Quality evaluation for compressed medical images: Statistical Issues

Pamela Cosman, Robert Gray, Richard Olshen

1 Introduction

In compressing a radiological image, the fundamental question is: will the compressed image still be as diagnostically useful as the original? In the previous chapter, we presented several clinical studies, experimental protocols, and statistical analysis techniques. Taken together, these provide a methodology for answering this type of question, and they also provide the answer to the question in the context of particular images sets, compression algorithms, and diagnostic tasks. There remain, however, a number of questions which must be addressed in any study of this type. For example, was the experiment designed well enough? Is the answer different for different radiologists? How does diagnostic utility relate to other measures of image quality? In this chapter we present various statistical approaches to these broad questions. We first discuss statistical size and power, and learning effects, both of which speak to the question of whether the clinical experiment was well designed. Next, we present a comparison of judges, and we discuss how diagnostic utility can be related to other measures of image quality.

2 Statistical Size and Power

The *size* of a test is the probability of incorrectly rejecting the null hypothesis if it is true. The *power* of a test is the probability of correctly rejecting the null hypothesis if it is false. For a given hypothesis and test statistic, one constrains the size of the test to be small and attempts to make the power of the test as large as possible.

Given a specificed size, test statistic, null hypothesis, and alternative, statistical power can be estimated using the common (but sometimes inappropriate) assumption that the data are Gaussian. As data are gathered, however, improved estimates can be obtained by modern computer intensive statistical methods. For example, the power and size can be computed for each test statistic described above to test the hypothesis that digital mammography of a specified bit rate is equal or superior to film screen mammography with the given statistic and alternative hypothesis to be suggested by the data. In the absence of data, we can only guess the behavior of the collected data to approximate the power and size. We consider a one-sided test with the "null hypothesis" that, whatever the criterion (management or detection sensitivity, specificity, or predictive value positive (PVP)), the digitally acquired mammograms or lossy compressed mammograms of a particular rate are worse than analog. The "alternative" is that they are better. In accordance with standard practice, we take our tests to have size .05. We here focus on sensitivity and specificty of management decisions, but the general approach can be extended to other tests and tasks.

Approximate computations of power devolve from the 2 by 2 agreement tables of the form of Table 1. In this table, the rows correspond to one technology (for example analog) and columns to the other (digital, say). "R" and "W" correspond to "right" (agreement with gold standard) and "wrong" (disagreement with gold standard). So, for example, the count $N(1,1)$ is the number of cases where a radiologist was right when reading both the analog and digital images. The key idea is twofold. In the absence of data, a guess as to power can be computed using standard approximations. Once preliminary data are obtained, however, more accurate estimates can be obtained by simulation techniques taking advantage of the estimates inherent in the data. Table 2 shows the possibilities and their corresponding probabilities. The right hand column and bottom row are sums of what lies, respectively, to the left and above them. Thus, ψ is the value for one technology and $\psi + h$ is the value for the other; $h = 0$ denotes no difference. It is the null hypothesis. The four entries in the middle of the table are parameters that define probabilities for a single study. They are meant to be average values across radiologists, as are the sums that were cited. Our simulations allow for what we know to be the case: radiologists are very different in how they manage and in how they detect.

Two fundamental parameters are γ and R. The first is the chance (on average) that a radiologist is "wrong" for both technologies; R is the number of radiologists. These key parameters can be estimated from the counts of the 2 by 2

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

Table 1: Agreement 2×2 Table

II\ I	Right	Wrong	
Right	$2\psi + h - 1 + \gamma$	$1 - \psi - h - \gamma$	ψ
Wrong	$1 - \psi - \gamma$	γ	$1 - \psi$
	$\psi + h$	$1 - \psi - h$	1

Table 2: Management Outcome Probabilities

agreement table resulting from the pilot experiment, and then improved as additional data are acquired.

In our small pilot study of management, we found sensitivity of about .60 and specificity about .55. The respective estimated values of h varied from more than .02 to about .07; γ was about .05. These numbers are all corrupted by substantial noise. Indeed, the variability associated with our estimation of them is swamped by the evident variability among radiologists. For a test of size .05, by varying parameters in amounts like what we saw, the power might be as low as .17 with 18 radiologists, or as high as 1.00 with only 9 radiologists. The power is very sensitive to the three parameters. No matter how many studies or how many radiologists we would have, one could always vary the parameters so that we would need more of either or both.

If we think of sensitivity for detection being .85, say, then at least for that quantity 400 studies and 9 radiologists seem ample. At this time one good recommendation would be to start with 400 studies, 12 radiologists, three at each of four centers, and find an attained significance level for a test of the null hypothesis that there is no difference between technologies. And, perhaps at least as important, estimate the parameters of Table 2. At that point possible numbers of required further radiologists or studies, if any, could be estimated for particular values of size and power that reviewers might require. The design could be varied so that the pool of studies would include more than 400, but no single radiologist would read more than 400. In this way we could assess fairly easily the impact of variable prevalence of adverse findings in the gold standard, though we could get at that issue even in the situation we study here.

