Measuring Quality in Computer-Processed Radiological Images

R. M. Gray, Dept. of Elect. Eng.
R. A. Olshen, Dept. of Health Res. and Pol.
D. Ikeda, Dept. of Radiology
Stanford University, Stanford, CA 94305
P.C. Cosman, Dept. of Elect. and Comp. Eng., Univ. of Calif. at San Diego, La Jolla, CA
S.M. Perlmutter, C. Nash, K.O. Perlmutter *, Department of Electrical Engineering, Stanford University, Stanford, CA 94305

Abstract

As digital imaging and digital image processing grow in their importance to medical and scientific applications, issues of image quality and utility in sensitive applications also gain importance. Analog images must be digitized if they are to be transmitted or stored in digital media or if they are to be subjected to digital image processing for assisting screening, diagnosis, and research. Digitization and digital signal processing such as lossy data compression, enhancement, or segmentation all entail changing an image from its original form. A key question for all concerned with radiology is to determine when such changes indeed improve the quality or utility of the resulting images. We discuss here some of the issues that arise in demonstrating that one image is as good as or better than another in a specific medical application.

1 Introduction

Radiology is becoming increasingly digital and hence has the capability of using digital communication links and storage facilities and digital image processing as potential aids to screening, diagnosis, and research. The vast majority of medical images in hospitals are X-rays, which are still acquired as analog images and must be digitized if they are to take advantage of the digital communication, storage, and processing systems. Digitization causes a loss of information, which might diminish the utility of an image. All digital images, whether acquired digitally or digitized, may be subject to changes in an attempt to speed communication, improve storage efficiency, or be rendered in a form useful for a radiologist viewer. All such changes can conceivably help or hurt in applications. Traditional engineering measures of image quality such as signal-to-noise ratios can be inadequate as a predictor of image quality and do not even make sense in some applications.

A now traditional approach to establishing quality and utility in specific applications is to simulate the application in a carefully designed experiment, gather necessary data in a way that interferes with the simulation as little as possible, and analyze the resulting data to prove or disprove a specific hypothesis, such as "image type A is at least as effective as image type B" in a specific diagnostic application. While the goal is traditional, the implementation is not.

2 Basic Principles

The following general principles for protocol design have evolved from earlier work on quality and utility evaluation [1, 2, 3, 4, 5, 6].

- The protocol should simulate ordinary clinical practice as closely as possible. Participating radiologists (judges, observers) should perform in a manner that mimics their ordinary practice. The studies should require little or no special training of their clinical participants.
- The clinical studies should include examples of images containing the full range of possible findings, all but extremely rare conditions.
- The findings should be reportable using the American College of Radiology (ACR) Standardized Lexicon.
- Statistical analyses of the trial outcomes should be based on assumptions as to the outcomes and sources of error that are faithful to the clinical scenario and tasks.
- "Gold standards" for evaluation of equivalence or superiority of algorithms must be clearly defined and consistent with experimental hypotheses.

*This work was supported by the Army Medical Research and Materiel Command under Grant DAMD17-94-J-4354 and by Kodak, Inc.

1058-6393/96 $5.00 © 1996 IEEE
Proceedings of ASILOMAR-29

• Careful experimental design should eliminate or minimize any sources of bias in the data that are due to differences between the experimental situation and ordinary clinical practice, e.g., learning effects that might accrue if a similar image is seen using separate imaging modalities.
• The number of studies should be sufficient to ensure satisfactory size and power for the principal statistical tests of interest.

We have argued in the cited references that traditional receiver operating characteristic (ROC) analyses violate several of the basic goals because of the requirement for confidence levels, the statistical assumptions of Gaussian or Poisson behavior, the difficulty dealing with nonbinary tasks, and the lack of care in distinguishing among possible notions of "ground truth" or "gold standard" in clinical experiments. We focus on three definitions of diagnostic truth as a basis of comparison for the diagnoses on all reproductions of that image. These are: **Personal:** Each judge's readings on an original analog image are used as the gold standard for the readings of that same judge on the digitized version of that same image, **Independent:** formed by the agreement of the members of an independent expert panel, and **Separate:** produced by the results of further imaging studies (including ultrasound, spot and magnification mammogram studies), surgical biopsy, and autopsy. The first two standards are conservative in that they are biased in favor of the modality used to establish the gold standards. Whenever a separate gold standard is available, it provides a more fair gold standard against which both old (analog) and new (digital) images can be compared. When histologic data are available, they can be used to establish a separate gold standard against which results based on both analog and digital images can be compared.

