AD_____

Award Number DAMD17-97-1-7193

TITLE: Methods for Evaluating Mammography Imaging Techniques

PRINCIPAL INVESTIGATOR: Carolyn M. Rutter, Ph.D.

CONTRACTING ORGANIZATION: Group Health Cooperative of Puget Sound
Seattle, Washington 98101-1448

REPORT DATE: June 1999

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

20000303 106

DTIC QUALITY INSPECTED 3

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE June 1999	3. REPORT TYPE AND DATES COVERED Annual (19 May 98 - 18 May 99)

4. TITLE AND SUBTITLE
Methods for Evaluating Mammography Imaging Techniques

5. FUNDING NUMBERS
DAMD17-97-1-7193

6. AUTHOR(S)
Carolyn M. Rutter, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Group Health Cooperative of Puget Sound
Seattle, Washington 98101-1448

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This Department of Defense Breast Cancer Research Program Career Development Award is enabling Dr. Rutter to develop biostatistical methods for breast cancer research. Dr. Rutter is focusing on methods for evaluating the accuracy of breast cancer screening. This four year program includes advanced training in the epidemiology of breast cancer, training in clinical detection of breast cancer, development of statistical methodology, and graduate teaching. During this second award year, Dr. Rutter has continued to attend Breast Cancer Surveillance Consortium (BCSC) meetings [1] and has presented her research to the BCSC's Statistical Coordinating Center. Ongoing participation in theBCSC has provided Dr. Rutter with important practical information about radiologists' interpretation of mammograms, and the timing and execution of diagnostic procedures. Dr. Rutter has also continued to expand her knowledge of statistical research methods for diagnostic and screening test assessment through participation in the Diagnostic Methods working group at the University of Washington's Department of Biostatistics. Dr. Rutter's statistical research focuses on receiver operating characteristic (ROC) analysis. During the second funding year, she has published on design of intervention studies aimed at improving mammographer accuracy and had developed methods for comparing mammographers' test performance to their clinical practice performance.

14. SUBJECT TERMS
Breast Cancer
receiver operating characteristic(ROC) curves, correlated observations, measurement error.

15. NUMBER OF PAGES
96

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

__✓__ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Carolyn B 6/15/1999

PI - Signature Date

Table of Contents

Standard Form (SF) 298, Report Documentation Page . . . . . page 2

Foreword . . . . . page 3

I. Introduction . . . . . page 5

II. Year 2 Achievements . . . . . page 6

Technical Objective 1:
Gain additional training in breast cancer epidemiology, detection and treatment. . . page 6
Statistical Research Aims:
Technical Objective 2: Statistical Research, Aim 1:
Develop methods for multiple patient assessments . . . . . page 6
Technical Objective 3: Statistical Research, Aim 2:
Extend exact methods for ordinal regression models . . . . . page 7
Technical Objective 4: Statistical Research, Aim 3:
Develop methods to adjusting for measurement error in disease status . . . . . page 7
Technical Objective 5:
Develop and teach a course in methods for assessing diagnostic tests . . . . . page 7

Progress Toward Other Grant Aims. . . . . page 7

III. Summary and Key Research Accompishments . . . . . page 8

IV. References . . . . . page 9

V. Appendices . . . . . page 11
A. Statement of Work. . . . . page 12
B. Bootstrap Estimation of Diagnostic Accuracy Using Patient-Clustered Data. . . . . page 13
C. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. . . . . page 33
D. Assessing Mammographers Accuracy: A comparison of clinical and test performance . . . . . page 63
E. Design of a Study to Improve Accuracy in Reading Mammograms . . . . . page 84

I. Introduction

The purpose of this Department of Defense Breast Cancer Research Program Career Development Award is enabling Dr. Rutter to develop biostatistical methods for breast cancer research. Dr. Rutter's focus is on evaluating the accuracy of breast cancer screening. This four year program includes advanced training in the epidemiology of breast cancer, training in clinical detection of breast cancer, development of statistical methodology, and graduate teaching. A basic knowledge of the epidemiology, disease process and detection of breast cancer will guide the development of statistical methods designed to address analysis problems encountered when evaluating mammography. Proposed statistical research focuses on receiver operating characteristic (ROC) analysis. ROC analysis provides accuracy measures for ordinal tests and is a more general analysis strategy than other methods devised for dichotomous test outcomes. Therefore, the proposed research will have implications for both ordinal scale and dichotomous test analyses. Additional research will explore accuracy measures specific to dichotomous test outcomes, including sensitivity, specificity, and positive and negative predictive values.

II. Year 2 Achievements

Technical Objective 1: Gain additional training in breast cancer epidemiology, detection and treatment
During this second award year, Dr. Rutter has continued to attend Breast Cancer Surveillance Consortium (BCSC) meetings [1], and has presented her research to the BCSC's Statistical Coordinating Center. The BCSC is a multi-site NCI-funded study evaluating the performance of mammography in a community setting. Ongoing participation in BCSC meetings has provided Dr. Rutter with important practical information about radiologists' interpretation of mammograms, and the timing and execution of diagnostic procedures.

Statistical Research Aims
Dr. Rutter has continued to expand her knowledge of statistical methods for diagnostic and screening test assessment. In particular, her thoughts have been clarified through participation in the Diagnostic Methods working group at the University of Washington's Department of Biostatistics.

Technical Objective 2: Statistical Research, Aim 1: Develop methods for multiple patient assessments.
Dr. Rutter has continued to work toward publishing her article describing bootstrap approaches for multi-site, multi-reader diagnostic test data. A brief version of this article was rejected by *Biometrics*. A more complete version, which includes percentile intervals and comparisons to an analytic estimator, has been submitted to *Academic Radiology* (see Appendix B). The analysis used for this bootstrap paper can be conducted using a relatively simple SAS macro. During her third funding year, Dr. Rutter will generalize this macro and examine ways to make it more broadly available, for example via the SAS users group webpage.

Dr. Rutter has no plans for further development of this research pathway. This reflects her increased knowledge of data collection and use by radiologists, and also reflects recent developments in statistical methodology. As described in her year one progress report, generalized estimating approaches for diagnostic data have been fully developed [2,3]. These models can accommodate correlated rating data, and estimation of models can be carried out using standard statistical software packages. In addition, the limited robustness of ROC curve analysis [4] limits the usefulness of robust covariance adjustments.