Computations of power apply equally well in our formulation to sensitivity and specificity. They are based on a sample of 400 studies for which prudent medical practice would dictate *return to screening* for 200, and something else (*six month followup*, *additional assessment needed*, or *biopsy*) for the other 200. Thus, there are 200 studies that figure in computation of sensitivity and the same number for specificity. All comparisons are in the context of “clinical management,” which can be “right” or “wrong.” It is a given that there is an agreed upon *gold standard*, independent or separate. For a given radiologist who has judged two technologies – here called I and II and meant to be digital and analog or analog and lossy compressed digital in application – a particular study leads to an entry in a 2 by 2 agreement table of the form of Table 1.

If the null hypothesis of “no difference in technologies” is true, then whatever be ψ and γ, $h = 0$. An alternative hypothesis would specify $h \neq 0$, and without loss (since we are free to call whichever technology we want I or II) we may take $h > 0$ under the alternative hypothesis that there is a true difference in technologies. Under the null, *given* $b + c$, b has a binomial distribution with parameters $b + c$ and 1/2. Under the alternative, given $b + c$, b is binomial with parameters $b + c$ and $(1 - \psi - \mathrm{h} - \gamma)/(2 - 2\psi - 2\gamma - \mathrm{h})$. The usual McNemar *conditional* test of the null hypothesis is based on $(b - c)^2/(b + c)$ having approximately a chi-square distribution with one degree of freedom.

In actual practice we intend to use R radiologists for $R = 9, 12, 15$, or 18, to assume that their findings are independent, and to combine their data by adding the respective values of their McNemar statistics. We always intend that the *size* = probability of Type I error is .05. Since the sum of independent chi-square random variables is distributed as chi-square with degrees of freedom the sum of the respective degrees of freedom, it is appropriate to take as the critical value for our test the number C, where $\Pr(\chi^2_R > C) = .05$. The four respective values of C are therefore 16.92, 21.03, 25.00, and 28.87.

Computation of power is tricky because it is *unconditional* since before the experiment, $b + c$ for each radiologist is random. Thus, the power is the probability that a non-central chi-square random variable with R degrees of freedom and non-centrality parameter $[(p_1 - .5)^2/p_1 q_1] \sum_{i=1}^{R}(b_i + c_i)$ exceeds $C/4p_1 q_1$, where $b_i + c_i$

	Sample 2			
		A	not A	
Sample 1	A	k	r	$k+r$
	not A	s	m	$s+m$
		$k+s$	$r+m$	N

Table 3: 2×2 table of pairs in the McNemar analysis

has a binomial distribution with parameters N and $2 - 2\psi - 2\gamma - h$; and the R random integers are independent; $p_1 = (1-\psi-h-\gamma)/(2-2\psi-2\gamma-h) = 1-q_1$. This entails that the non-centrality parameters of the chi-square random variable that figures in the computation of power is itself random. Note that a non-central chi-square random variable with R degrees of freedom and non-centrality parameter Q is the distribution of $(G_1 + Q^{1/2})^2 + G_2^2 + \cdots + G_R^2$, where $G_1, ...G_R$ are independent, identically distributed standard Gaussians. On the basis of previous work and pilot study, we have chosen to compute the power of our size .05 tests for N always 200, ψ from .55 to .85 in increments of .05; $\gamma = .03, .05, .10$; and, as was stated, $R = 9, 12, 15$, and 18. The simulated values of power can be found in [2] and code for carrying out these computations is in Appendix C of [1]. These form the basis of our earlier estimates for the necessary number of patients and will be updated as data is acquired.

3 Analysis of Learning Effects

In experiments of this type, the radiologists see an image at many compression levels during the course of the study. One needs to ascertain whether learning effects are significant. Learning and fatigue are both processes that might change the score of an image depending on when it was seen. For the CT study, the McNemar test [8] was used to examine this possibility.

Suppose one has N observations of *paired* data. The members of the pair are called sample 1 and sample 2. Each member of the pair can be described as being “A” or “not A.” There are clearly 4 types of pairs: those with both members of type “A,” those with neither member of type “A,” those where the first member of the pair is of type “A” but the second is not, and those where the second member is of type “A” but the first is not. The last two types are referred to as disparate or “untied” pairs. We denote the number of occurrences of each type of these four pairs by k, m, r, and s, as shown in a 2×2 table, in Table 3.
The proportions of individuals of type “A” in the 2 samples are

$$\frac{k+r}{N} \quad \text{and} \quad \frac{k+s}{N},$$

and the difference between the two proportions is

$$\frac{r-s}{N}.$$

The null hypothesis that there is no difference between the proportions of type “A” individuals in the two populations is

$$E\left(\frac{r-s}{N}\right) = 0 \qquad \Rightarrow \qquad E(r) = E(s).$$

Denote $r + s$ by n. If the null hypothesis holds, given n disparate or “untied” pairs, the number of pairs of type 2 (or of type 3) would follow a binomial distribution with parameter equal to 1/2. Typically, a large sample test is obtained by regarding the quantity

$$u = \frac{r - \frac{n}{2}}{\frac{1}{2}\sqrt{n}}$$

as a standardized normal deviate. However, in this study we make no assumption of normality. This McNemar analysis is applied to study intrasession learning effects in the CT study as follows: In each session, each image was seen at

