Evaluating Quality and Utility of Digital Mammograms and Lossy Compressed Digital Mammograms

C. N. Adams, M.S.[a], A. Aiyer, M.S.[a], B.J. Betts, M.S.[a], J. Li, M.S.[a], P. C. Cosman, Ph.D.,[b]
S. M. Perlmutter, Ph.D.,[c] M. Williams, Ph.D.,[d] K. O. Perlmutter, Ph.D.[c], D. Ikeda, M.D.,[e]
L. Fajardo, M.D.[d], R. Birdwell, M.D.[e], B. L. Daniel, M.D.[e], S. Rossiter, M.D.[e],
R. A. Olshen, Ph.D.,[f] and R. M. Gray, Ph.D.[a] *

[a]Information Systems Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA

[b]Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA

[c]Johnson-Grace Company, Inc., 2 Corporate Plaza, Suite 150, Newport Beach, CA

[d]Department of Radiology, University of Virginia, Charlottesville, VA

[e]Department of Radiology, Stanford University, Stanford, CA

[f]Departments of Health Research and Policy, Statistics, and Electrical Engineering, Stanford University, Stanford CA.

Quality and utility are critical whenever medical images are changed by image processing such as analog to digital conversion and lossy compression. These issues are of concern to those who conduct research and development of digital imaging devices and algorithms intended to improve quality or utility of images, and to regulatory agencies that define protocols for validating new devices. This paper is a summary of principles for the design of clinical experiments to quantify quality and utility of medical images along with preliminary results of a pilot experiment. It has evolved as a result of cooperative research among engineers, statisticians, radiologists, and medical physicists. The protocol and experiment provide a context for discussion of a “strawman” experimental design for approval of digital mammography devices promoted by the Center for Devices and Radiological Health of the Federal Drug Administration (CDRH/FDA).

1. Introduction

Radiology is increasingly digital and is correspondingly capable of using digital communication links, storage facilities, and image processing. Digitization of analog images and most image processing algorithms change images and therefore might reduce their utility. A now traditional approach to establishing quality and utility in specific applications is to simulate the

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

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 assess the validity of a specific hypothesis, such as "image type A is not different from image type B" in a specific diagnostic application. For example, in the United States, new devices such as full frame digital mammography (FFDM) systems must receive approval from the CDRH/FDA if they are to be marketed commercially. In September, 1995 the CDRH/FDA called for comments on a *Draft Guidance* describing a proposed protocol for clinical studies that might lead to such approval. The protocol was designed to test the hypothesis that FFDM is at least as effective as traditional film/screen (F/S) mammography in screening applications. We believe that the experimental design and statistical analysis of the "Guidance" could be improved. The proposed requirement that more than 11,000 patients be studied seems vastly excessive for adequately testing sensitivity and specificity of decisions regarding patient management. If adopted, the proposal would place an extreme burden in time and money on companies seeking approval for such devices. This in turn could delay the application of promising new technologies, could have a chilling effect on research, and could discourage companies from remaining in or entering the field. The controversy focuses on fundamental issues of how to quantify quality and utility of medical images that have been altered by digitization, digital acquisition, or image processing.

We describe fundamental principles we have developed in part as a response to our concerns regarding applications to radiology of traditional receiver operating characteristic (ROC) methods. The principles provide a context in which to describe the evolution of the CDRH/FDA "strawman" protocol and of our proposed alternative, aspects of which have been incorporated into the *Draft Guidance*. A new version of the *Guidance* will soon be released which may show further convergence or divergence from this work.

2. Background

The following general principles for protocol design have evolved from earlier work on quality and utility evaluation [1–5]: 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. "Gold standards" for evaluation of equivalence or superiority of algorithms must be clearly defined and 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. 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. The number of studies should be sufficient to ensure satisfactory size and power for the principal statistical tests of interest. 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, it is common to constrain the size of the test to be small, and to attempt to make the power of the test as large as possible.

Receiver Operating Characteristic (ROC) analysis has been the dominant technique for evaluating the suitability of radiologic techniques for real applications [6–10]. We have argued that many ROC analyses violate the stated goals because of the requirement for confidence levels, the statistical assumptions of Gaussian or Poisson behavior, the difficulty of dealing with non-binary tasks, and the lack of care in distinguishing among possible notions of "ground truth." Our research has involved three definitions of diagnostic truth as a basis of comparison for the diagnoses on all versions an image. *Personal:* each judge's readings on an original analog image constitute 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 follow-up, biopsy, or autopsy. The second of these is our principal concern here.

