EVALUATING QUALITY AND UTILITY IN DIGITAL MAMMOGRAPHY

R. M. Gray, R. A. Olshen, D. Ikeda, P.C. Cosman, S. Perlmutter, C. Nash, and K. Perlmutter

Departments of Electrical Engineering, Health Research and Policy, and Radiology
Stanford, CA 94305

ABSTRACT

Image quality and utility become crucial issues for engineers, scientists, patients, regulators, administrators, insurance companies, and lawyers whenever there are changes in the technology by which medical images are produced. Examples of such changes include analog-to-digital conversion, lossy compression for efficient transmission and storage, image enhancement, and computer-aided methodology for diagnosis that affects the appearances of images. This paper is a summary of some principles for designing protocols for clinical experiments to quantify the relative qualities and utilities of different images, here analog, digital, and lossy compressed digital mammograms. The talk will supplement this paper with a status report on the specific experiment described which is scheduled to be conducted during summer 1995.

1. INTRODUCTION

Digital mammography holds promise for earlier detection of breast cancer than typically is possible at present because it will improve the efficiency and accuracy of screening and provide rapid and reliable access to mammograms by radiologists. Digital images can be stored in digital media and transmitted over digital communication networks such as the expanding National Information Infrastructure. They also permit the application of methods for digital image processing so that regions of clinical interest are enhanced or highlighted, and the implementation of modern techniques for classification and pattern recognition, thereby giving automatic second opinions for screening and diagnosis. Digital image processing has as its goals improvements in the appearance and usefulness of images. A critical issue is how such quality or utility can be quantified in a manner convincing to the medical community and to the Food and Drug Administration (FDA). Such validation is necessary for demonstrating that technologies alleged to be advances in fact provide improvements over well established methods.

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

2. GOALS

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

- The protocol should simulate ordinary clinical practice as closely as possible. Participating radiologists should perform in a manner that mimics their ordinary practice. The experiments 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 anomalies, all but extremely rare conditions.
- The findings should permit summary within the American College of Radiology (ACR) Standardized Lexicon.
- Statistical analyses of the outcomes should be based on assumptions as to the outcomes and sources of error that are faithful to the clinical scenario and tasks.
- The numbers of patients and images should be sufficient to ensure satisfactory size and power for the principal statistical tests of interest.
- "Gold standards" for evaluation of equivalence or superiority of algorithms must be defined clearly and be consistent with experimental hypotheses.
- 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.

3. ROC ANALYSIS

ROC analysis is the dominant technique for evaluating the suitability of radiologic techniques for real applications [6, 7, 8, 9]. Its origins are in the theory of signal detection: a filtered version of signal plus Gaussian noise is sampled and compared to a threshold. If

the threshold is exceeded, then the signal is said to be there. As the threshold varies, the probability of erroneously declaring a signal absent and the probability of erroneously declaring a signal there when it is not vary too, and in opposite directions. The plotted curve is a summary of the tradeoff in these two quantities; more precisely, it is a plot of *true positive rate* or *sensitivity* against *false positive rate*, the complement of *specificity*. Summary statistics, such as the area under the curve, can be used to summarize overall quality. Aspects of analyses are facilitated when data are Gaussian. In typical implementations, radiologists are asked to assign integer confidence ratings to their diagnoses, and thresholds in these ratings are used in computing the curves. This approach generally differs from clinical practice and requires special training. Further, as image data are nonGaussian, methods that rely on Gaussian assumptions are suspect. Modern computer-intensive statistical sample reuse techniques can help get around the failures of Gaussian assumptions, but in fact difficulties with ROC in this specific context are more fundamental. For clinical studies that involve other than binary tasks, specificity does not make sense because it has no natural or sensible denominator, as it is not possible to say how many abnormalities are absent. This can be done for a truly binary diagnostic task, for if the image is not normal then exactly one abnormality is absent. Previous studies were able to make use of ROC analyses by focusing on detection tasks that were truly binary or could be rendered binary. Extensions to ROC to permit consideration of multiple abnormalities have been developed [10], but these still require the use of confidence ratings as well as Gaussian or Poisson assumptions on the data, and we believe that alternative methods are preferable.

4. TRUTH

There are many ways to define "ground truth" or "gold standard" in clinical experiments. We focus on three as providing the diagnostic truth of each original image and as a basis of comparison for the diagnoses on all versions 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, or autopsy.