		2nd occurrence		
		Perfect	Not Perfect	
First occurrence	Perfect	53	4	57
	Not Perfect	9	5	14
		62	9	71

Table 4: Judge 1, pairing for CT lung nodules

exactly 2 levels, and the ordering of the pages ensured that they never appeared with fewer than 3 pages separating them. For each judge J and each session S and each image I, we pair the judge's reading for a given compression level L_1 with the same judge's reading for compression level L_2 for the same image and same session, where L_1 was seen *before* L_2. For each member of the pair, the reading is either perfect (sensitivity = 1 and PVP = 1, type "A") or not perfect (type "not A"). For example, Judge 1 in evaluating lung nodules over the course of 3 sessions saw 71 pairs of images, in which an image seen at one compression level in a given session is paired with the same image seen at a different level in the same session. Of the 71 pairs, 53 times both images in the pair were judged perfectly, and 5 times both images were judged incorrectly.

We concern ourselves with the other 13 pairs: 9 times the image seen first was incorrect while the second one was correct, and 4 times the image seen second was incorrect when the first one was correct. If it did not matter whether an image was seen first or second, then conditional on the numbers of the other two types, these would have a binomial distribution with parameters 13 and $1/2$. This example is shown in Table 4. The probability that a fair coin flipped 13 times will produce a heads/tails split at least as great as 9 to 4 is 0.267, thus this result is not significant. These calculations were carried out for $2 \times 4 \times 2 = 16$ different subsets of the data (lungs vs. mediastinum (2), Judges 1, 2, 3 considered separately or pooled together (4), consensus or personal gold standards (2)), and in no case was a significant difference found at the 5% significance level (p-values ranged from 0.06 to 1.0). An analysis of variance using the actual sensitivity and PVP observations (without combining them into "perfect" and "not perfect") similarly indicated that page order and session order had no significant effect on the diagnostic result.

Learning Effects for the Mammography Experiment: In the mammography experiment, the radiologists saw each study at least 5 times during the complete course. These 5 versions were the analog originals, the digitized versions, and the 3 wavelet compressed versions. Some images would be seen more than 5 times, as there were JPEG compressed images, and there were also some repeated images, included in order to be able to directly measure intra-observer variability.

In this work, we looked for whether learning effects were present in the management outcomes using what is known in statistics as a "runs" test. We illustrate the method with an example. Suppose a study was seen exactly five times. The management outcomes take on four possible values (RTS, F/U, C/B, BX). Suppose that for a particular study and radiologist, the observed outcomes were BX three times and C/B two times. If there were no learning, then all possible "words" of length five with three BX's and two C/B's should be equally likely. There are 10 possible words that have three BX's and two C/B's. These words have the outcomes ordered by increasing session number; that is, in the chronological order in which they were produced. For these 10 words, we can count the number of times that a management outcome made on one version of a study differs from that made on the immediately previous version of the study. The number ranges from one (e.g., BX BX BX C/B C/B) to four (BX C/B BX C/B BX). The expected number of changes in management decision is 2.4, and the variance is 0.84. If the radiologists had learned from previous films, one would expect that there would be fewer changes of management prescription than would be seen by chance. This is a conditional runs test, which is to say that we are studying the conditional permutation distribution of the runs.

We assume that these "sequence data" are independent across studies for the fixed radiologist, since examining films for one patient probably does not help in evaluating a different patient. So we can pool the studies by summing over studies the observed values of the number of changes, subtracting the summed (conditional) expected value, and dividing this by the square root of the sum of the (conditional) variances. The attained significance level (p-value) of the resultant Z value is the probability that a standard Gaussian is $\leq$ Z.

Those studies for which the management advice never changes have an observed number of changes 0. Such studies are not informative with regard to learning, since it is impossible to say whether unwavering management advice is the result of perfect learning that occurs with the very first version seen, or whether it is the result of the obvious alternative, that the study in question was clearly and independently the same each time, and the radiologist simply interpreted it the same way each time. Such studies, then, do not contribute in any way to the computation of the statistic. The JPEG versions and the repeated images, which are ignored in this analysis, are believed to make this analysis and p-values actually conservative. If there were no learning going on, then the additional versions make no difference. However, if there were learning, then the additional versions (and additional learning) should mean that there would be even fewer management changes among the 5 versions that figure in this analysis.

The runs test for learning did not find any learning effect at the 5% significance level for the management outcomes in the mammography experiment. For each of the 3 judges, approximately half of the studies were not included in the computation of the statistic, since the management decision was unchanging. For the 3 judges, the numbers of studies retained in the computation were 28, 28, and 27. The Z values obtained were -0.12, -0.86, and -0.22, with corresponding p-values of 0.452, 0.195, and 0.413.