3 Protocols

We have proposed a protocol for comparing full screen digital mammography (FDDM) with traditional analog film/screen mammagraphy (F/S) and for comparing FDDM with lossy compressed versions. This protocol is an expanded version of that reported in [6] and concentrates on a screening application with diagnostic aspects. We have proposed studies using 200 normal and 200 abnormal patients and nine radiologist judges. Two views will be provided of each breast (CC and MLO), so four views will be seen simultaneously for each patient. Each of the judges will view all the images in an appropriately randomized order over the course of several sessions. Two sessions should be held every other week, with a week off in between. A clear overlay should be provided for the judge to mark on the image without leaving a visible trace. For each image, the judge either should indicate that the image is normal, or, if something is detected, should have an assistant fill out an Observer Form in Figure 1 using the American College of Radiology (ACR) Standardized Lexicon by circling the appropriate answers or filling in blanks as directed. The form is intended to capture the essential information of screening with supporting detail regarding detection and assessment in a form useful for statistical analysis. This is done using the ACR lexicon so as to approximate ordinary procedures as much as possible and obviate special training. The use of the form is described in the Instructions in Figure 2. The judges should be asked to use a grease pencil to circle the detected item. The judges should be allowed to use a magnifying glass to examine the films.

4 Statistical Analysis

Detection accuracy: Once a gold standard is established, a value can be assigned to the sensitivity, the probability that something is detected given that it is present in the gold standard. Predictive value positive (PVP, also called PPV), the chance an abnormality is actually present given that it is marked, fills the role of quantifying false positive reporting. Sensitivity and PVP should be measured separately for each specific lesion type. They should also be measured for the collection of all anomalies. For this case specificity also makes sense. Mean values for both quantities for both analog and digital images will be determined together with the two-sided 95% confidence regions for the difference. Because such data are neither Gaussian nor binary, some care is required in summarizing them and forming confidence intervals for their "true values." We strongly recommend the use of the BC_a bootstrap technique to compute these intervals [7, 8, 5]. It should be noted that the 95% BC_a confidence intervals for a difference in our basic parameters cover 0 if, and only if, a companion test with α-level 0.05 cannot reject the null hypothesis of "no difference." Differences in sensitivity or PVP between analog and digital images should be analyzed using the permutation distribution of the Behrens-Fisher (Welch) statistic. These comparisons should be conducted for both personal and independent gold standards to demonstrate both consistency and accuracy. Sensitivity and PVP for the masses, calcifications, and other abnormalities can be evaluated both separately and combined.

Management: 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. Statistical analysis should quantify the degree, if any, to which any such differences exist. As for detection, counts can be used to estimate sensitivity, PVP, and specificity with respect to the personal and independent gold standards. Standard statistical methods (including simple χ^2 tests) can be used to quantify any significant differences between the management judgements of each type and as a whole.

An ROC-style curve can be produced by plotting the (sensitivity, specificity) pairs for the management decision for the levels of suspicion. Sample reuse methods (rather than common Gaussian assumptions) can be applied to provide confidence regions around the sample points [9].

Statistical Power: There is little experimental data upon which to base precise computations of size and power in the present mammographic context. Hence we can provide only coarse approximations. It should be emphasized that "power" alone is not the issue. It makes sense only in the context of a specific size, test statistic, null hypothesis, and alternative. Once some preliminary data are available, the power and size can be computed for each test statistic described above to test the hypothesis that digital mammography 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 (sensitivity, specificity, or predictive value positive), the digitally acquired mammograms are worse than analog. The "alternative" is that they are better. In accordance with standard practice, we take our tests to have size .05.

