a personal gold standard in general accounts for a measurement bias attributed to an individual radiologist, thereby providing a more consistent result among the measurements of each judge at the different compression levels. The personal gold standard thus eliminates the interobserver variablity present with the independent gold standard. However, it does not allow us to compare performance at compressed bit rates to performance at the original bit rates since the standard is determined from the original bit rates, thereby giving the original images zero error. As before, we first consider visual trends and then quantify differences between levels by statistical tests.

Figure 7 shows average pme vs. mean bit rate for the 5 compressed levels for each judge separately and for the judges pooled, whereas Figure 8 is a display of the actual pme vs. actual achieved bit rate for all the data points. The data for the judges pooled are the measurements from all judges, images, levels, and vessels, with each judge's measurements compared to her or his personal gold standard. In each case, quadratic splines with a single knot at 1.0 bpp were fit to the data. Figures 9 and 10 are the corresponding figures for the apme. As expected, with the personal gold standard the pme and the apme are less than those obtained with the independent gold standard. The graphs indicate that whereas both Judges 2 and 3 overmeasured at all bit rates with respect to the independent gold standard, only Judge 3 overmeasured at the compressed bit rates with respect to the personal gold standard.

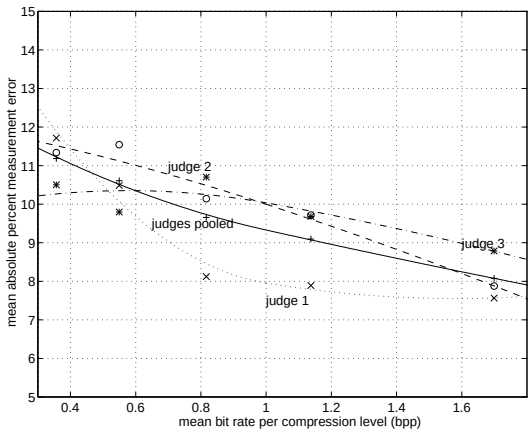

Figure 9: Mean apme vs. mean bit rate using the personal gold standard. The dotted, dashed, and dash-dot curves are quadratic splines fit to the data points for Judges 1, 2, and 3, respectively. The solid curve is a quadratic spline fit to the data points for all judges pooled.

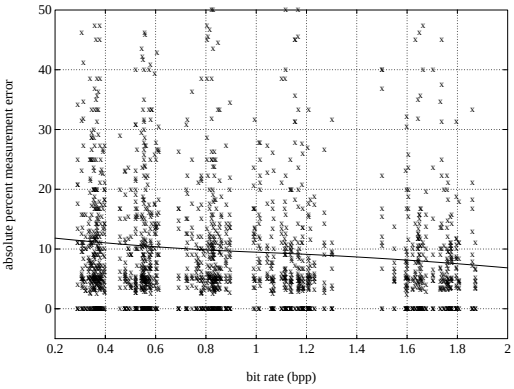

Figure 10: Apme vs. actual bit rate using the personal gold standard. The x's indicate data points for all images, pooled across judges and compresssion levels. The solid curve is a quadratic spline fit to the data.

The t-test results indicate that levels 1 (0.36 bpp) and 4 (1.14 bpp) have significantly different pme associated with them than does the personal gold standard. The results of the Wilcoxon signed rank test on percent measurement error using the personal gold standard are similar to those obtained with the independent gold standard. In particular, only level 1 at 0.36 bpp differed significantly from the originals. Furthermore, levels 1, 3, and 4 were significantly different from level 5.

Since the t-test indicates that some results are marginally significant when the Wilcoxon signed rank test indicates the results are not significant, a Bonferroni simultaneous test (union bound) was constructed. This technique uses the significance level of two different tests to obtain a significance level that is simultaneously applicable for both. For example, in order to obtain a simultaneous significance level of $\alpha\%$ with two tests, we could have the significance of each test be at $\alpha/2\%$. With the simultaneous test, the pme at level 4 (1.14 bpp) is not significantly different from the uncompressed level. As such, the simultaneous test indicates that only level 1 (0.36 bpp) has significantly different pme from the uncompressed level. This agrees with the corresponding result using the independent gold standard. Thus, pme at compression levels down to 0.55 bpp does not seem to differ significantly from the pme at the 9.0 bpp original.