The first two gold standards are usually established using the analog original films. As a result, they are extremely biased in favor of the established modality, i.e., the original analog film. Thus statistical analysis arguing that a new modality is equal to or better than the established modality will be conservative since the original modality is used to establish "ground truth." The personal gold standard is in fact "hopelessly" biased in favor of the analog films, since each judge's determination for the analog image will be defined to be correct in the comparison against that judge's reading of the digital version. Therefore, it is impossible for the personal gold standard to be used to show that digital images are *better* than analog ones. Any noise in the diagnostic decision will be counted against the modality that does not define truth. The personal gold standard is often useful however, for giving some indication of the diagnostic consistency of an individual judge. The independent gold standard is also biased in favor of the analog images, but not "hopelessly" so, as it is at least possible for the readings of an individual judge on either the digital or analog images to differ from the analog gold standard provided by the independent panel. Whenever a separate gold standard is available, it provides a more fair benchmark 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.

5. AN EXAMPLE: MAMMOGRAPHIC SCREENING/DIAGNOSIS

Hypothesis: Digitized mammograms (12 bits per pixel, 50 micron spot size) and lossy compressed digitized mammograms are equivalent to traditional film/screen mammography for the indication of screening asymptomatic women provided that the bit rate is sufficient (the particular value to be estimated conservatively).

Questions to be answered by the study:

1. Do digital mammograms and lossy compressed digital mammograms provide equal or superior values for important parameters in comparison with film screen mammography? Particular parameters of interest are sensitivity, predicted value positive (PVP or PPV), and, when it can be defined, specificity.
2. Are there any significant statistical differences between the assessment and resulting management recommendations made in clinical trials based on analog, digital, and lossy compressed digital mammograms?

Methods: Images will be viewed on hardcopy film on a lightbox in a manner closely simulating ordinary screening practice. Two views will be provided of each breast (CC and MLO), so four views will be seen si-

multaneously for each patient. Each of four judges will view all the images in an appropriately randomized order over the course of seven sessions. Two sessions will be held every other week, with a week off in between. An acetate overlay will be provided for the judge to mark on the image without leaving visible trace. For each image, either the judge will indicate that the image is normal, or, if something is detected, will fill out the form in Figure 1 using the American

Reader initials: _____
Mammograms were of (**Left, Right, Both**) breast(s).
Subjective rating for diagnostic utility (sharpness, contrast)?
(bad) 1 2 3 4 5 (good)

Findings (detection):
Breast Density:

1. almost entirely fat
2. scattered fibroglandular densities that could obscure a lesion
3. heterogeneously dense (possibly lowering mammographic sensitivity)
4. extremely dense (lowering the mammographic sensitivity)

Implants present? **No, Prepectoral, Retropectoral**
Abnormality side? **Neither, Left, Right, Both** (Use separate sheet for women with abnormalities in both breasts.)
Projection in which abnormality is seen: **craniocaudal (CC), mediolateral oblique (MLO), CC and MLO**
Location: **1) upper outer quadrant (UOQ) , 2) upper inner quadrant (UIQ), 3) lower outer quadrant (LOQ), 4) lower inner quadrant (LIQ), 5) 12:00, 6) 3:00 , 7) 6:00 , 8) 9:00, 9) retroareolar, 10) central, 11) axillary tail**
Lesion type: **1) mass, 2) calcifications, 3) both mass and calcifications, 4) architectural distortion, 5) solitary dilated duct, 6) asymmetric breast tissue, 7) focal asymmetric density**
Size: _____ cm long axis by _____ cm short axis seen on (CC,MLO) view.
Distance from the nipple: _____ cm from the nipple on CC view.
Distance from the nipple: _____ cm from the nipple on MLO view.
Assessment/Management:
The (mass, calcifications, abnormality) is/are

(A) indeterminate, additional assessment needed

(1) negative

(2) benign (also negative but with benign findings)

(3) probably benign finding requiring 6-month followup

(4) suspicion of malignancy (low)

(4) suspicion of malignancy (moderate)

(4) suspicion of malignancy (high)

(5) radiographic malignancy

Management: (A,4,5) further study, possible biopsy. (1,2,3) routine follow-up.

Figure 1: Observer Form

College of Radiology (ACR) Standardized Lexicon by circling the appropriate answers or filling in blanks. The form is intended to capture the essential information of screening in a manner that facilitates statistical analysis. The form will be completed for each detected item, so there may be several filled out for one patient. Ellipses drawn around clusters should include all microcalcifications seen, as if making a recommendation for surgery. Masses should be outlined carefully to 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 will be allowed to use a magnifying glass to examine the films.

Although the judging form is not a standard form, 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.

Clinical history of the patient will not be provided. Judges will not be supplied with prior films, and will not know the patient's age. In this way each image will be judged on its own merits.