4 Comparison of judges

In the CT study, comparisons of judges to each other were carried out using the permutation distribution of Hotelling's paired T^2 statistic applied to the consensus gold standard results. T^2 as we used it is a generalization of (the square of) a univariate paired t statistic. We illustrate its use by an example. Suppose that Judges 1 and 2 are compared for their sensitivities on compressed lung images. The vector for comparison is 6-dimensional, one coordinate for each level of compression. Each image (i) and bit rate (b) evaluated by both judges gives rise to a difference $d(i,b)$ of the sensitivities, Judge 1 $-$ Judge 2, and to a sample mean $\bar{d}(b)$ and sample variance $s^2(b)$. Each image i for which both judges evaluated at bit rates b and b' contributes a term

$$(d(i,b) - \bar{d}(i,b))(d(i,b') - \bar{d}(i,b'))$$

to the sample covariance $s(b,b')$. Write $\bar{D}$ for the column vector with bth coordinate $\bar{d}(b)$, and $\hat{S}$ for the 6 × 6 matrix with b, b' coordinate $s(b,b')$. The version of T^2 we use is

$$T^2 = \bar{D}'\hat{S}^{-1}\bar{D}.$$

It differs from the usual version [7] by a norming constant that implies an F distribution for T^2 when $\{d(i,b)\}$ are jointly Gaussian and the numbers of (b,b') pairs are equal. As our data are decidedly nonGaussian, computations of attained significance are again based on the permutation distribution of T^2 [7], though only on 999 permutations plus the unpermuted value and not on the full distribution, which is neither computationally feasible nor necessary.

The permutation distribution is motivated by the fact that, were there no difference between the judges, then in computing the difference $d(i,b)$, it should not matter whether we compute Judge 1 $-$ Judge 2 or vice versa, or whether we randomize the choice with a fair coin toss. The latter is exactly what we do, but we constrain the randomization so that for fixed i, the signs of $\{d(i,b)\}$ are all the same. The constraint tends to preserve the covariance structure of the set of differences, at least when the null hypothesis of no difference is approximately true. (Unconstrained randomization would render the signs of $d(i,b)$ and $d(i,b')$ independent, and this is clearly not consistent with the data.) After randomizing the signs of all differences, we compute T^2 again; the process is repeated a total of 999 times. There results a list of 1000 T^2 values, the "real" (unpermuted) one and 999 others. Were there no difference between the judges, the 1000 values would be (conditional on the data) independent and identically distributed. Otherwise, we expect the "real" value to be larger than at least most of the others. The attained significance level for the test of the null hypothesis that there is no difference between the judges is (k+1)/1000, where k is the number of randomly permuted T^2 values that exceed the "real" one.

Some comparisons we hoped to make with T^2 were not possible to compute because not only are the $d(i,b)$ not independent, but also $\hat{S}$ was singular. We could have extended the domain of applicability of the Hotelling T^2 approach to the case when the covariance matrix is not invertible by making an arbitrary choice of a pseudo-inverse. This is not a customary approach to T^2 in the usual Gaussian case, and also the inferences we draw are quite clear without resorting to that technique.

The actual p-value for the comparison of Judges 1 and 2 for their sensitivity in finding lung nodules is not significant, and the same is true for comparisons of Judges 1 and 3 and that of Judges 2 and 3. Two of the three comparisons of predictive value positive for the lung are not significant; for the other (1 versus 2) it is not possible to compute because $\hat{S}$ is singular. The analogous comparisons for the mediastinum give rather different results. Judge 2 seems to differ from both other judges in sensitivity (both p-values about .04). Judge 2 also seems to differ from Judge 3 in predictive value positive at the same p-value. Similar results were obtained from a 7-dimensional comparison, in which the additional coordinate comes from data on the original images. The basic message is that judges seem to differ from one another in judging the mediastinum but not the lung.

5 Relationships between quality measures

As image quality can be quantified by diagnostic accuracy, subjective ratings, or computable measures such as signal-to-noise ratio (SNR), one key question concerns the degree to which these different measures agree. Verifications of medical image quality by perceptual measures require the detailed, time-consuming, and expensive efforts of human observers, typically highly trained radiologists. Therefore, it is desirable to find computable measures that strongly correlate with or predict the perceptual measures.

In previous sections we have studied how certain parameters such as percent measurement error and subjective scores appear to change with bit rate. It is assumed that bit rate has a direct effect on the likely measurement error or subjective score and therefore the variables are correlated. In this sense, bit rate can also be viewed as a “predictor”. For instance, a low bit rate of 0.36 bits per pixel (bpp) may “predict” a high percent measurement error or a low subjective score. If the goal is to produce images which lead to low measurement error, parameters which are good predictors of measurement error are useful for evaluating images as well as for evaluating the effect of image processing techniques. A “good” predictor is a combination of an algorithm and predictor variable which estimates the measurement error within a narrow confidence interval.

Percent measurement error can be predicted from other variables besides bit rate. The graphs below give an indication of whether subjective scores, SNR, or image distortion are good predictors of measurement error. For instance, does a high subjective score or high SNR generally lead to low percent measurement error? We plot percent measurement error against each predictor variable of interest in Figures 1, 2, and 3. Subjective scores and SNR are as defined in previous chapters and MSE distortion is taken to be the average non-normalized squared distortion between the original and compressed image.