In summary, with both the independent and personal gold standards, the t-test and the Wilcoxon signed rank test indicate that pme at compression levels down to 0.55 bpp did not differ significantly from the pme at the 9.0 bpp original. This was shown to be true for the independent gold standard by a direct application of the tests. For the personal gold standard, this was resolved by using the Bonferroni test for simultaneous validity of multiple analyses. The status of measurement accuracy at 0.36 bpp remains unclear, with the t-test concluding no difference and the Wilcoxon indicating significant difference in pme from the original with the independent gold standard, and both tests indicating significant difference in pme from the original with the personal gold standard. Since the model for the t-test is fitted only fairly to moderately well by the data, we lean towards the more conservative conclusion that lossy compression by our vector quantization compression method is not a cause of significant measurement error at bit rates ranging from 9.0 bpp down to 0.55 bpp, but it does introduce error at 0.36 bpp.

A radiologist's subjective perception of quality changes more rapidly and drastically with decreasing bit rate than does the actual measurement error. Radiologists evidently believe that the usefulness of images for measurement tasks degrades rapidly with decreasing bit rate. However, their actual measurement performance on the images was shown by both the t-test and Wilcoxon signed rank test (or the Bonferroni simultaneous test to resolve differences between the two) to remain consistently high down to 0.55 bpp. Thus, the radiologist's opinion of an image's diagnostic utility seems not to coincide with its utility for the clinical purpose for which the image was taken. The radiologist's subjective opinion of an image's usefulness for diagnosis should not be used as the sole predictor of actual usefulness.

3.2 Discussion

There are issues of bias and variability to consider in comparing and contrasting gold standards. One disadvantage of an independent gold standard is that since it is determined by the measurements of radiologists who do not judge the compressed images, significant differences between a compressed level and the originals may be due to differences between judges. For example, a biased judge who tends to overmeasure at all bit rates may have high pme which will not be entirely reflective of the effects of compression. In our study, we determined that two judges consistently overmeasured relative to the independent gold standard. The personal gold standard, however, overcomes this difficulty. A personal gold standard also has the advantage of reducing pme and apme at the compressed levels. This will result in a clarification of trends in a judge's performance across different compression levels. Differences will be based solely on compression level and not on differences between judges. Another argument in favor of a personal gold standard is that in some clinical settings a fundamental question is how the reports of a radiologist whose information is gathered from compressed images compare to what they would have been on the originals. Indeed, systematic biases of a radiologist are sometimes well recognized and corrected for by the referring physicians.

One disadvantage with the personal gold standard, however, is that by defining the measurements on the original images to be "correct," we are not accounting for the inherent variability of a judge's measurement on an uncompressed image. For example, if a judge makes an inaccurate measurement on the original and accurate measurements on the compressed images, these correct measurements will be interpreted as incorrect. Thus the method is biased against compression. An independent gold standard reduces the possibility of this situation occurring since we need an agreement by two independent radiologists on the "correct" measurement.

The analysis previously presented was based on judges, vessels, and images pooled. Other analyses in which the performances of judges on particular vessels and images are separated demonstrate additional variability. Judges seem

to have performed significantly differently from each other. Judges 2 and 3 consistently overmeasured. As a result, the Wilcoxon signed rank test using the independent gold standard indicates significant differences between the gold standard and the measurements of Judges 2 and 3 at all compression levels, including the original. Judge 1, however, does not have any significant performance differences between the gold standard and any compression levels. In addition, certain vessels and images had greater variability in pme than others. To examine the validity of pooling the results of all judges, vessels, and images, an Analysis of Variance (ANOVA) [21] was used to assess whether this variability is significant. The ANOVA took the judges, vessels, and images to be random effects and the levels to be fixed effects, and separated out the variance due to each effect. For technical reasons it is not feasible here to use direct F-tests on each of the variances estimated. Thus, we obtained confidence regions for each component of variance using a jackknife technique [21]. In particular, if zero falls within the 95% confidence interval of a certain effect, then the effect is not considered significant at the 5% level. Using the jackknife technique, the ANOVA indicates that the variability in judges, vessels, images, and levels were not significantly different from zero, thereby validating the pooling.

4 Mammography study: example of management accuracy

X-ray-mammography is the most sensitive technique for detecting breast cancer [2], with a reported sensitivity of 85–95% for detecting small lesions. Most non-invasive ductal carcinomas, or DCIS, are characterized by tiny non-palpable calcifications detected at screening mammography [29, 9, 16]. Traditional mammography is essentially analog photography using X-rays in place of light and analog film for display. Mammography machines based on direct digital acquisition exist, and the review is in process for FDA approval for market. The study discussed here, however, employed only digitized analog films. The studies were digitized using a Lumisys Lumiscan 150 at 12 bpp with a spot size of 50 microns. After compression, the images were put on hardcopy film. The films were printed using a Kodak 2180 X-ray film printer, a 79 micron 12 bit greyscale printer which writes with a laser diode of 680 nm bandwidth.