More importantly, multi-site data are not likely to be used when assessing screening mammography. In the screening setting, laterality is important but quadrant location within the breast is less important because women go on to further diagnostic assessment. Because screening assessments affect the entire woman, rather than the breast, statistics that deal with data at the woman-level are most informative. The best approach to these data combines breast-level ratings in conjunction with disease state. When a woman has bilateral disease, the lowest (least likelihood of disease) rating given to her two breasts is used for analyses, capturing potential undercalling of disease. When a woman has unilateral disease, the rating given to her diseased breast is used for analyses. Although this ignores overcalling in the non-diseased breast, it incorporates critical disease detection information. When a woman does not have disease, the maximum (most likelihood of disease) rating is used, capturing overcalling of disease.

In the diagnostic setting, both laterality and quadrant are important. These data can be analyzed using GEE methods [2,3]. Unfortunately, rating data are generally not collected at the quadrant level. Analysis at the quadrant level can also be limited by the accuracy of localization by the gold standard (e.g., pathology reports or cancer registry outcomes).

Technical Objective 3: Statistical Research, Aim 2: Extend exact methods for ordinal regression models
Development of methods for small samples has been deferred to year 3. Instead, Dr. Rutter has focused on methods for adjusting for measurement error in disease status.

Technical Objective 4: Statistical Research, im 3: Develop methods to adjust for measurement error in disease status
Several authors have explored methods for estimating test accuracy when there are multiple test outcomes with no true gold standard.[5-11] Some articles have described methods that allow estimation of accuracy in the complete absence of gold standard information.[10,11] Over the last year, this topic has been of great interest to the Diagnostic Methods working group. Methods that handle missing disease status rely on latent variable approaches when the 'definitive' diagnosis is uncertain. These solutions are not satisfying because they hinge on a latent, unobserved, disease state. In this case, the referent populations are unknown, making comparisons across studies, or from studies to clinical practice, extremely difficult. Consider a situation where misclassification can be extreme: screening for alcohol abuse and dependence. Suppose there was a new blood test for alcohol abuse and dependence. To test the accuracy of this new blood test, the natural reference standard is the Diagnosis and Statistical Manual (DSM) definition of alcohol abuse and dependence [12], assessed using a questionnaire. This reference standard is likely to misdiagnose some people. However, one of the key purposes of the DSM diagnostic guidelines is standardization that allows comparability of independent research. New statistical methods allow estimation of sensitivity and specificity relative to an unobserved true state. Unfortunately these approaches leave the research community to ponder the meaning of these sensitivities and specificities. Exactly what the test detects is unclear because it is essentially undefined, rendering these estimates useless.

An alternative approach is to clearly define the reference standard, and to improve reference standards as necessary. In the context of screening mammography, the accepted standard is biopsy and two years of follow-up data. Currently, a cancer is "missed" by screening if it occurs within two years of a disease-negative screening. Incorporation of stage-of-disease information could improve this gold standard. In this case, a cancer is "missed" by screening if it occurs within two years of a disease-negative *and* the woman is node positive. Other information, such as tumor size or grade, could be incorporated into this definition. This approach acknowledges the goal of screening mammography: early detection of disease.

Technical Objective 5: Develop and teach a course in methods for assessing diagnostic tests.
As the 1999 Genentech Distinguished Professor, Dr. Margaret Pepe will teach a special topics course on Medical Diagnostic Testing at the University of Washington's Department of Biostatistics in Spring 2000. Dr. Rutter will work with Dr. Pepe developing course materials and lectures, and will guest lecture.

Progress Toward Other Grant Aims

Dr. Rutter is working toward publishing her article (co-authored with Constantine Gatsonis) describing models for meta-analysis of diagnostic test data (see Appendix C). These methods were developed as a way to appropriately combine sensitivities and specificities from several studies, but can also be applied to sensitivity/specificity results from several sites within a multi-site study or from several mammography centers within a study. This article was recently rejected by *Statistics in Medicine*, and Dr. Rutter is in the process of revising the article for resubmission to an alternate journal.

Dr. Rutter has examined the correlation between test and clinical performance measures of mammographic interpretation (see Appendix D). This article was submitted to *Journal of Clinical Epidemiology* and is in the process of a second review following an encouraging "revise and resubmit". Direct estimation of mammographers' clinical accuracy requires the ability to capture screening assessments and correctly identify which screened women have breast cancer. This clinical information is often unavailable and when it is available its observational nature can cause analytic problems. Problems with clinical data have led some researchers to evaluate mammographers using a single set of films. Research based on these test film sets implicitly assumes a correspondence between mammographers' accuracy in the test setting and their accuracy in a clinical setting. However, there is no evidence supporting this basic assumption. Dr. Rutter used hierarchical models and data from 27 mammographers to directly compare accuracy estimated from clinical practice data to accuracy estimated from a test film set. There was no evidence of correlation between clinical and test accuracy. These findings raise important questions about how mammographer accuracy should be measured. Dr. Rutter has presented these findings to the BCSC, and plans to present to a wider audience at the International Conference on Health Policy Statistics: Methodologic Issues in Health Services and Outcomes Research in December 1999.

During her year two funding period, Dr. Rutter also coauthored a paper describing the design of the mammography rereading studies.[13] (see Appendix E)

In the last two years of grant funding, Dr. Rutter plans to shift her research goals somewhat, to better align them with current needs in mammography research. One new goal is development of methods that handle data collected using the Breast Imaging Reporting and Data System (BI-RADS).[14] This standardized set of mammographic interpretations proscribed by the American College of Radiology lexicon improves data collection by virtue of standardizaton. However, the inclusion of an interpretive code for additional work-up complicates evaluation of mammographic accuracy. The additional work-up category does not fit neatly into an ordinal outcome scale. These cases include a mix of women, for example, it could naturally include both women with suspected cysts (benign disease) and women with suspicious findings that need additional evaluation. Models need to be developed to handle these kinds of data. One possible approach to these data is extension of two-part models employed in econometrics.[15] The first part of the model would estimate the probability of an interpretation based on the current mammogram (i.e., additional workup not requested). The second part of the model would describe ordinal outcomes among observations with an interpretation of the current mammogram. Inference is drawn from the combined results from these two model steps.