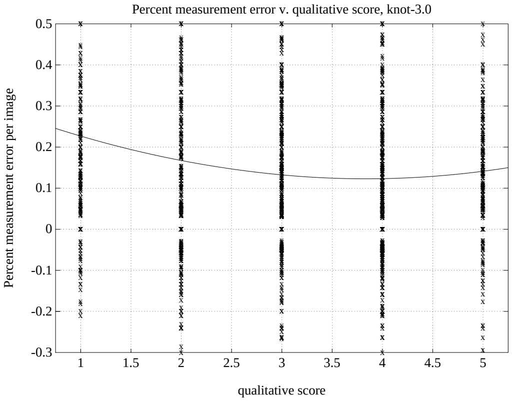

Figure 1: Percent measurement error vs. subjective score for the MR study

How does one quantify whether or not some variable is a good predictor? In the remainder of this section, we examine the usefulness of SNR as a predictor of subjective quality for the MR data set. Our work suggests that cross-validated fits to the data using generalized linear models can be used to examine the usefulness of computable measures as predictors for human-derived quality measures. In the example studied below, the computable measure is SNR, and the human-derived measure is subjective ratings, but the method presented is applicable to other types of prediction problems.

In the classical linear regression model, the “predictor” x is related to the outcome y by

$$y = \beta^t x + \epsilon, \tag{1}$$

where β is a vector of unknown coefficients, and the error ϵ at least has mean zero and constant variance, or may even be normally distributed. In the regression problem of using SNR to predict subjective quality scores, the response

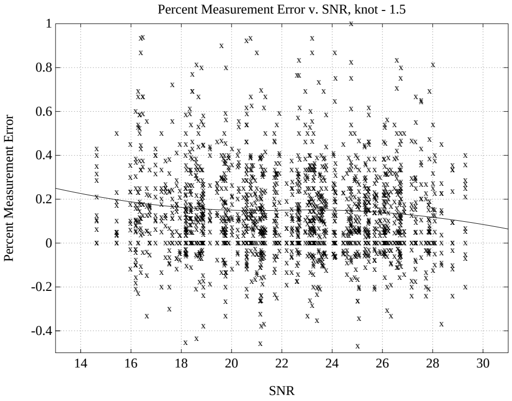

Figure 2: Percent measurement error vs. SNR for the MR study

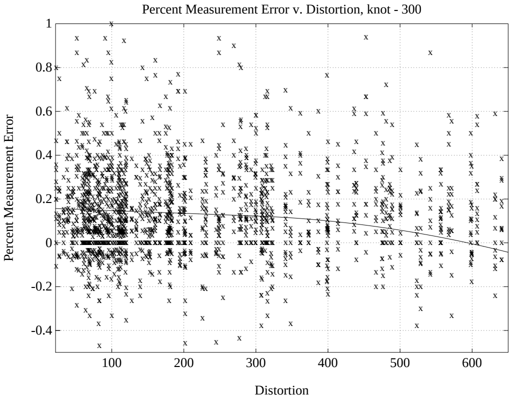

Figure 3: Percent measurement error vs. MSE for the MR study

variable y takes on integer values between 1 and 5, and so the assumption of constant variance is inappropriate because the variance of y depends on its mean. Furthermore, y takes on values only in a limited range, and the linear model does not follow that constraint without additional untenable assumptions. We turn to a generalized linear model that is designed for modeling binary and, more generally, multinomial data [6].

A generalized linear model requires two functions: a link function that specifies how the mean depends on the linear predictors, and a variance function that describes how the variance of the response variable depends on its mean. If $X_1, X_2, ...X_n$ are independent Poisson variables, then conditional upon their sum, their joint distribution is multinomial. Thus the regression can be carried out with the Poisson link and variance functions:

$$\beta^t x = \ln \mu \quad \text{and} \quad var(y) = \mu \tag{2}$$

in which case the mean of the response variable is

$$\mu = e^{\beta^t x}. \tag{3}$$

The results of this approach are shown in Figure 4. The predictors are a quadratic spline in SNR:

$$x = (1, snr, snr^2, [max(snr - snr_0, 0)]^2) \tag{4}$$

where the spline knot snr_0 was chosen to be 22.0 (the average SNR value of the data set). In Figure 4, the x symbols denote the raw data pairs (subjective score, SNR) for the judges pooled, and the curve is the regression fit. The o symbols denote the 95% confidence intervals obtained from the bootstrapped BC_a method [4, 5]. This method is outlined below. The null deviance (a measure of goodness of fit) of the data set is 229 on 449 degrees of freedom, and the residual deviance of the fit is 118 on 446 degrees of freedom, indicating a useful fit. The model parameters were estimated using the statistical software S, which uses iteratively reweighted least-squares to generate a maximum-likelihood estimate. The data for all 3 judges were pooled because an analysis of variance (ANOVA) determined that the effect of judges was not significant at the 5% level. In the ANOVA, judges, images, and bit rates were taken to be fixed effects.