Images were viewed on hardcopy film on an alternator by judges in a manner that simulates ordinary screening and diagnostic practice as closely as possible, although patient histories and other image modalities were not provided. Two views were provided of each breast (CC and MLO), so four views were seen simultaneously for each patient. Each of the judges viewed all the images in an appropriately randomized order over the course of nine sessions. Two sessions were held every other week, with a week off in between. A clear overlay was provided for the judge to mark on the image without leaving a visible trace. For each image, the judge either indicated that the image was normal, or, if something was detected, had an assistant fill out the Observer Form using the American College of Radiology (ACR) Standardized Lexicon by circling the appropriate answers or filling in blanks as directed. The Observer Form is given in Figures 11–13 below. The instructions for assistants and radiologists along with suggestions for prompting and a CGI web data entry form may be found at the project Web site `http://www-isl.stanford.edu/~gray/army.html`. The judges used a grease pencil to circle the detected item. The instructions to the judges specified that ellipses drawn around clusters should include all microcalcifications seen, as if making a recommendation for surgery, and outlines drawn around masses should include the main tumor as if grading for clinical staging, without including the spicules (if any) that extend outward from the mass. This corresponds to what is done in clinical practice except for the requirement that the markings be made on copies. The judges were allowed to use a magnifying glass to examine the films.

Although the judging form is not standard, the ACR Lexicon is used to report findings, and hence the judging requires no special training. The reported findings permit subsequent analysis of the quality of an image in the context of its true use, finding and describing anomalies and using them to assess and manage patients.

To confirm that each radiologist identifies and judges a specific finding, the location of each lesion is confirmed both on the clear overlay and the judging form. Many of these lesions were judged as 'A' (assessment incomplete), since it is often the practice of radiologists to obtain additional views in two distinct scenarios: (1) to confirm or exclude the presence of a finding, that is, a finding that may or may not represent a true lesion, or (2) to further characterize a true lesion, that is, to say a lesion clearly exists but is incompletely evaluated.

The judging form allows for two meanings of the 'A' code. If the judge believes that the finding is a possible lesion, this is indicated by answering "yes" to the question "are you uncertain if the finding exists?" Otherwise, if the lesion is definite, the judges should give their best management decision based on the standard two-view mammogram.

The initial question requesting a subjective rating of diagnostic utility on a scale of 1-5 is intended for a separate

ID number _____ Session number _____ Case number _____
Reader initials _____
Mammograms were of (**Left** **Right** **Both**) breast(s).

Subjective rating for diagnostic quality:

(bad) 1 – 5 (good):

Left CC	Left MLO	Right CC	Right MLO

If any rating is < 4 the problem is:
1) sharpness 2) contrast 3) position 4) breast compression
5) noise 6) artifact 7) penetration

Recommend repeat? **Yes** **No**

Breast Density: Left **1** **2** **3** **4** Right **1** **2** **3** **4**

1) almost entirely fat 2) scattered fi broglandular densities
3) heterogeneously dense 4) extremely dense

Findings: **Yes** **No**

Note: If there are NO fi ndings, the assessment is: **(1) (N) negative - return to screening**

Figure 11: Observer Form for mammograms: this part is completed for each case

evaluation of the general subjective opinion of the radiologists of the images. The degree of suspicion registered in the Management portion also provides a subjective rating, but this one is geared towards the strength of the opinion of the reader regarding the cause of the management decision. It is desirable that obviously malignant lesions in a gold standard should also be obviously malignant in the alternative method.

4.1 Statistical Analysis

We focus here on patient management, the decisions that are made based on the radiologists' readings of the image [23, 1, 14]. Management is a key issue in digital mammography. There is concern that artifacts could be introduced, leading to an increase in false positives and hence in unnecessary biopsies. The management categories we emphasize are the following four, given in order of increasing seriousness:

RTS incidental, negative, or benign with return to screening

F/U probably benign but requiring six month follow-up

C/B call back for more information, additional assessment needed

BX Immediate biopsy.

These categories are formed by combining categories from the basic form of the Appendix: RTS is any study that had assessment = 1 or 2, F/U is assessment = 3, C/B is assessment = indeterminate/incomplete with best guess either unsure it exists, 2 or 3, and BX is assessment = indeterminate/incomplete with best guess either 4L, 4M, 4H or 5, or assessment = 4L, 4M, 4H or 5.

We also consider the binarization of these four categories into two groups: Normal and Not Normal. But there is controversy as to where the F/U category belongs, so we make its placement optional with either group. The point is to see if lossy compression makes any difference to the fundamental decision made in screening: does the patient return to ordinary screening as normal, or is there suspicion of a problem and hence the demand for further work?