III. Summary

Dr. Rutter remains on track with her stated goals, making significant progress towards proposed research goals. At this point in her career development award, she has shifted some of her proposed research in response to her increased knowledge of breast cancer screening and advances in statistical methodology. Statistical approaches for dealing with errors in the gold standard no longer seem feasible. Instead, the focus of research should be on development of more adequate reference standards. New research problems have come to the fore. Dr. Rutter has addressed the validity of assessing mammographers accuracy using test film sets and in the future plans to address analytic problems related to the BI-RADS data collection system.

Key Research Accomplishments, Year 2:

- Published article: Pepe MS, Urban N, Rutter C, Longton G "Design of a study to improve accuracy in reading mammograms," *Journal of Clinical Epidemiology,* 50: 1327-38, 1997.
- Submitted article: Rutter CM. Bootstrap estimation of diagnostic accuracy using patient-clustered data, *Academic Radiology*.
- Submitted article: Rutter CM, Taplin S. Assessing mammographers' accuracy: A comparison of clinical and test performance, *Journal of Clinical Epidemiology*.
- Developed a clear plan for teaching and developing a course on Medical Diagnostic Testing at University of Washington with Dr. Margaret Pepe.
- Ongoing participation in Diagnostic Methods working group at the University of Washington's Department of Biostatistics
- Ongoing participation in Breast Cancer Surveillance Consortium meetings

IV. References

1. Ballard-Barbash R, Taplin SH, Yankaskas BC, Ernster VL, Rosenberg RD, Carney PA, Barlow WE, Geller BM, Kerlikowske K, Edwards BK, Lynch CF, Urban N, Chrvala CA, Key CR, Poplack SP, Worden JK, Kessler LG. "Breast Cancer Surveillance Consortium: A National Mammography Screening and Outcomes Database," *American Journal of Roentgenology,* 169: 1001-1008, 1997.
2. Leisenring W, Pepe MS, Longton G. "A Marginal Regression Modelling Framework for Evaluating Medical Diagnstotic Tests," *Statistics in Medicine,* 16: 1263-1281, 1997.
3. Pepe MS. "Three approaches to regression analysis of receiver operating characteristic curves for continuous test results," *Biometrics,* 54: 124-35, 1998.
4. Walsh SJ. "Limitations to the Robustness of Binormal Roc Curves: Effects of Model Misspecification and Location of Decision Thresholds on Bias, Precision, Size and Power," *Statistics in Medicine,*16: 669-679, 1997.
5. Walter SD, Irwig LM. "Estimation of Test Error Rates, Disease Prevalence and Relative Risk from Misclassified Data: A Review," *Journal of Clinical Epidemiology,* 41: 923-937, 1988.
6. Epstein LD, Muñoz A, He D. "Bayesian Imputation of Predicitve Values When Covariate Information is Available and Gold Standard Diagnosis is Unavailable," *Statistics in Medicine,* 15: 463-476, 1996.
7. Lu Y, Keying Y, Mathur AK, Hui S, Feurst TP, Genant HK. "Comparative Calibration Without a Gold Standard," *Statistics in Medicine,* 16: 1889-1905, 1997.
8. Qu Y, Hadgu A. "A Model for Evaluating Sensitivity and Specificity for Correlated Diagnostic Tests in Efficacy Studies With an Imperfect Reference Test," *Journal of the American Statistical Society,*93: 920-928, 1998.
9. Torrance- Rynard VL, Walter SD. "Effects of Dependent Errors in the Assessment of Diagnostic Test Performance," *Statistics in Medicine,*16: 2157-2175, 1997.
10. Joseph L, Gyorkos TW, Coupal L. "Bayesian Estimation of Disease Prevalence and the Parameters of Diagnostic Tests in the Absence of a Gold Standard," *American Journal of Epidemiology,*141: 263-272, 1995.
11. Joseph L, Gyorkos TW. "Inferences for Likelihood Ratios in the Absence of a 'Gold Standard'," *Medical Decision Making,* 16: 412-417, 1996.
12. Diagnostic and statistical manual of mental disorders : DSM-IV; Washington, DC : American Psychiatric Association, 1994.
13. Pepe MS, Urban N, Rutter C, Longton G "Design of a study to improve accuracy in reading mammograms," *Journal of Clinical Epidemiology,* 50: 1327-38, 1997.
14. Linver MN, Osuch JR, Brenner RJ, Smith RA. "The Mammography Audit: A Primer for the Mammography Quality Standards Act (MQSA)," *American Journal of Roentgenology,165:19-25, 1995*
15. Judge GG, Griffiths WT, Hill RC, Lütkepohl H, Lee T. *The Theory and Practice of Econometrics, Second Edition,* John Wiley and Sons: New York, 1980.

V. Appendices

Appendix A. Statement of Work

Technical Objective 1: Gain additional training in breast cancer epidemiology, detection and treatment.

Task 1: Months 1-4: Review of information on the epidemiology, diagnosis and treatment of breast cancer as suggested by Dr. Margaret Mandelson.

Task 2: Months 1-48: Attend seminars sponsored by the Seattle Breast Cancer Research Program.

Technical Objective 2: Statistical research, aim 1: develop methods for multiple patient assessments.

Task 3: Month 6: Review current research for generalized estimating equation and random effect approaches for nonlinear models.

Task 4: Months -11: Test bootstrap, robust covariance adjustment and generalized estimating equation methods for breast-level analyses using simulation studies.

Task 5: Months 12-21: Develop methods for woman-level analysis, possibly including software development for random effects in generalized ordinal regression models.

Technical Objective 3: Statistical research, aim 2: extend exact methods for ordinal regression models

Task 6: Month 22: Review current research in exact methods.

Task 7: Months 23-34: Extend exact methods and write computational algorithms and programs to compute distributions of sufficient statistics.