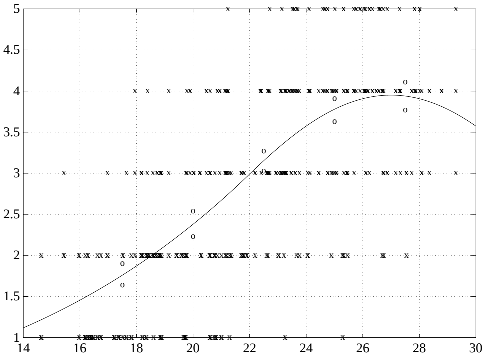

Figure 4: Expected subjective score (y-axis) vs. SNR (x-axis) (Permission for reprint, courtesy Society for Information Display)

Instead of doing a fit directly to the expectation of the response, a second way to approach this problem looks for the probability p_i of obtaining the response i, for each of the five possible responses ($i = 1, \ldots, 5$). The expectation can then be calculated from the probabilities. We can transform the responses y into binary outcomes:

$$y_i = \begin{cases} 1 & \text{if } y = i \\ 0 & \text{otherwise.} \end{cases}$$

The binary response variables y_i can then each be fitted using the *logit* link:

$$\beta^t x = \ln \frac{\mu}{1 - \mu} \tag{5}$$

in which case the mean of the response variable is

$$\mu = \frac{e^{\beta^t x}}{1 + e^{\beta^t x}} \tag{6}$$

which guarantees that μ is in the interval [0,1]. The logit link together with the binomial variance function $\mu(1-\mu)$ defines the *logistic regression* model. For each y_i the predictor x was a quadratic spline in SNR, with the knots located in each case at the mean value of the SNRs which produced that response (18.2, 20.12, 22.57, 24.61, 25.56). The probabilities p_i are shown in Figure 5 with vertical offsets so they are not superimposed.

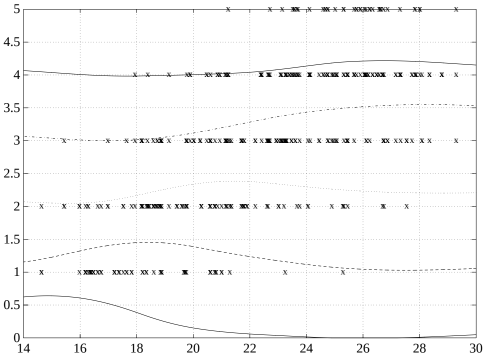

Figure 5: Response probabilities (y-axis) vs. SNR (x-axis) (Permission for reprint, courtesy Society for Information Display)

As the five probabilities have been determined from separate regressions, they need to be scaled so that they add to one before calculating $E(y)$ from them. The logistic gives a value for the $\log(odds)$, that is, for a given value of SNR = s we obtain

$$t = \log \frac{p_i}{p(\text{not } i)}.$$

Exponentiating both sides and rearranging terms yields

$$p_1 - e^t(p_2 + p_3 + p_4 + p_5) = 0 \text{ when } i = 1.$$

For that value of SNR, similar equations can be found for p_2, p_3, p_4, and p_5. Additionally, we know that $\sum_i p_i = 1$. This system can be solved and the expectation calculated from these scaled probabilities:

$$E(y) = \sum_i i p_i.$$

For some of the p_i there are slight edge effects from the spline fit. For example, p_1 dips very slightly below zero at SNR = 24.9, and then becomes slightly positive again for SNRs > 27.2, although there are no further reponses of 1 at those SNRs. Until we have made a further study of these edge effects, they are dealt with simply by setting p_i identically equal to zero beyond the point where it first crosses zero. The expectation is then calculated from these windowed probabilities. The expectation is almost indistinguishable from the curve of Figure 4, thereby validating the Poisson model.

Having established the appropriateness of the Poisson model, we use it to compare SNR against segmental SNR in their ability to predict subjective quality ratings. Segmental SNR, often used in speech quality evaluation, compensates for the under-emphasis of weak-signal performance in conventional SNR. An image is divided into blocks, the SNR is calculated for each block on a log scale, thresholded below at 0 and above at 45, and the values are averaged. By converting component SNR values to dB values prior to averaging, very high SNR values corresponding to well-coded large-signal segments do not camouflage coder performance with the weak segments, as in conventional SNR. We examined block sizes of all powers of 2 between 2×2 and 256×256. Since the images are of size 256×256, the segmental SNR for that block size equals the conventional SNR. The usefulness of the computable metric in predicting subjective quality was examined as follows: For $n = 20$ times, the 30 MR images were put in a different random order. Each time, a 10-fold cross-validation was performed in which 3 images at a time were left out, and the other 27 images were used to fit the model. All judges and levels corresponding to those 27 images were used. The 3 images not involved in determining the parameters of the fit comprise 45 data points (3 images $\times$ 3 judges $\times$ 5 compression levels). For these data we compute the mean outcome and the sum of squared deviation from this overall mean. This value is called S_1. Then we calculate the fitted values for these data, and take the sum of squared deviations of observed and fitted, called S_2. If the model is good and the test set of 3 images is not unlike the set of 27 images used to fit the model, we expect S_2 to be smaller than S_1. The percent reduction in mean squared error that owes to fitting the model (beyond fitting an overall constant) is a statistic that summarizes the model's predictive power:

$$M = 100(1 - (S2/S1))\%.$$

This statistic is a cross-validated analogue of the multiple correlation. The results are presented in Table 5.