Truth is determined by agreement with a gold standard. The raw results are plotted as a collection of 2×2 tables, one for each category or group of categories of interest and for each radiologist. As will be discussed, the differences

Findings (detection): **Dominant** **Incidental, focal** **Incidental, diffuse**

Individual finding side: **Left** **Right** **Both/Bilateral** Finding #_____ of_____

Finding type: **(possible, definite)**

1) mass	**2) clustered calcifications**
3) mass containing calcifications	**4) mass with surrounding calcs**
5) spiculated mass	**6) ill defined mass**
7) architectural distortion	**8) solitary dilated duct**
9) asymmetric breast tissue	**10) focal asymmetric density**
11) breast edema	**12) multiple scattered and occasionally**
13) occasional scattered benign appearing calcs	**14) multiple benign appearing masses**
15) skin lesion	**16) milk of calcium**
17) plasma cell mastitis/secretory calcs	**18) oil cysts**
19) lymph node	**20) fibroadenoma**
21) calcified fibroadenoma	**22) vascular calcs**
23) dermal/skin calcs	**24) post biopsy scar**
25) reduction mammoplasty	**26) implants**
27) benign mass	**28) other**

Location:

1) UOQ	**5) 12:00**	**9) outer/lateral**	**13) whole breast**	**17) both breasts/**
2) UIQ	**6) 3:00**	**10) inner/medial**	**14) central**	**bilateral**
3) LOQ	**7) 6:00**	**11) upper/cranial**	**15) axillary tail**	
4) LIQ	**8) 9:00**	**12) lower/inferior**	**16) retroareolar**	

View(s) in which finding is seen: **CC** **MLO** **CC and MLO**

Associated findings include: (p= possible, d= definite)

1) breast edema	**(p , d)**	**8) architectural distortion**	**(p , d)**
2) skin retraction	**(p , d)**	**9) calcs associated**	**(p , d)**
3) nipple retraction	**(p , d)**	**with mass**	
4) skin thickening	**(p , d)**	**10) multiple similar masses**	**(p , d)**
5) lymphadenopathy	**(p , d)**	**11) dilated veins**	**(p , d)**
6) trabecular thickening	**(p , d)**	**12) asymmetric density**	**(p , d)**
7) scar	**(p , d)**	**13) none**	**(p , d)**

Figure 12: Observer Form for mammograms: this part is completed for each *finding* in a case

among radiologists prove to be so large an effect that extreme care must be taken when doing any pooling or averaging of results across radiologists. A typical table is shown in Table 2.

The columns correspond to image modality or method I and the rows to II; I could be original analog and II original digitized, or I could be original digitized and II compressed digitized. "R" and "W" correspond to "right" (agreement with gold standard) and "wrong" (disagreement with gold standard). The particular statistics could be, for example, the decision of "normal," i.e., return to ordinary screening. Regardless of statistic, the goal is to quantify the degree, if any, to which differences exist.

One way to quantify the existence of statistically significant differences is by an exact McNemar test, which is based on the following argument. If there are $N(1,2)$ entries in the (1,2) place and $N(2,1)$ in the (2,1) place, and the technologies are equal, then the conditional distribution of $N(1,2)$ given $N(1,2)+N(2,1)$ is binomial with parameters $N(1,2)+N(2,1)$ and 0.5; that is,

$$P(N(1,2)=k|N(1,2)+N(2,1)=n)=\left(\begin{array}{c} n \\ k \end{array}\right)2^{-n};\ k=0,1,\ldots,n.$$

This is the conditional distribution under the null hypothesis that the two modalities are equivalent. The extent to which $N(1,2)$ differs from $(N(1,2)+N(2,1))/2$ is the extent to which the technologies were found to be different

Assessment: **The finding is**

(A) indeterminate/incomplete, additional assessment needed
What? 1) spot mag 2) extra views 3) U/S 4) old films 5) mag
What is your *best guess* **as to the finding's 1–5 assessment? _____ or are you uncertain if the finding exists?** *Y*

(1) (N) negative – return to screening

(2) (B) benign (also negative but with benign findings) – return to screening

(3) (P) probably benign finding requiring 6-month followup

(4L) (S) suspicion of malignancy (low), biopsy

(4M) (S) suspicion of malignancy (moderate), biopsy

(4H) (S) suspicion of malignancy (high), biopsy

(5) radiographic malignancy, biopsy

Comments: _____

Measurements:
CC View Size: _____ cm long axis by _____ cm short axis
Distance from center of finding to: nipple _____ cm
left edge _____ cm, top edge _____ cm

MLO View Size: _____ cm long axis by _____ cm short axis
Distance from center of finding to: nipple _____ cm
left edge _____ cm, top edge _____ cm

Figure 13: Observer Form for mammograms: this assessment portion is completed for each *finding* in a case

in the quality of performance with their use. Let $B(n, 1/2)$ denote a binomial random variable with these parameters. Then a statistically significant difference at level .05, say, will be detected if the observed k is so unlikely under the binomial distribution that a hypothesis test with size .05 would reject the null hypothesis if k were viewed. Thus if

$$\Pr(|B(n,1/2) - \frac{n}{2}| \geq |N(1,2) - \frac{n}{2}|) \leq .05,$$

then we declare a statistically significant difference has occurred.