Technical Objective 4: Statistical research, aim 3: Develop methods to adjust for measurement error in disease status

Task 8: Month 36: Review current research in errors-in-measurement models.

Task 9: Months 37-48: Develop simple combined corrections for verification and follow-up bias. These methods will be extended to allow adjustments in general ordinal regression models.

Technical Objective 5: Develop and teach a course in methods for assessing diagnostic tests.

Task 10: Months 1-24: Collect relevant references and outlining lectures for the methods course. During this time, specific lectures may be presented in other University of Washington courses.

Task 11: Months 25-36: Offer methods course at University of Washington through the Department of Biostatistics.

Appendix B

Bootstrap Estimation of Diagnostic Accuracy
using Patient-Clustered Data

CM Rutter

submitted to *Academic Radiology*

Bootstrap Estimation of Diagnostic Accuracy using Patient-clustered Data

Carolyn M. Rutter, Ph.D.

Address correspondence to:

Carolyn Rutter

Group Health Cooperative, Center for Health Studies

1730 Minor Avenue, Suite 1600

Seattle, WA 98101

email: rutter.c@ghc.org

phone: 206.287.2190

fax: 206.287.2871

Bootstrap Estimation of Diagnostic Accuracy using Patient-clustered Data

Abstract

Rationale and Objectives: This article describes simple asymptotically consistent bootstrap estimation of test accuracy statistics. Unlike most other methods, the bootstrap approach can account for correlation due to multiple diagnostic modalities, multiple readers, and assessment at multiple body sites. Bootstrap methods are easy to apply, even in complicated settings.

Methods: The performance of bootstrap estimates is evaluated and compared to analytic estimates using a simulation study. Bootstrapping is demonstrated using data from a study comparing two angiography methods.

Results: Analytic and bootstrap estimators had similar coverage rates. Bootstrap estimates were slightly better in some cases, and analytic estimators were slightly better in others. Bootstrap percentile intervals had better coverage than asymptotic normal bootstrap intervals.

Conclusions: Bootstrapping is a useful method for estimating confidence intervals for the area under the ROC curve, sensitivity and specificity when data are correlated.

Keywords: area under the receiver operating characteristic curve (AUC), sensitivity, specificity.

1. Introduction

Diagnostic evaluation often requires simultaneously assessing disease at multiple body sites. Examples of multi-site diagnostic assessments include screening mammography to detect breast cancer, computed tomography of the liver to detect metastatic colorectal cancer(Zerhouni et al, 1996), and magnetic resonance angiography of leg vessels to detect occlusive peripheral vascular disease (Baum et al, 1995). Although the accuracy of these multi-site tests can be estimated using information from a single body site, studies that use all available information have more statistical power. Reducing site level data to patient level data is the simplest approach to multi-site diagnostic assessment. However, composite patient-level measures of true state and test outcome reduce the amount of information about test accuracy contained in multi-site assessments. These composite measures also ignore disease localization, information that can be more important for treatment decisions than global determination of disease presence.

Estimates of diagnostic accuracy that use multi-site data need to account for within patient correlation. Methods for handing multiple assessment of a single site, by different modalities or readers, are well developed. Song (1997) gives an overview of current approaches. These methods require that patients are either diseased or not diseased, and do not allow true state to across the different sites within patients.

When disease state is dichotomous, logistic regression models can be used to estimate the relationship between true state and test outcomes (e.g., Baum et al, 1995). When data are clustered within patients, standard methods can be used to adjust the logistic regression coefficient

covariance matrix for within patient correlation (Lipsitz and Harrington, 1990). The logistic model conditions on test results and estimates their association with disease state. These models do not result in standard accuracy measures, making comparisons to other studies difficult.

Obuchowski described a method for estimating standard errors for the area under the empirical receiver operating characteristic curve based on sums of squares (Obuchowski, 1997). Obuchowski's method allows estimation of the standard error of the AUC for a single test, or the standard error of the difference between AUC statistics for two tests. Obuchowski's approach requires definition and calculation of appropriate sums of squares, and this can become complicated when there are multiple sources of correlation. For example, when patients are evaluated at multiple sites by more than one test with each test independently evaluated by more than one reader.

Pepe recently proposed a general regression methodology that allows multi-site assessments (Pepe, 1998). This regression approach estimates the effects of covariates on the receiver operating characteristic (ROC) curve. The interpretation of the regression model depends on the functional form chosen for the ROC curve. Coefficients estimated from a logistic model are interpretable as the log-odds of correctly classifying a diseased subject for a fixed false positive rate. Pepe suggests using bootstrap resampling to estimate standard errors of regression coefficients when correlated data are included in these models.

This article demonstrates a very simple bootstrap approach for estimating true positive rates, false positive rates, and the area under the ROC curve for multi-site test outcome data. This bootstrap approach is useful for simple comparisons between tests, when there are no covariates. When using regression approaches, bootstrap estimates can provide supplemental descriptive

statistics. The bootstrap approach is easy to use when there are multiple sources of correlation and resulting confidence intervals are asymptotically consistent.

2. Nonparametric Accuracy Statistics: sensitivity, specificity, AUC

The accuracy of imaging tests is based on radiologists' interpretations of disease state. These interpretations are typically measured using a 5-point ordinal scale that ranges from 'definitely not diseased' to 'definitely diseased'. True positive rates, false positive rates, and the area under the receiver operating characteristic curve are the basic statistics used to measure test accuracy. These statistics condition on true disease state, treating disease state as fixed and known and treating test outcomes (ratings) as randomly distributed outcomes. When disease state is known without error, these accuracy statistics are independent of disease prevalence.

When test outcomes are dichotomous, sensitivity and specificity measure test accuracy. Sensitivity is the probability of a positive test outcome (indicating presence of disease) when the target disease is present. Specificity is the probability of a negative test outcome when the disease is absent. When test outcomes are ordinal, sensitivity and specificity can be calculated by dichotomizing outcomes. However, a single sensitivity-specificity pair cannot completely describe the accuracy of an ordinal test because both rates depend on test stringency. Receiver operating characteristic (ROC) curve analysis accounts for the tradeoff in these rates as test stringency varies. Suppose the ordinal outcome of a diagnostic test, t_i, takes values in $\{1, 2, \ldots, K\}$ with increasing values of t_i corresponding to stronger evidence of disease. There are $K + 1$ possible ways to dichotomize the ordinal test, including 'all positive' and 'none positive', and each is associated with a sensitivity-specificity pair. The empirical ROC curve is drawn by plotting

pairs of observed rates, (1-specificity) versus sensitivity, and connecting the $K+1$ consecutive points with straight lines. The empirical ROC curve provides a simple graphical description of test performance.