Approximate computations of power devolve from the distributions of off-diagnoal elements in a table listing counts of "right" and "wrong" calls (with respect to any of the gold standards) for the two image modalities. Approximate analysis suggests that for a single judge, for a test of size .05 (5%), the power is approximately .76 for detecting the difference by our test based on the (conditional) binomial computation for our 400 overall subjects, of which, 200 are normal. Changing the parameters a bit does not alter the basic conclusion that we have reasonable power for detecting differences in sensitivity. If the data from four judges can be combined, then power increases to .999+ (for our size .05 test). If six judges could be combined, then we could lower size to nearly 0 and have power nearly 1.

Specificity is a more delicate issue and here our approach is rather different from the approach that we have taken regarding sensitivity. Sensitivity is a "breast by breast" issue in that one commits an egregious mistake by missing disease in a single breast. Each woman was assumed in the computations thus far to contribute two breasts to the computation of sensitivity except regarding diagnoses in which asymmetry is the defining parameter. With specificity, the egregious mistake is to take a woman to biopsy of either breast when she does not require it. Here, the units for computation are individuals, and the effective sample sizes therefore are much smaller than before. The values of the parameters are quite different as well. For an individual judge, the power of a test of our null hypothesis for which the size is .05 is only .27. If, however, we can combine the results of four judges, then the power of the size .05 test rises to .71, while if we can combine the results of all six judges, then the power increases to .83.

References

[1] R. A. Olshen, P. C. Cosman, C. Tseng, C. Davidson, L. Moses, R. M. Gray, and C. Bergin, "Evaluating compressed medical images," in *Proceedings COMCON III*, (Victoria, B.C., Canada), pp. 830–840, Oct. 1991.

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

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

[4] S. Perlmutter, C. Tseng, P. Cosman, K. . Li, R. Olshen, and R. Gray, "Measurement accuracy as a measure of image quality in compressed mr chest scans," in *Proceedings of the IEEE 1994 International Symposium on Image Processing*, vol. 1, (San Antonio, Texas), pp. 861 – 865, October 1994.

[5] P. Cosman, R. Gray, and R. Olshen, "Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy," *Proceedings of the IEEE*, June 1994. to appear.

[6] R. M. Gray, R. A. Olshen, D. Ikeda, P. Cosman, S. Perlmutter, C. Nash, and K. Perlmutter, "Evaluating quality and utility in digital mammography," in *Proceedings of the IEEE 1995 International Conference on Image Processing*, (Washington, D.C.), October 1995.

[7] B. Efron, "Better bootstrap confidence intervals and bootstrap approximations," *J. Amer. Stat. Assoc.*, vol. 82, pp. 171–185, 1987.

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

[9] A. Garber, R. Olshen, H. Zhang, and E. Venkatraman, "Predicting high-risk cholesterol levels," *International Statistical Review*, vol. 62, no. 2, pp. 203–228, 1994.

ID number _____ Session number _____ Case number _____

Reader initials: _____
Mammograms were of (Left, Right, Both) breast(s).

Subjective rating for diagnostic quality (sharpness, contrast)?

(bad) 1 – 5 (good):	Left CC	Left MLO	Right CC	Right MLO

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

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

Finding side: Neither, Left, Right, Both

Findings (detection):

Individual finding side: Left, Right Finding # _____ of _____

Projection in which finding is seen: CC MLO **CC and MLO**

Location:

1) UOQ	5) 12:00	9) retroareolar	13) inner
2) UIQ	6) 3:00	10) central	14) upper
3) LOQ	7) 6:00	11) axillary tail	15) lower
4) LIQ	8) 9:00	12) outer	16) whole breast

Finding type: (possible, definite)

1) mass	5) architectural distortion	9) breast edema
2) calcifications	6) solitary dilated duct	10) other
3) mass containing calcifications	7) asymmetric breast tissue	___
4) mass with surrounding calcs	8) focal asymmetric density	___

CC View	MLO View
Size: _____ cm long axis by _____ cm short axis	Size: _____ cm long axis by _____ cm short axis
Distance from the nipple: _____ cm	Distance from the nipple: _____ cm