Whether and how to agglomerate the multiple tables is an issue. Generally speaking, we stratify the data so that any test statistics we apply can be assumed to have sampling distributions that we could defend in practice. It is always interesting to simply pool the data within a radiologist across all gold standard values, though it is really an analysis of the off-diagonal entries of such a table that is of primary interest. If we look at such a 4×4 table in advance of deciding upon which entry to focus, then we must contend with problems of multiple testing, which would lower the power of our various tests. Pooling the data within gold standard values but across radiologists is problematical because our radiologists are patently different in their clinical performances. This is consistent with what we found in the CT and MR studies. Thus, even if one does agglomerate, there is the issue of how. Minus twice the sum over tables of the natural logarithms of attained significance levels has, apart from considerations of discreteness of the binomial distribution, a chi-square distribution with degrees of freedom twice the number of summands if the null hypothesis of no difference is true for each table and if the outcomes of the tables are independent. This method was made famous by R.A. Fisher. Then again, $(N(1,2)-N(2,1))^2/(N(1,2)+N(2,1))$ has, under the null hypothesis of no difference,

II\ I	R	W
R	N(1,1)	N(1,2)
W	N(2,1)	N(2,2)

Table 2: Agreement 2×2 Table

approximately at least, a chi-square distribution with one degree of freedom, if the null hypothesis of no difference in technologies is correct for the table. One can sum across tables and compare with chi-square tables where the degrees of freedom are the number of summands, a valid test if tables are independent.

4.2 Results and Discussion

The clinical experiment took place at Stanford University Hospital during spring 1996. The gold standard was established by E. Sickles, M.D., Professor of Radiology, University of California at San Francisco, and Chief of Radiology, Mt. Zion Hospital, and D. Ikeda, Assistant Professor and Chief, Breast Imaging Section, Department of Radiology, Stanford University, an independent panel of expert radiologists, who evaluated the test cases and then collaborated to reach agreement. The majority of the detected items were seen by both radiologists. Any findings seen by only one radiologist were included. The other type of discrepancy resolved was the class of the detected lesions. Since the same abnormality may be classified differently, the two radiologists were asked to agree on a class.

The focus of the statistical analysis is the screening and management of patients and how it is affected by analog vs. digital and lossy compressed digital. In all, there were 57 studies that figure in what we report. According to the gold standard, the respective numbers of studies of each of the four types management types RTS, F/U, C/B, and BX were 13, 1, 18, and 25, respectively.

For each of the four possible outcomes, the analog original is compared to each of four technologies: digitized from analog original, and wavelet compressed to three different levels of compression (1.75 bpp, 0.4 bpp, and 0.15 bpp). So the McNemar 2×2 statistics based on the generic table of Table 2 for assessing differences between technologies were computed 48 times, 16 per radiologist, for each competing image modality (original digital and the three lossy compressed bit rates). For example, the 2×2 tables for a single radiologist (A) comparing analog to each of the other four modalities is shown in Table 3

For none of these tables for any radiologist was the exact binomial attained significance level (p-value) .05 or less. For our study and for this analysis, there is nothing to choose in terms of being "better" among the analog original, its digitized version, and three levels of compression, one rather extreme. We admit freely that this limited study had insufficient power to permit us to detect small differences in management. The larger the putative difference, the better our power to have detected it. Table 4 summarizes the performance of each radiologist on the analog vs. uncompressed digital and lossy compressed digital using the independent gold standard. In all cases, columns are "digital" and rows "analog". Table 4(A) treats analog vs. original digital and Tables 4(B)–(D) treat analog vs. lossy compressed digital at bit rates of 1.75 bpp, 0.4 bpp, and 0.155 bpp, respectively.

Consider as an example the analog vs. digital comparison of (A). Radiologist A made 23 "mistakes" of 57 studies from analog, and 20 from digital studies. The most frequent mistake, seven for both technologies, was classifying a gold standard "biopsy" as "additional assessment". Radiologist B made 26 "mistakes" from analog studies, and 28 from digital. In both cases, the most frequent mistake was to "biopsy" what should, by the gold standard, have been "additional assessment". There were 15 such mistakes with analog and 14 with digital. Radiologist C made 19 "mistakes" from analog studies and 19 from digital. With the former, the most frequent mistake occurred eight times when "biopsy" was judged when "additional assessment" was correct. With digital, the most frequent mistakes were for what was judged "additional assessment", but that should have been "biopsy" for five and "return to screening" for five. On this basis, we cannot say that analog and digital are different beyond chance. However, we note here, as elsewhere, that radiological practice varies considerably by radiologist.