The overall accuracy of an ordinal test can be summarized by the area under the ROC curve (AUC). The AUC estimates the probability of correctly ranking a randomly selected (diseased,not-disease) pair on the ordinal test scale; It ranges from 0 to 1, with the value 1 corresponding to a perfect diagnostic test. A test that is no better than chance has an AUC equal to one half. The AUC statistic is unbiased and asymptotically normally distributed. The test of H_o : AUC $= 1/2$ based on the asymptotic distribution is equivalent to a Mann-Whitney test (Hanley and McNeil, 1982). The AUC test is essentially a test for differences in the distribution of test outcomes in diseased and not-diseased groups.

3. Bootstrap estimation of sensitivity, specificity, AUC

Sensitivity, specificity, and the area under the receiver operating curve (AUC) are all generalized U-statistic of order 1: Each of these statistic is a sum of functions of statistically independent quantities (Lee, 1990). Because sensitivity, specificity and the AUC are U-statistics, bootstrap resampling provides consistent point and interval estimates (Bickel and Freedman, 1981; Arcones and Giné, 1992).

Let $\mathbf{t}_i = (t_{i1}, t_{i2}, \ldots, t_{im})'$ be the vector of ordinal test outcomes across m sites for the ith subject and let $\mathbf{d}_i = (d_{i1}, d_{i2}, \ldots, d_{im})'$ be the corresponding vector of true states, where $d_{ij} = 1$ if the jth site of the ith patient is diseased and $d_{ij} = 0$ otherwise. Written in U-statistic form,

sensitivity and specificity for the kth cutpoint are:

$$\text{sensitivity}_k = \frac{1}{n_D}\sum_i \phi_k(\mathbf{t}_i, \mathbf{d}_i) \qquad \text{and} \qquad \text{specificity}_k = \frac{1}{n_{\overline{D}}}\sum_i (1 - \phi_k(\mathbf{t}_i, (1 - d_i)))$$

with kernel function $\phi_k(\mathbf{t}_i, \mathbf{d}_i) = \sum_j \delta_k(t_{ij}) d_{ij}$ where $\delta_k(t) = 1$ if $t \geq k$, and $\delta_k(t) = 0$ otherwise. The associated sample sizes are $n_D = \sum_i \sum_j d_{ij}$ and $n_{\overline{D}} = \sum_i \sum_j (1 - d_{ij})$. Here D indicates presence of disease and $\overline{D}$ indicates absence of disease.

The AUC statistic is given by:

$$\text{AUC} = \frac{\sum_{(i,j)\in D} \sum_{(i',j')\in \overline{D}} \psi(t_{ij}, t_{i'j'})}{n_D n_{\overline{D}}}$$

with kernel function

$$\psi(t_{ij}, t_{i'j'}) = \begin{cases} 1 & \text{if} \quad t_{ij} > t_{i'j'} \\ \frac{1}{2} & \text{if} \quad t_{ij} = t_{i'j'} \\ 0 & \text{if} \quad t_{ij} < t_{i'j'} \end{cases}$$

When both diseased and not diseased sites can occur within a patient, the sum corresponding to the AUC statistic includes functions of correlated pairs of diseased and not-diseased observations, violating U-statistic properties. However, relatively few correlated $(D, \overline{D})$ pairs are included in the sum. Let p_p be the patient-prevalence of disease, and let p_s be the expected proportion of sites with disease given a patient has disease. If all patients with disease have the same number of affected sites, then the proportion of correlated $(D, \overline{D})$ pairs is

$$\frac{(1 - p_s)}{(1 - p_p p_s)} \frac{1}{N}.$$

When all patients have a single disease state, $p_s = 1$, there are no correlated $(D, \overline{D})$ pairs. The number of correlated pairs is maximized at $1/N$ when all N patients have disease ($p_p = 1$). When

there are correlated pairs, the U-statistic properties of the AUC statistic can be maintained by excluding correlated pairs from AUC sums. In most cases this exclusion is unnecessary because the number of correlated comparisons quickly becomes negligible as sample size increases. This article examines bootstrap resampling that is directly applied directly to AUC statistics, without excluding correlated pairs.

Bootstrap samples are constructed by stratifying patients on overall disease state (any or none) and drawing patients, the independent units, with replacement from these strata. Resampling patient-level data incorporates all sources of within patient variability. Stratifying the bootstrap samples by patient-level disease state corresponds to conditioning on true disease state, a property of the accuracy statistics sensitivity, specificity and AUC, and reflects the sampling strategy often used when evaluating diagnostic test performance. Accuracy statistics are calculated for each bootstrap sample. Point estimates are simple averages of statistics. The accuracy of two tests can be compared by calculating the difference in accuracy statistics for each bootstrap sample, incorporating between test correlation. Standard errors are estimated by observed standard errors across bootstrap samples. Standard errors should be based on at least 100 draws. Confidence intervals can be estimated using bootstrap estimated standard errors, with a normal approximation. Confidence intervals can also be estimated using percentiles, though this requires at least 1,000 bootstrap draws.