The initial question requesting a rating of diagnostic utility on a scale of 1-5 is not itself used to quantify actual diagnostic utility. Rather, it is intended for a separate evaluation of the general subjective opinion of the radiologists of the images. The degree of suspicion registered in the Management portion also provides a kind of subjective rating. It is desirable that obviously malignant lesions in a gold standard should also be obviously malignant in the alternative method.

6. 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. Sensitivity makes sense for non-binary detection tasks, and is a crucial statistic that quantifies results. *Predictive value positive* (PVP, also called PPV), the chance an abnormality is actually present given that it is marked, fills the role of specificity in penalizing false positive reporting. Sensitivity and PVP can be measured separately for each specific lesion type, and for the collection of all anomalies, i.e., for the identification of any of the listed lesions as opposed to none. For this case specificity also makes sense as a statistic.

Mean values for both quantities for both analog and digital images will be determined together with the two-sided 95% confidence regions. Because such data are neither Gaussian nor binary, some care is required in summarizing them and forming confidence intervals

for their "true values." Computer-intensive schemes such as permutation statistics and bootstrapping [11] can be adapted to form valid confidence intervals for these two fundamental parameters [5].

Relative to the independent gold standard, sensitivity and PVP for the findings of the judging radiologists will be determined by whether or not their outlined sites largely contain the smaller circles of the independent panel (taking into account possible positioning differences on the digital mammograms). Differences in sensitivity or PVP between analog and digitized images will be analyzed using the permutation distribution of the Behrens-Fisher (Welch) t-statistic. The test is a variation of the two-sample t-test that takes account of differences in sample variances. As we implement the test with its permutation distribution, the test is exact in a certain sense, and does not rely on Gaussian assumptions that would patently be false for this data set. These comparisons will 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.

Both of the subjective ratings in the questionnaire provide a score with which ROC curves can be found. The first rating of purely subjective quality clearly separates the overall rating from the specifics of the detection and management, whereas the second rating assessing suspicion does not. We believe that other statistics provide a better indication of the questions posed by the study than do such ROC curves, i.e., are important parameters that summarize the performance of radiologists using digital or lossy compressed digital equal to or better than those that summarize traditional film/screen mammography? Nonetheless, ROC curves can be produced from the data acquired.

Management: Management is a key issue in digital mammography. Digital artifacts *might* lead to an increase in false positives and hence in unnecessary biopsies. Statistical analysis should quantify the degree, if any, to which such differences exist. One way to analyze the management portion of the task is to record the management decisions of (ordinary followup, further study [spot mammo, magnification mammo, other imaging]) for the two modalities in a 2-D array of all possible pairs of the two essential decisions. The 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 judgments of each type and as a whole. A McNemar test then can be applied to test for significant differences in management decisions [2].

Statistical Power: As data are acquired we will use sample-reuse methods, when necessary, to estimate the size and power for various of our test statistics. We conclude with brief summaries of some aspects of power for McNemar tests in which sensitivity and specificity are compared for two methods for management, and where the information comes from judgments of four radiologists whose data can be combined. The example applies traditional assumptions to a population size suggested for an FDA "straw man" protocol for comparing analog and digital mammograms: 400 patients with 200 normal studies, 110 mammographically detected breast cancers, 75 benign findings, and 15 breast edemas. When the two methods are being compared for their sensitivities, both are assumed to have sensitivity 80% under the null hypothesis, and one has sensitivity 85% under the alternative. If the methods are simultaneously wrong only 5% of the time, then the power of the size .05 test of equality of sensitivities for the combined data is .99+. For technical reasons, comparing specificities is more complicated. Here, suppose that what were 80% and 5% are now 50% and 25%. Then the power is .71. If our four judges become six, then .71 becomes .83.

7. 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*, pp. 830–840, Victoria, B.C., Canada, 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, no. 4, 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 Conference on Image Processing*, Vol. 1, pp. 861—8, Austin, Texas, November 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*, vol. 82, pp. 919–932, June 1994.

[6] C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, vol. VIII, no. 4, pp. 282–298, Oct. 1978.

[7] J. A. Swets, "ROC analysis applied to the evaluation of medical imaging techniques," *Investigative Radiology*, vol. 14, pp. 109–121, March–April 1979.

[8] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC)curve," *Diagnostic Radiology*, vol. 143, pp. 29–36, 1982.

[9] B. McNeil and J. Hanley, "Statistical approaches to the analysis of receiver operating characteristic (ROC) curves," *Medical Decision Making*, vol. 4, pp. 137–150, 1984.

[10] D. Chakraborty and L. Winter, "Free-response methodology: alternate analysis and a new observer-performance experiment," *Radiology*, vol. 174, no. 3, pp. 873–881, 1990.

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