The primary conclusion from these data and analyses is that variabilities among judges exceed by a considerable amount, in their main effects and interactions, the variability in performance that owes to imaging modality or compression within very broad limits. In other words, the differences among analog, digital, and lossy compressed images are in the noise of the differences among radiologists, and are therefore more difficult to evaluate. This suggests variations in statistical analysis that will be explored in other experiments.

II\ I	R	W
R	8	1
W	2	2

RTS

II\ I	R	W
R	0	0
W	0	1

F/U

II\ I	R	W
R	7	3
W	3	5

C/B

II\ I	R	W
R	15	1
W	2	7

BX

(A) Analog vs. Digital Original

II\ I	R	W
R	4	4
W	0	4

RTS

II\ I	R	W
R	0	0
W	0	1

F/U

II\ I	R	W
R	3	7
W	4	4

C/B

II\ I	R	W
R	11	4
W	4	5

BX

(B) Analog vs. Digital Lossy Compressed: 1.75 bpp

II\ I	R	W
R	7	2
W	1	3

RTS

II\ I	R	W
R	0	0
W	0	1

F/U

II\ I	R	W
R	6	4
W	2	6

C/B

II\ I	R	W
R	13	2
W	4	5

BX

(C) Analog vs. Digital Lossy Compressed: 0.4 bpp

II\ I	R	W
R	6	3
W	1	3

RTS

II\ I	R	W
R	0	0
W	0	1

F/U

II\ I	R	W
R	8	2
W	2	6

C/B

II\ I	R	W
R	14	2
W	1	8

BX

(D) Analog vs. Digital Lossy Compressed: 0.15 bpp

Table 3: Agreement 2×2 Tables For Radiologist A

	RTS	F/U	C/B	BX
RTS	12	0	5	0
F/U	0	0	0	0
C/B	3	0	12	6
BX	0	0	2	17

	RTS	F/U	C/B	BX
RTS	3	0	1	0
F/U	0	1	0	0
C/B	3	0	3	3
BX	1	0	5	37

	RTS	F/U	C/B	BX
RTS	9	0	5	1
F/U	0	0	0	0
C/B	0	0	11	1
BX	0	0	7	23

A: Analog versus Digital

	RTS	F/U	C/B	BX
RTS	8	0	7	0
F/U	0	0	0	0
C/B	3	1	9	8
BX	1	0	6	11

	RTS	F/U	C/B	BX
RTS	3	1	0	0
F/U	0	1	0	0
C/B	3	0	3	2
BX	1	1	5	35

	RTS	F/U	C/B	BX
RTS	7	0	7	0
F/U	0	0	0	0
C/B	0	0	9	3
BX	0	0	9	20

B: Analog versus Lossy Compressed Digital: 1.75 bpp

	RTS	F/U	C/B	BX
RTS	10	0	5	2
F/U	0	0	0	0
C/B	3	0	9	9
BX	1	0	1	16

	RTS	F/U	C/B	BX
RTS	1	0	2	1
F/U	0	1	0	0
C/B	1	0	4	4
BX	2	1	5	35

	RTS	F/U	C/B	BX
RTS	8	0	6	1
F/U	0	0	0	0
C/B	1	0	9	2
BX	0	0	3	27

C: Analog versus Lossy Compressed Digital: 0.4 bpp

	RTS	F/U	C/B	BX
RTS	11	0	6	0
F/U	0	0	0	0
C/B	2	0	15	4
BX	1	0	2	16

	RTS	F/U	C/B	BX
RTS	2	0	1	1
F/U	0	1	0	0
C/B	2	1	3	3
BX	1	0	3	38

	RTS	F/U	C/B	BX
RTS	11	0	4	0
F/U	0	0	0	0
C/B	1	1	8	2
BX	1	0	4	25

D: Analog versus Lossy Compressed Digital: 0.15 bpp

Radiologist A Radiologist B Radiologist C

Table 4: Radiologist Agreement Tables

Management Sensitivity and Specifi city: The means and variances of the sensitivity and specificity and the mean of the PVP of the management decisions with respect to the independent gold standard are summarized in Table 5. Level 1 refers to the analog images, level 2 to the uncompressed digital, and levels 3, 4, and 5 refer to those images where the breast section was compressed to 0.15, 0.4 and 1.75 bpp respectively (and where the label was compressed to .07 bpp). In this table, sensitivity, specificity, and PVP are defined relative to the independent gold standard. The table does not show any obvious trends for these parameters as a function of bit rate. Sensitivity is the ratio of the number of cases a judge calls "positive" to the number of cases actually "positive" according to the independent gold standard. Here "positive" is defined as the union of categories F/U, C/B, and BX. A "negative" study is RTS. Sensitivity and specificity can be thought of as binomial issues, and so if the sensitivity is p, then the variance associated with that sensitivity is $p(1-p)$. The standard deviation calculation for PVP is somewhat more complicated and is not included here; because PVP is the ratio of two random quantities (even given the gold standard), an analytical approach to the computation of variance requires approximate statistical methods related to what is called "propagation of errors" in physics.