4. Angiography Study

Contrast angiography (CA) is the usual method for mapping vascular occlusion prior to bypass surgery in patients with peripheral vascular disease. Magnetic resonance angiography (MRA)

is an alternative method for obtaining the same diagnostic information. MRA is less invasive than CA because it does not require injection of contrast materials. The ability of CA and MRA to correctly identify open vessel segments was compared using a prospective study, with intraoperative angiography used as the gold standard.(Baum et al, 1995) Patients were evaluated at 15 sites (vessel segments). Analyses were based on 96 patients with peripheral vascular disease who had intraoperative angiography results and at least one preoperative angiographic tests (eleven patients were missing an MR assessment). On average, 33% of each patient's vessel segments were occluded. Overall, 335 of 932 segments with gold standard information were occluded (36%). Study radiologists rated the occlusion of each vessel segments using a five-point scale: 1) normal; 2) minimal disease (<50% stenosis); 3) stenotic (a single lesion with $\geq$ 50% stenosis but not fully occluded); 4) diffuse disease (multiple lesions with $\geq$ 50% stenosis but not fully occluded); 5) fully occluded. Ratings within patients were moderately correlated, with similar degrees of correlation for the two tests. Overall, correlation (based on Kendall's τ) was 0.20 for CA and 0.19 for MRA. Correlation between sites with the same disease state was 0.49 for CA and 0.46 for MRA. Correlation between sites with different disease states was -0.34 for CA and -0.36 for MRA. Correlation between CA and MRA ratings of the same site was 0.62.

Original analyses examined both detection of near-normal (patent) vessel segments (ratings 1 and 2) and detection of open segments (ratings 1 through 4). Both methods were similar in their ability to detect open vessel segments: both had 81% specificity, CA had an 83% sensitivity and MRA had an 85% sensitivity. In detecting patent segments, CA was less sensitive than MRA (77% versus 82%) but more specific (92% versus 84%). Based on these descriptive data and statistical tests for differences in odds ratios, the original investigators concluded that CA

and MRA had similar diagnostic ability. Bootstrap estimation allows us estimate AUC statistics for patent segments, to examine whether differences are likely due to a threshold effect, and to place confidence intervals on estimated sensitivity and specificity. We report bootstrap percentile intervals based on 1000 bootstrap samples. Bootstrap estimates of sensitivity were CA: 76% with 95% CI (70.5,81.8) and MRA: 82% with 95% CI (76.8,87.0). Bootstrap estimates of specificity were CA: 93% false positive rate with 95% CI (89.8,95.9); MRA: 84% with 95% CI (79.4,88.1). CA and MRA had similar empirical AUC statistic. The empirical AUC for CA was 0.879 with 95% confidence interval (0.847,0.910). The empirical AUC for MRA was 0.874 with 95% confidence interval (0.844,0.904). The bootstrap estimate of the difference in AUC statistics was 0.005 with 95% confidence interval (−0.035,0.044).

5. Simulation Study

This simulation study describes characteristics of bootstrap accuracy and compares them to Obuchowski's analytic estimates (Obuchowski, 1997). Bootstrap confidence intervals based on normal approximations were based on 100 bootstrap samples. Bootstrap confidence intervals based on percentiles were based on 1000 bootstrap samples. Comparisons focus on the observed coverage of 95% confidence intervals for differences between two AUC statistics. The description of bootstrap estimates also includes coverage rates for estimated false positive rates.

Simulated data represent comparisons between two tests (A and B) with outcomes on a 5-point ordinal scale. Test A has an empirical AUC equal to 0.8 and specificities equal to (0.5, 0.7,0.9.0.95). Test B has the same specificities and an AUC statistic equal to 0.80 or 0.85. The bootstrap's ability to handle multiple sources of variability was evaluated by simulating outcomes

for two readers per test. The overall diagnostic accuracy of each test was based on the average of the two readers' AUC statistics. Data simulated for two readers assumes that readers evaluating the same test had equal accuracy, with the same specificities and the same AUC statistics. Two reader bootstrap AUC estimates were calculated by estimating each reader's AUC statistic then averaging these within each bootstrap sample.

Ordinal test outcomes were simulated by categorizing continuous multivariate normal (MVN) psuedodeviates. One MVN pseudodeviate of length $4m$ was generated for each patient-observation, where m is the number of sites within patients. Each independent MVN pseudodeviate represents a single patient's unobservable continuous test outcome for 2 tests and 2 readers. Within patient correlation was induced on the continuous scale. Simulations examine the characteristics of estimators for three within subject correlation structures: independent, compound symmetry, and disease-dependent. Under the compound symmetry structure, multiple observations within subjects are equicorrelated, with correlation equal to 0.50. The disease-dependent structure is identical to the compound symmetry structure with one exception: Under the disease-dependent structure, observations from sites with different disease states (i.e., $(D,\overline{D})$ pairs) are negatively correlated, with correlation equal to -0.50.

Simulations examine 3 sampling scenarios. Under the first scenario (small N) 100 patients, 50 with disease and 50 without, are evaluated at 4 sites. Under the second scenario (large m) 100 patients, all with disease, are evaluated at 15 sites. Under the third scenario (large N), 500 patients, 250 with disease and 250 without, are evaluated at 4 sites. For all scenarios, patients with disease are expected to have disease at half of the sites. The number of disease-positive sites for each patient was simulated using a binomial random number generator. Ordinal

ratings were derived from MVN deviates by assuming an underlying bivariate normal ROC model (Hanley, 1989). That is, 'cut points' for each for the five rating categories are set equal to $\theta_0 = -\infty$, $\theta_k = \Phi^{-1}(1 - \text{specificity}_k)$, $k = 1, \ldots, 4$, and $\theta_5 = +\infty$. Given μ and θ, sensitivities were $\text{sensitivity}_k = \Phi(\theta_k + \mu)$. Desired empirical AUC statistics were obtained with μ =1.29 for AUC=0.80 and μ =1.949 for AUC=0.85. For disease negative sites, the ordinal rating corresponding to the MVN deviate y is equal to k when $\theta_{k-1} < y < \theta_k$. For disease positive sites, the MVN deviates are first shifted by an appropriate μ, with simulated ratings based on categorizing $y + \mu$.

Simulation results were based on 5,000 simulated data sets for each combination of AUC_B (0.80 or 0.85), sampling scenario (small N, large m, or large N), and correlation structure (independent, equicorrelated, and disease-dependent).

6. Simulation Results:

Table 1 shows the observed within-patient correlation in the simulated categorical data. These rating data are inherently correlated, since diseased sites were more likely to have high scores than not diseased sites.