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

1) breast edema	(p , d)	5) lymphadenopathy	(p , d)	**9) multiple similar masses**	**(p , d)**
2) skin retraction	(p , d)	6) trabecular thickening	(p , d)	**10) dilated veins**	**(p , d)**
3) nipple retraction	(p , d)	7) architectual distortion	(p , d)	**11) asymmetric density**	**(p , d)**
4) skin thickening	(p , d)	8) calcs associated with mass	(p , d)		

Assessment: The finding is

(A) indeterminate, additional assessment needed
What? 1) spot mag 2) extra views 3) U/S 4) old films
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

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

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

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

(5) radiographic malignancy, biopsy

Comments: _____

Figure 1: Observer Form

Instructions to mammogram readers

You have been invited to participate in a reading of mammograms to detect breast abnormalities as seen on analog and digital studies. The study has been designed to simulate the clinical scenario as closely as possible. The films have been hung so that you will not be able to identify the patient names, and separate study numbers have been assigned to each patient for purposes of the study. A clear overlay has been taped to each film, but this should not interfere with your reading of the image. You may use a magnifying glass and you may use a bright light as you would ordinarily in clinical practice. The reading of the films is not timed.

An assistant will be assigned to you to prompt you for specific answers to questions on breast density, location, and suspicion of breast findings as stated on a questionnaire. You will also be asked to circle the abnormalities on the clear overlays with a grease or wax pencil and number them. You will also be asked to mark the location of the nipple on each film. Please be as specific as possible and follow these guidelines:

1. Please rate each mammogram for its sharpness and contrast as based on the technique of the year it was obtained. Rate each individual view for quality, e.g., "The right CC is good (5), and all the others are pretty good (4)." Note motion unsharpness in the comments.
2. Rate the right and left breast densities separately, for example the left breast could be rated as 1 and the right breast could be rated as 2.

Abnormalities:

1. Tell the assistant how many abnormalities are present in each breast, then describe each abnormality individually, e.g., "There are two lesions in the left breast. Lesion 1 of 2 is ...." The student will fill out extra forms when there are lesions in both breasts, or multiple lesions in one breast. The assistant will not re-fill out the ratings for diagnostic quality or breast density for each abnormality.
2. Circle all abnormalities, whether benign or malignant (i.e. circle fibroadenomas, fat necrosis, benign appearing clustered calcifications as well as malignant appearing calcifications). Please also note the location of the nipple by a grease or wax pencil mark on the clear overlay.
3. For each abnormality, rate it as a definite or possible abnormality. Possible abnormalities are those in which you are not sure that a lesion exists, for example, possible architectural distortion for which you would get additional views to confirm or exclude a lesion. Definite abnormalities are ones that are conclusively present, such as a mass or focal asymmetric density.
4. If you can only see an abnormality on one view, please circle it only on that view.
5. Circle spiculated masses such that you include the body of the mass but not its tiny extensions. For architectural distortion that may not have a central mass, include the spiculations.
6. Note and encircle architectural distortion, even when you think it is due to post-biopsy change and include the spiculations in your outline.
7. If you are unsure whether an apparent lesion exists, encircle it and judge the assessment as 'A' (assessment incomplete), and note your uncertainty by circling the *Y*. Here extra views are needed to confirm or exclude the presence of the abnormality.
8. If you are sure an apparent lesion exists and is a true mass, calcification, calcification cluster, or other finding, but the assmenent is 'A' because ultrasound or extra views are needed to evaluate mass borders or calcifications shapes, or to determine if the finding is a cyst, please mark down your BEST GUESS as to whether the lesion is benign or malignant using the ACR lexicon codes.
9. If the lesion has a differential, such as post-biopsy change vs. cancer, or cyst, fibroadenoma or well-circumscribed cancer, and you would like to note it, please do so in the comments section.

Thank you for your participation in this study. If you have questions or comments, please direct them to Debra M. Ikeda, M.D. at (415) 723-7672.

Figure 2: Observer Form Instructions