Block Size	M
256 × 256	42.96
128 × 128	42.29
64 × 64	34.49
32 × 32	46.48
16 × 16	47.72
8 × 8	48.10
4 × 4	46.62
2 × 2	47.21
l_1	38.60
l_3	35.08

Table 5: Comparison of computable quality measures

It appears that segmental SNR at several different block sizes outperforms conventional SNR. The best of these (on 8 × 8 blocks) produced a 48% reduction compared to the 43% reduction for SNR. One could examine the statistical significance of these differences by sampling from the permutation distribution, and it would be of interest to compare SNR against perceptually based computable quality measures.

In studies like ours, one frequently wants a measure of the predictive power of the model, as well as measures of its goodness of fit. One diagnostic as to the appropriateness of the Poisson regression model (how median-biased it is) is z_0 (described in Section 13); zero is a "good" value. For us, values for z_0 for our five confidence intervals ranged from -0.043 to 0.059, with a median of -0.012. The correlation between observed and fitted values is a statistic that summarizes the model's predictive power. But the number computed from the data that gave rise to the model (0.70) can be overly optimistic. There are many approaches to getting around that optimism, a simple one being 10-fold cross-validation, as in [3]. To implement 10-fold cross-validation one divides the data set at random into 10 distinct parts. Nine of them would be used to fit the generalized linear model, and the correlation coefficient between actual and fitted values would be calculated for the tenth part, that is, the part that did not figure in the fitting. This would be repeated 10 times in succession, and the resulting ten values of the correlation averaged. Other sample reuse methods can be used to accomplish the same task.

5.1 BC$_a$ confidence intervals

The BC$_a$ confidence intervals for fixed values of SNR were obtained by a bootstrapping method in which images are the sampling units. Suppose that there are B bootstrap samples. We took B = 2000. Each bootstrap sample was generated by sampling randomly (with replacement) 450 times from our set of 450 (subjective score, SNR) pairs. Huge computational savings can be realized in that, for the same set of images being sampled, one bootstrap sample can be used simultaneously for different SNRs. For a fixed SNR, $\hat{E}$ is the fitted expected subjective score based on the model as computed for the original data, and $\hat{E}^{*(b)}$ is the value computed for the bth bootstrap sample. The 100(1-2α)% BC$_a$ confidence interval will be of the form

$$(\hat{E}^{*(\alpha_1)}, \hat{E}^{*(\alpha_2)}),$$

where $\hat{E}^{*(\alpha)}$ is the 100αth percentile of the bootstrap distribution of $\hat{E}$; α_1 and α_2 are defined by

$$\alpha_1 = \Phi(z_0 + \frac{z_0 + z^{(\alpha)}}{1 - a(z_0 + z^{(\alpha)})}),$$

$$\alpha_2 = \Phi(z_0 + \frac{z_0 + z^{(1-\alpha)}}{1 - a(z_0 + z^{(1-\alpha)})}).$$

Φ is the standard normal cumulative distribution function, and $z^{(\beta)}$ is the 100βth percentile (so, for example, $z^{(.95)}$ = 1.645). The "bias correction" z_0 and "acceleration constant" a remain to be defined. $\hat{E}^{*(b)}$ is the value of $\hat{E}$ for the bth bootstrap sample, and $\hat{E}$ is the value computed for the original data. Then

$$z_0 = \Phi^{-1}(\frac{\#\{\hat{E}^{*(b)} < \hat{E}\}}{B}),$$

where Φ^{-1} is the Gaussian quantile function (so, for example, Φ^{-1}(.95) = 1.645). Suppose that there are n images in all, and let $\hat{E}_{(i)}$ be the computed value of $\hat{E}$ when the ith image is deleted (so the computation is done on $n-1$ images). Let

$$\hat{E}_{(.)} = \frac{1}{n}\sum_{i}^{n} \hat{E}_{(i)}.$$

Then

$$a = \frac{\sum_{i=1}^{n}[\hat{E}_{(.)} - \hat{E}_{(i)}]^3}{6\{\sum_{i=1}^{n}[\hat{E}_{(.)} - \hat{E}_{(i)}]^2\}^{3/2}}.$$

There are two main differences between the BC_a confidence intervals described here and the Scheffé confidence intervals described in the previous chapter. The Scheffé method produces a *simultaneous* confidence interval, that is, one that provides upper and lower limits for the entire curve at once. The BC_a method supplies *pointwise* intervals, valid for specific points along the x-axis. It is not currently known how to extend the BC_a method to simultaneous intervals. The second difference is that the Scheffé intervals are always symmetric about the curve, regardless of whether there are any constraints on the range of the variables. Sensitivity and PVP, for example, have a maximum value of 1. The expected value of sensitivity at a particular bit rate may be very close to that upper limit, and when obtaining a Scheffé confidence interval for that curve, the confidence interval may exceed 1, since it is necessarily symmetric about the curve. In that case, the upper confidence curve must be thresholded at 1. The BC_a method has the advantage of providing intervals that are not necessarily symmetric, but respect the fact that the values of the response variable lie within a small constrained range.