Table 2 shows coverage rates of 95% confidence intervals for the difference between AUC_A and AUC_B based on Obuchowski's analytic estimator, the single reader bootstrap percentile interval and the two reader bootstrap percentile interval. Coverage rates for normal-approximation bootstrap intervals are not shown because they were similar to percentile intervals, with slightly poorer coverage properties. In general, coverage rates of the percentile interval fell between coverage rates for the analytic and bootstrap percentile intervals. Both the sampling scenario and the correlation structure affected the observed coverage rates, but true differences between

the two AUC statistics did not affect coverage. Within the small N and large m sampling schemes, intervals estimated from data with compound symmetry correlation tended to have coverage rates that were further from the 95% level than estimates based on independent data The bootstrap and estimates had very similar coverage rates for the small N and large N scenarios. The analytic estimator had better coverage for the large m scenario. Single reader and two reader bootstrap estimates had similar coverage rates.

The analytic estimator and the single reader bootstrap estimator had similar mean squared errors (MSE's). Across simulated data sets, the MSE of the bootstrap estimate for one reader was less than 0.1% higher than the MSE of the analytic estimate. The MSE of two reader bootstrap estimates were approximately half the MSE of either single-reader estimate.

Table 4 shows coverage rates of bootstrap percentile interval estimates for specificities. Coverage rates were generally less than the nominal level, but improved as the specificity decreased from 0.95 to 0.50 and as the amount of data available for estimation increased Percentile intervals had better coverage rates than asymptotic normal intervals (not shown). When specificity was 0.95, a few (less than 1%) of the asymptotic normal bootstrap intervals fell outside of the $(0, 1)$ range.

7. Discussion

Diagnostic evaluation often involves testing patients at multiple sites. Bootstrap and analytic estimation methods allow simple comparisons of AUC statistics based on clustered patient data. These methods are asymptotically consistent. However, evaluation of diagnostic tests often involve relatively small sample sizes. We used a simulation study to evaluate the small sample

characteristics of Obuchowski's analytic AUC estimator and bootstrap AUC estimators applied to ordinal test data. When comparing two tests with one reader per test, the bootstrap and analytic estimators had very similar performance. Both methods produced confidence intervals with observed coverage rates below the nominal level. Coverage rates of bootstrap percentile confidence intervals were nearly identical to asymptotic normal intervals for AUC statistics. However, percentile intervals had better coverage than asymptotic normal intervals for proportions. Although these methods are asymptotically consistent, simulations suggest that when test outcomes are ordinal and tests are relatively accurate, large samples are needed before asymptotic results hold.

The simulations presented in this article demonstrated poorer performance for Obuchowski's estimator than was originally reported. There are important difference in the simulations in this article and those presented by Obuchowski. Two key differences are site-level prevalence of disease and the scale of the test outcome. In the smallest sample setting, patient level disease prevalence was 50% and among patients with disease an average of 50% of sites were affected, resulting in a 25% overall site-level prevalence. Obuchowski simulated data with an overall 50% site-level prevalence. Obuchowski also generated outcomes on a continuous 0 to 100 scale, rather than the 5-point ordinal scale more commonly found in radiology. The continuous scale allows more variability in true positive (tp) and false positive (fp) rates. A comparison between continuous scales is also more informative than a comparison between corresponding ordinal scales because there are no ties.

Simulation studies examine the behavior of estimators in specific settings. The simulation study examined plausible scenarios. In radiology research tests outcomes are often measured on 5-point ordinal scale, and these tests can be highly accurate withe relatively high specificity.

Overall sample sizes are often small, including 100 or fewer subjects. There were some important assumptions that may limit the conclusions that can be drawn from simulation study findings. One important assumption made for these simulations was that the two tests compared had the same underlying sensitivities. Perhaps the strongest assumption made for simulated data was that when two readers were involved they each had identical ROC curves. In real life settings, readers ROC curves will almost certainly differ. In this context, the investigator must determine whether there is value in the estimate of the average AUC statistic.

References

[1] Zerhouni EA, Rutter CM, Hamilton SR, Balf DM, Megibow AJ, Francis IR, Moss AA, Heiken JP, Tempany CMC, Aisen AM, Weinreb J, Gatsonis C, McNeil BJ. CT and MRI imaging in the staging of colorectal carcinoma: Report of the Radiologic Diagnostic Oncology Group II, *Radiology* **1996;** 200: 443–51.

[2] Baum RA, Rutter CM, Sunshine JH, Blebea JS, Blebea J, Carpenter JP, Dickey KW, Quinn SF, Gomes AS, Grist TM, McNeil BJ for the American College of Radiology Rapid Technology Assessment Group. Multi-center trial to evaluate peripheral vascular magnetic resonance angiography, *Journal of the American Medical Association***1995;** 274: 875-880.

[3] Song HH. Analysis of Correlated ROC Areas in Diagnostic Testing, *Biomet-*

rics **1997;** 53: 370–382.

[4] Lipsitz LR, Harrington DP. Analyzing correlated binary data using SAS, *Computers and Biomedical Research***1990;** 23: 268-282.

[5] Obuchowski NA. Nonparametric Analysis of Clustered ROC Curve Data *Biometrics* **1997;** 53: 567-578.

[6] Pepe MS. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results, *Biometrics* **1998;** 54: 124-135.

[7] Hanley JA and McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology* **1982;** 143: 29–36.

[8] Lee AJ. *U-Statistics, Theory and Practice,* New York: Marcel Decker, **1990.**

[9] Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art, *Critical Reviews in Diagnostic Imaging***1989;** 29: 307-335.

[10] Bickel PJ, Freedman DA. Some asymptotic theory for the bootstrap, *The Annals of Statistics***1981;** 9: 1196–1217, 1981.

[11] Arcones MA, Giné E. On the bootstrap of U and V statistics, *The Annals of Statistics* **1992;** 20: 655–674.