3. Hypotheses

The nominal hypothesis is that digitized mammograms 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. We also consider diagnostic tasks so as to evaluate the possible effects of digitization on diagnostic aspects of the images when they are considered without the context of past history and other modalities. The more fundamental hypothesis is that the nominal hypothesis can be tested by using a protocol consistent with the stated principles and that this can fulfill the underlying goals of the proposed CDRH/FDA protocol for digital vs. analog comparisons with an order of magnitude fewer patient studies. The current pilot study and the preliminary results reported here are consistent with these hypotheses, but the study is too small to be definitive. The requirements for a definitive study are discussed later. The protocol is applicable to any image modality.

4. Methods

Images, Digitization, and Lossy Compression Our primary goal has been the demonstration that FFDM is not less effective than F/S mammography. This pilot study has focused instead on digitized F/S mammograms for digital images because at the time we gathered the data this approach provided the best digital images for purposes of comparison and compression. The studies were digitized at the University of Virginia using a Lumisys Lumiscan 150 at 12 bpp with a spot size of 50 microns. Films were printed using a Kodak 2180 X-ray film printer. The 57 studies included a variety of normal images and images containing benign and malignant objects. For our experiment we selected a popular and efficient algorithm for compression in terms of the tradeoffs among bit rate, distortion, and complexity — an embedded zerotree wavelet coding scheme due to Said and Pearlman [11].

Study Design Our protocol for comparing FFDM with traditional F/S and for comparing FFDM with lossy compressed versions [12,5] concentrates on a screening application with diagnostic aspects. At the time of the Third International Workshop on Digital Mammography, the observer trials for a pilot study involving 57 patient studies and three radiologist judges have been completed at Stanford University and have begun at the University of Virginia with an additional three radiologist judges. The preliminary results reported here are based on the Stanford portion of the experiment with an independent gold standard. This pilot study involves comparisons with lossy compressed digital images as well as the basic analog vs. digital comparison. We have proposed a future definitive study using 200 normal and 200 abnormal patients and nine radiologist judges using the same basic protocol. In our estimation the larger study will suffice to provide good statistical size and power for principal tests of interest, as discussed later.

To simulatate screening mammography, two views are provided of each breast (CC and MLO), so four views are seen simultaneously for each patient. Each judge views all of the images in an appropriately randomized order over the course of several sessions. Two sessions are held every other week, with a week off in between. For each image, the judge either indicates that the image is normal, or, if something is detected, has an assistant fill out an observer form using the American College of Radiology (ACR) Standardized Lexicon by circling the appropriate answers or filling in blanks as directed. The judge uses a grease pencil to circle the detected item on a clear overlay to the film, and is allowed to use a magnifying glass to examine the films. Clinical history of the patient (including age) is not provided, and judges are not supplied with prior films. The form along with instruction forms for readers and assistants, a prompting form for the assistants, a description of the data set, and a sample compressed image may be seen on the Web (`http://www-isl.stanford.edu/∼gray/army.html`).

Statistical Analysis The gold standard was established by E. Sickles, M.D., Professor of Radiology, UC San Francisco, and Chief of Radiology, Mt. Zion Hospital, and D. Ikeda, Professor and Chief, Breast Imaging Section, Dept. 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.

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

Our primary analyses are based upon simple 2×2 tables of the form where the columns correspond to 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, detection accuracy for specific lesions or management decisions. In these preliminary results we focus on the latter. Regardless of statistic, the goal is to quantify the degree, if any, to which differences exist. Tables can be combined when suitable; but we shall argue this is not a good idea for the present results. We begin with a stratifying of the data according to the values of the independent gold standard (four categories) and by radiologists (three). For a pair of technologies, there are thus $12 = 4 \times 3$ tables. Each study contributes a count to the table. The off-diagonal entries are evaluated with a McNemar test, the exact (conditional) version of which involves a binomial computation. Thus, 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, say, N(1,2) given N(1,2)+N(2,1) is binomial with parameters N(1,2)+N(2,1) and 0.5. 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 in the quality of performance with their use. Whether and how to agglomerate the 12 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 in keeping with what we found in an earlier study of the measurement of vessel sizes in chest MR [3,4]. 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, 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.