6 Philosophical issues

There are many different perspectives from which these different measures of image quality can be viewed. They vary in the extent to which they explicitly consider the application for which the images are used. At one extreme are the computable measures such as SNR, which in no way take account of the medical nature of the images. Subjective ratings in which a radiologist is asked to rate the medical usefulness of an image begin to address the issue. ROC analysis, which includes both a (generally) binary diagnostic decision and a subjective confidence ranking associated with that diagnosis, are serious attempts to capture the medical interest of the images through their diagnostic value. Studies such as the CT detection task and MR measurement task attempt to reproduce very closely some actual clinical diagnostic tasks of radiologists, and to ask the fundamental question of whether a diagnosis made on a compressed image is as good as one made on an original. By this measure, an image has high quality if the number and locations of lesions one finds there precisely match the number and locations one finds on the original (or what the independent panel finds on the original). But is that really the fundamental question? A diagnosis is made on a patient's scan in order to make a decision about medical care for that patient, so perhaps image quality could be defined in terms of medical care. That is, an image has high quality if the decision on medical care is unchanged from that determined upon the original. So if the original image has 6 nodules and the compressed one has 9, that may still be an extremely high quality image according to this particular measure, because the decision regarding medical care may be unaltered in the case of many tumors with a few more or less. One can step back further to look at patient outcome rather than decision regarding medical care. Suppose hypothetically that one designs a classification scheme to highlight suspected tumors in an image. And perhaps, unbeknownst to the designers, pre-cancerous cells which have an overlapping intensity distribution with that of cancerous cells also tend to get highlighted, causing the surgeon to make a wider resection and have lower recurrence rates. Then the processed image might rate as poorer quality than an original based on the previous measures (because both diagnosis and medical care decision would be different from those based on the original image), yet the processed image would rate as top quality according to the measure of improved patient outcome. No one would seriously propose these as measures of image quality. The decision on medical care and the patient outcome both depend on far too many factors other than just image quality. And yet, if one considers the true measure of medical image quality to be simply whether a diagnosis on the processed image is unchanged from the diagnosis on the original, one denies the possibility that the processing may in fact *enhance* the image. This is not a worrisome consideration with image compression, although there is some indication that in fact slightly vector quantized images are superior to originals because noise is suppressed by a clustering algorithm. However, this may soon be a difficult issue in evaluating the quality of digitally processed medical images where the processing is, for example, a highlighting based on pixel classification, or a pseudo-colored superposition of images obtained from different modalities. There is a need to develop image evaluation protocols for medical images that explicitly recognize the possibility that the processed image can be *better*.

In addition to the advantages which the evaluation protocol confers on the originals, physician training also provides a bias for existing techniques. Radiologists are trained in medical school and residency to interpret certain kinds of images, and when asked to look at another type of image (e.g., compressed or highlighted) they may not do as well just because they were not trained on those. Highly compressed images have lower informational content than do originals, and so even a radiologist carefully trained on those could not do as well as a physician looking at original images. But with image enhancement techniques or slightly compressed images, perhaps a radiologist trained on those would do better when reading those than someone trained on originals would do reading originals.

In this series of three chapters, we have presented several different ways of evaluating medical image quality. Simple computable measures have a role in the design algorithms and in the evaluation of quality simply because they are quickly and cheaply obtainable, and tractable in analysis. The actual diagnostic quality is determined by various statistical protocols that enable the evaluation of diagnostic accuracy in the context of specific detection and measurement tasks. The analysis of subjective quality is of interest mostly for the fact that it shows a different trend from actual diagnostic quality, which can reassure physicians that diagnostic utility is retained even when a compressed image is peceptually distinguishable from the original. There is considerable future work to be done both in evaluation studies of image quality for different types of images and diagnostic tasks, and in searching for computable measures of image quality that can accurately predict the outcome of such studies, and perhaps be incorporated into algorithms for designing codes that yield better quality compression.

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] B.J. Betts. *A Statistical Analysis of Digital Mammography*. PhD thesis, Stanford University, Department of Electrical Engineering, 1999.

[2] B.J. Betts, J. Li, A. Aiyer, S.M. Perlmutter, P.C. Cosman, R.M. Gray, R.A. Olshen, D. Ikeda, R. Birdwell, M. Williams, K.O. Perlmutter, S. Horine, C.N. Adams, L. Fajardo, and B.L. Daniel. Image quality in digital mammography. Technical report, Stanford University, November 1998. revision of final report to

the Army Medical Research and Material Command and Compression and Classification of Digital Mammograms for Storage and Transmission, and Computer Aided Screening. Available as a pdf file at http://www-isl.stanford.edu/˜gray/armyfinal.pdf.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984.

[4] B. Efron. Better bootstrap confidence intervals (with discussion). *J. Amer. Stat. Assoc.*, 82(397):171–200, March 1987.

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

[6] T.J. Hastie and D. Pregibon. Generalized linear models. In J.M. Chambers and T.J. Hastie, editors, *Statistical Models in S*, Cole Advanced Books and Software, chapter 6, pages 196–246. Wadsworth & Brooks, Pacific Grove, CA, 1992.

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

[8] I. McNemar. Note on the sampling errors of the differences between correlated proportions of percentages. *Psychometrika*, 12:153–157, 1947.