level	judge	sensitivity		specificity		PVP
		mean	stdev	mean	stdev	mean
1	A	0.826	0.379	0.692	0.462	0.905
1	B	1.000	0.000	0.308	0.462	0.836
1	C	0.913	0.282	0.846	0.361	0.955
2	A	0.886	0.317	0.769	0.421	0.929
2	B	0.955	0.208	0.385	0.487	0.840
2	C	0.932	0.252	0.462	0.499	0.854
3	A	0.814	0.389	0.333	0.471	0.814
3	B	0.953	0.211	0.417	0.493	0.854
3	C	0.977	0.151	0.500	0.500	0.875
4	A	0.860	0.347	0.615	0.487	0.881
4	B	0.955	0.208	0.154	0.361	0.792
4	C	0.977	0.149	0.615	0.487	0.896
5	A	0.841	0.366	0.538	0.499	0.860
5	B	0.953	0.211	0.231	0.421	0.804
5	C	0.932	0.252	0.769	0.421	0.932

Table 5: Sensitivity, specificity and PVP

5 Concluding Remarks

The application of the ideas of the previous chapter to three specific medical image modalities and diagnostic applications provides a quantitative and graphic appreciation of the effects of lossy compression on tasks closely simulating clinical practice. These results alone, however, do not resolve all of the issues involved nor provide a definitive characterization of conditions under which lossy compression might or might not be acceptable. In the next Chapter we conclude our survey of quality evaluation by raising and discussing a variety of underlying statistical issues that arose during these experiments.

Acknowledgements

The authors gratefully acknowledge the essential assistance of many colleagues who participated in and contributed to both the performance of the research described here and the writing of the papers and reports on which these chapters are based. In particular we acknowledge and thank C.N. Adams, A. Aiyer, C. Bergin, B.J. Betts, R. Birdwell, B.L. Daniel, H.C. Davidson, L. Fajardo, D. Ikeda, J. Li, K.C.P. Li, L. Moses, K.O. Perlmutter, S.M. Perlmutter, C. Tseng, and M.B. Williams.

References

[1] C.N. Adams, A. Aiyer, B.J. Betts, J. Li, P.C. Cosman, S.M. Perlmutter, M. Williams, K.O. Perlmutter, D. Ikeda, L. Fajardo, R. Birdwell, B.L. Daniel, S. Rossiter, R.A. Olshen, and R.M. Gray. Evaluating quality and utility of digital mammograms and lossy compressed digital mammograms. In *Proceedings 3rd Intl. Workshop on Digital Mammography*, Chicago, IL, June, 1996.

[2] I. Andersson. Mammography in clinical practice. *Med Radiography and Photography*, 62(2):2, 1986.

[3] P. Armitage. *Statistical Methods in Medical Research*. Blackwell Scientific Publications, Oxford, 1971.

[4] C. Battilana, H. Zhang, R.A. Olshen, L. Wexler, and B.D. Myers. PAH extraction and the estimation of plasma flow in diseased human kidneys. *American Journal of Physiology*, 30:F726–F733, 1991.

[5] J.M. Bramble, L.T. Cook, M.D. Murphey, N.L. Martin, W.H. Anderson, and K.S. Hensley. Image data compression in magnification hand radiographs. *Radiology*, 170:133–136, 1989.

[6] R.O. Brandenburg, V. Fuster, E.R. Giuliani, and D.C. McGoon. *Cardiology: Fundamentals and Practice*. Year Book Medical Publishers, Inc., Chicago, 1987.

[7] P.C. Cosman, H. C. Davidson, C.J. Bergin, C. Tseng, L. E. Moses, E.A. Riskin, R.A. Olshen, and R.M. Gray. Thoracic CT images: effect of lossy image compression on diagnostic accuracy. *Radiology*, 190:517–524, 1994.

[8] P.C. Cosman, C. Tseng, R.M. Gray, R.A. Olshen, L. E. Moses, H. C. Davidson, C.J. Bergin, and E.A. Riskin. Tree-structured vector quantization of CT chest scans: image quality and diagnostic accuracy. *IEEE Trans. Medical Imaging*, 12(4):727–739, Dec. 1993.