Several other approaches are planned, including estimating sensitivity, predictive value positive (PVP), and, when appropriate, specificity of detection and management statistics, estimated by counts with bootstrapped confidence regions for each modality [13,14]. This will be done for personal and separate gold standards as well as the independent standard treated here. 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 points on the curve or functionals of the curve such as the area in a wedge [15].

5. Results

We here focus on preliminary results regarding the screening and management of patients. There are four possible outcomes: “RTS”, by which we mean “incidental, negative, or benign with return to screening”; “F/U”, meaning “probably benign but requiring six month follow-up”; “C/B”, or “additional assessment needed”; and “BX”, which is to say “biopsy”. 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 were 13, 1, 18, and 25.

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 for assessing differences between technologies were computed 48 times, 16 per radiologist. For none of these tables was the exact binomial attained significance level (p-value) .05 or less, though for two of these comparisons, the chi-square approximation to the p-value was significant at $p < .05$. However, these were both for a single radiologist, and the signals in the two comparisons are in opposite directions. Thus, in the comparison of analog with uncompressed digital, this one radiologist made five mistakes in digital mode for what was RTS by the gold standard, classifying each of the five as C/B; the five were classified correctly when they were viewed as analog. However, for studies identified as C/B for the gold standard, the radiologist made only one mistake on the uncompressed digital studies, but seven on the analog original. The mistakes balance, and there is no clear pattern regarding analog original and its digitized version. Furthermore, with 48 comparisons and a .05 chance of Type I error, we expect 2.5 tables to be “significant” by chance alone. We obtained zero or two, depending on how one computes attained significance. That is to say, for our study and for the particular set of its data we were able to analyze in time for this report, 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. Since it is really the analog to digital comparison that seems most timely to report, we offer now the tables for each radiologist in which his or her performance is summarized using the classification of our independent gold standard. In all cases, columns are “digital” and rows “analog”. 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.

	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

Radiologist A

	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

Radiologist B

	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

Radiologist C

Radiologist Agreement Tables: Digital versus Analog

The primary conclusion from the initial data and analysis is that variabilities among judges exceed by a considerable amount, in their main effects and interactions, 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 fact suggests variations in statistical analysis that will be explored this summer in the context of additional data from the second (Virginia) half of the experiment.

6. Discussion

Prevalence: The pilot data set of 57 images has two obvious shortcomings: it is too small to have good power for tests of reasonable size for those tests proposed, and the prevalence of abnormalities in this data set does not accurately reflect that of a normal screening population. This violates the literal goals of accurate simulation and representative statistics for a screening application. The first shortcoming can be resolved by a larger study, although it is a serious and controversial issue as to how large the study must be. The second problem, however, is unavoidable with any study of reasonable sample size as prevalence in a screening population can vary widely at different locations. We argue, however, that relevant conclusions can be drawn for the true prevalence based on a carefully constructed study using different proportions. In order to well simulate the proportion of normal images to ones containing pathology that actually would be found in a screening situation, we would require thousands of studies as there are only 6–8 cancers/1000 asymptomatic women screened. In our approach we do not *directly* estimate overall statistics for detection (sensitivity, PVP) and management (sensitivity, specificity). This would result in little power for some of the statistics without unreasonably large patient numbers or unreasonably large size to the tests. It would also involve incorporating somewhat arbitrarily abnormality prevalence values reflecting the "general population." A purely prospective screening study using commonly assumed prevalence values can require more than 11,000 patients, as reported by NCI statistician Dr. L.G. Kessler at a 6 March 1995 meeting of the Radiological Devices Panel Meeting to consider protocols for demonstrating substantial equivalence of film/screen mammography and FFDM (chaired by Francine Halberg, M.D.). Our "retrospective/prospective" approach, reported as an alternative protocol at the 6 March Panel meeting [5], allows us to compute estimates of our statistics conditional on the presence or absence of abnormalities and to estimate separately size and power for both conditional populations. This then yields by straightforward algebra overall statistics by suitably weighting the conditional statistics to reflect estimated prevalence. The specific numbers of patients needed for good size and power will be estimated in a cumulatively improving manner as the data are gathered and the experiments performed. As considered below, preliminary analysis based on standard approximations suggests that this will be far fewer than many thousands.

One reasonable concern about not attempting to simulate accurately a population prevalence is that radiologists might behave differently if they knew that the prevalences in an experiment were different from that ordinarily encountered in a clinic. This effect could be analyzed in a quantifiable manner by varying the prevalence at different sites in a controlled manner not known to the judges or assistants.

Statistical Size and Power: There is little experimental data upon which to base precise computations of size and power in the present mammographic context. Crude approximations [12,5] can be improved based on prevalence data from the pilot study. A particularly conservative approach [4] is to consider a one-sided test with the "null hypothesis" that, whatever the criterion (sensitivity, specificity, or PVP), the digitally acquired mammograms are worse than analog. The "alternative" is that they are better. Approximate computations of power can be derived from standard approximations based on estimates of the distributions of the off-diagonal elements in the agreement/disagreement table for two image modalities [5,12]. The pilot experiment, however, made two points clear. The digital images often performed better than the analog originals in either sensitivity or specificity, and hence the conservative one-sided test should be replaced by a two-sided test. Second, the estimates for sensitivity in particular were so close that estimated power for distinguishing the two modalities was low even when 6 judges are assumed. The differences among the image modalities are dwarfed by the differences among the radiologists which necessitates great care in producing statistics that can be safely pooled across radiologists. We are in the process of improving our estimates of sensitivity and specificity from the pilot data for use in predicting size and power as a function of the data set size and the number of radiologists, but thus far the implication is that at least six judges and possibly nine will be needed for good size and power for distinguishing specificity, and at least nine or possibly more judges for distinguishing sensitivity on a data set of roughly 400 images.

Acknowledgements

The authors gratefully acknowledge the many hours of help contributed by Edward Sickles, M.D., and by Sarah Horine and Dalia Gomez.

REFERENCES

1. 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.
2. 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.
3. 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 of the IEEE 1994 International Conference on Image Processing*, volume 1, pages 861—8, Austin, Texas, November 1994.
4. P.C. Cosman, R.M. Gray, and R.A. Olshen. Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy. *Proceedings of the IEEE*, 82:919–932, June 1994.
5. R.M. Gray, R.A. Olshen, D. Ikeda, P.C. Cosman, S.M. Perlmutter, C.L. Nash, and K.O. Perlmutter. Full breast digital imaging (FBDI) as a screening device: Summary of clinical protocol. Technical report, March 1995. Report submitted to Panel on Digital Mammography, FDA, 6 March 1995 and described in the transcription "Radiological Devices Panel Meeting" available from CASET Associates, Ltd., 10201 Lee Highway, Suite 160, Fairfax, VA 22030.
6. C. E. Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, VIII(4):282–298, Oct. 1978.
7. J. A. Swets. ROC analysis applied to the evaluation of medical imaging techniques. *Investigative Radiology*, 14:109–121, March–April 1979.
8. J.A. Hanley and B.J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC)curve. *Diagnostic Radiology*, 143:29–36, 1982.
9. B.J. McNeil and J.A. Hanley. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Medical Decision Making*, 4:137–150, 1984.
10. D.P. Chakraborty and L.H.L. Winter. Free-response methodology: alternate analysis and a new observer-performance experiment. *Radiology*, 174(3):873–881, 1990.
11. A. Said and W.A. Pearlman. A new fast and efficient image codec based on set partitioning in hierarchical trees. *IEEE Trans. on Circuits and Systems for Video Technology*, 1996. to appear.
12. 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 Second Conference on Image Processing*, volume II, pages 5–8, Washington, D.C., October 1995.
13. J. Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 20:213–220, 1968.
14. Y.M.M. Bishop, S.E. Feinberg, and P.W. Holland. *Discrete Multivariate Analysis*. MIT Press, Cambridge, Mass, 1975.
15. A. Garber, R.A. Olshen, H. Zhang, and E.S. Venkatraman. Predicting high-risk cholesterol levels. *International Statistical Review*, 62(2):203–228, 1994.