[9] D.D. Dershaw, A. Abramson, and D.W. Kinne. Ductal carcinoma in situ: mammographic findings and clinical implications. *Radiology*, 170:411–415, 1989.

[10] B. Efron. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia, 1982.

[11] B. Efron. Better bootstrap confidence intervals and bootstrap approximations. *J. Amer. Stat. Assoc.*, 82:171–185, 1987.

[12] B. Efron and R. Tibshirani. *An introduction to the bootstrap*, volume 57 of *Monographs on Statistics and applied probability*. Chapman & Hall, New York, 1993.

[13] D.A. Freedman. Bootstrapping regression models. *Annals of Statistics*, 9:1218–1228, 1981.

[14] R.M. Gray, R.A. Olshen, D. Ikeda, P.C. Cosman, S. Perlmutter, C. Nash, and K. Perlmutter. Evaluating quality and utility in digital mammography. In *Proceedings ICIP-95*, volume II, pages 5–8, Washington, D.C., October 1995. IEEE, IEEE Computer Society Press.

[15] E.J. Halpern, J.H. Newhouse, E.S. Amis, H.M. Levy, and H.W. Lubetsky. Evaluation of a quadtree-based compression algorithm with digitized urograms. *Radiology*, 171:259–263, 1989.

[16] D.M. Ikeda, I. Andersson, C. Wattsgård L. Janzon, and F. Linell. Radiographic appearance and prognostic consideration of interval carcinoma in the Malmö mammographic screening trial. *Amer. J. Roentgenology*, 159:287–294, 1992.

[17] J.K.T. Lee, S.S. Sagel, and R.J. Stanley. *Computed Body Tomography with MRI correlation*, volume 2. Raven Press, New York, 1989. Second Edition.

[18] E.L. Lehmann. *Testing Statistical Hypotheses*. John Wiley & Sons, New York, 1986. Second Edition.

[19] H. MacMahon, K. Doi, S. Sanada, S.M. Montner, M.L. Giger, C.E. Metz, N. Nakamori, F. Yin, X. Xu, H. Yonekawa, and H. Takeuchi. Data compression: effect on diagnostic accuracy in digital chest radiographs. *Radiology*, 178:175–179, 1991.

[20] R.G. Miller, Jr. *Simultaneous Statistical Inference*. Springer-Verlag, New York, 1981. Second Edition.

[21] R.G. Miller, Jr. *Beyond ANOVA, Basics of Applied Statistics*. John Wiley and Sons, New York, 1986.

[22] R.A. Olshen, E.N. Biden, M.P. Wyatt, and D.H. Sutherland. Gait analysis and the bootstrap. *Annals of Statistics*, 17:1419–1440, 1989.

[23] S.M. Perlmutter, P.C. Cosman, R.M. Gray, R.A. Olshen, D. Ikeda, C.N. Adams, B.J. Betts, M. Williams, K.O. Perlmutter, J. Li, A. Aiyer, L. Fajardo, R. Birdwell, and B.L. Daniel. Image quality in lossy compressed digital mammograms. *Signal Processing*, 59:189–210, June 1997.

[24] S.M. Perlmutter, P.C. Cosman, C. Tseng, R.A. Olshen, R.M. Gray, K.C.P. Li, and C.J. Bergin. Medical image compression and vector quantization. *Statistical Science*, 13(1):30–53, Jan. 1998.

[25] S.M. Perlmutter, C. Tseng, P.C. Cosman, K.C.P. Li, R.A. Olshen, and R.M. Gray. Measurement accuracy as a measure of image quality in compressed MR chest scans. In *Proceedings ICIP-94*, volume 1, pages 861–865, Austin, TX, Nov. 1994. IEEE Computer Society Press.

[26] J. Sayre, D. R. Aberle, M. I. Boechat, T. R. Hall, H. K. Huang, B. K. Ho, P. Kashfian, and G. Rahbar. Effect of data compression on diagnostic accuracy in digital hand and chest radiography. In *Proceedings of Medical Imaging VI: Image Capture, Formatting, and Display*, volume 1653, pages 232–240. SPIE, Feb. 1992.

[27] G. W. Snedecor and W. G. Cochran. *Statistical Methods*. Iowa State University Press, Ames, Iowa, 1989.

[28] D.D. Stark and W.G. Bradley, Jr. *Magnetic Resonance Imaging*. Mosby-Year Book, Inc., St. Louis, 1992. Second Edition.

[29] P.C. Stomper, J.L. Connolly, J.E. Meyer, and J.R. Harris. Clinically occult ductal carcinoma in situ detected with mammography: analysis of 100 cases with radiographic-pathologic correlation. *Radiology*, 172:235–241, 1989.