Table 1. Average observed correlation of simulated rating data (Kendall's τ) between tests when both have $AUC = 0.80$ and between readers evaluating Test A.

type of correlation	correlation structure		
	independence	compound symmetry	disease dependent
between sites, same disease state	0.199	0.478	0.478
between sites, different disease state	0.164	0.457	0.457
between tests, same site	0.000	0.348	-0.240
between readers, same site	0.200	0.478	0.478

Table 2. Observed coverage rates of 95% confidence intervals. In all cases $AUC_A = 0.8$. Small N simulations generate data from 50 diseased patients and 50 not diseased patients, each evaluated at 4 sites by both tests. Large m simulations generate data from 100 diseased patients, each evaluated at 15 sites by both tests. Large N simulations generate data from 250 diseased patients and 250 not diseased patients, each evaluated at 4 sites by both tests.

correlation structure	estimator	sampling design small N	large m	large N
independent	analytic	0.934	0.932	0.946
	1 reader bootstrap	0.939	0.944	0.946
	2 reader bootstrap	0.935	0.947	0.949
equicorrelated	analytic	0.936	0.934	0.952
	1 reader bootstrap	0.939	0.946	0.950
	2 reader bootstrap	0.939	0.947	0.950
disease dependent	analytic	0.939	0.936	0.950
	1 reader bootstrap	0.940	0.943	0.948
	2 reader bootstrap	0.938	0.944	0.950

Table 3. Observed coverage rates of 95% confidence intervals. In all cases $AUC_A = 0.8$. Small N simulations generate data from 50 diseased patients and 50 not diseased patients, each evaluated at 4 sites by both tests. Large m simulations generate data from 100 diseased patients, each evaluated at 15 sites by both tests. Large N simulations generate data from 250 diseased patients and 250 not diseased patients, each evaluated at 4 sites by both tests.

correlation structure	small N	large m	large N
		fp=0.95	
independent observations			
asymptotic normal	0.929	0.936	0.941
percentile	0.942	0.941	
equicorrelated, correlation=0.5			
asymptotic normal	0.912	0.923	0.936
percentile	0.926	0.926	
correlation dependent on disease state			
asymptotic normal	0.911	0.912	0.940
percentile	0.926	0.928	
		fp=0.50	
independent observations			
asymptotic normal	0.942	0.942	0.946
percentile	0.944	0.949	
equicorrelated, correlation=0.5			
asymptotic normal	0.941	0.942	0.948
percentile	0.948	0.947	
correlation dependent on disease state			
asymptotic normal	0.941	0.941	0.950
percentile	0.948	0.941	

Appendix C

A hierarchical regresssion approach to meta-analysis of diagnostic test evaluations

CM Rutter and CA Gatsonis

UNPUBLISHED MANUSCRIPT

Funding number: DAMD17-97-1-7193

A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations

Carolyn M. Rutter and Constantine A. Gatsonis

Carolyn M. Rutter

Group Health Cooperative, Center for Health Studies
1730 Minor Avenue, Suite 1600
Seattle, WA 98101
e-mail: rutter.c@ghc.org
phone: 206.287.2190
fax: 206.287.2871

Constantine A. Gatsonis

Center for Statistical Sciences
Brown University
Box G-H
Providence, RI 02912
email: gatsonis@stat.brown.edu
phone: 401.863.9183
fax: 401.863.9182

A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations

Summary

An important quality of meta-analytic models for research synthesis is their ability to account for both within- and between- study variability. Currently available meta-analytic approaches for studies of diagnostic test accuracy work primarily within a fixed-effects framework. In this paper we describe a hierarchical regression model for meta-analysis of studies reporting estimates of test sensitivity and specificity. The model allows more between- and within-study variability than fixed-effect approaches, by allowing both test stringency and test accuracy to vary across studies. It is also possible to examine the effects of study specific covariates. Estimates are computed using Markov Chain Monte Carlo simulation, allowing flexibility in the choice of summary statistics. We demonstrate our modelling approach using a recently published meta-analysis comparing three tests used to detect nodal metastasis of cervical cancer.

keywords: summary ROC curve, Bayesian methods, sensitivity, specificity.

1. Introduction

The need for systematic review and synthesis of published evidence on the accuracy of diagnostic tests has increased in recent years. The information from such reviews is a key element of clinical and health policy decision making regarding the use of diagnostic tests; it is also essential for guiding the process of technology development and evaluation in diagnostic medicine [1, 2].

Statistical methods for meta-analysis of diagnostic test evaluations have focused on the analysis of studies reporting estimates of test sensitivity and specificity, the most commonly used measures of diagnostic performance, and have worked within the fixed-effects framework [1, 2, 3, 4, 5, 6, 7, 8, 9]. A fundamental concern in the synthesis of such studies is derivation of summary measures of test performance. These measures must account for the tradeoff between sensitivity and specificity as the threshold for positivity varies along some explicit or latent scale. This tradeoff has been widely recognized in the evaluation of diagnostic tests and has led to the development of Receiver Operating Characteristic (ROC) methodology [10]. In the context of meta-analysis, simple averaging or pooling across studies can provide misleading conclusions, as can be readily seen from a simple example. If three studies report the following estimates of test sensitivity and specificity: (.10, .90), (.80, .80), and (.90, .10), the average pair of sensitivity and specificity is (.60, .60) and lies completely outside the domain of the original studies (see also [1, 2, 6]).

Differences in positivity threshold constitute an important source of variation across studies evaluating a diagnostic modality. Study characteristics, such as technical aspects of the diagnostic test, patient and disease cohorts, study settings, experience of readers, and sample size are also potential contributors to between-studies variations in the estimates of diagnostic performance.

In the fixed effects setting, regression models have been proposed for exploring these sources of variability [4, 5]. The use of regression models provides a flexible and powerful framework for meta-analysis. However, the number of covariates that can be accommodated in such models is limited. In addition, these fixed-effects regression models may not provide realistic accounts of the uncertainty associated with covariate estimates.

In this paper, we expand on earlier work [5] and present a hierarchical model formulation of the problem of combining information across studies reporting estimates of test sensitivity and specificity. The structure of the model is similar to that of models proposed for the meta-analysis of treatment studies [11, 12, 13]. The observed variation is partitioned into *within-* and *between* studies components. Each component consists of a *systematic* part and a *random* part, with the former attributed to covariates and the latter to unexplained variation. The hierarchical model makes it possible to pool information across studies and derive smoothed estimates of covariate effects, components of variance and individual study quantities. In addition, simple extensions of the hierarchical structure can incorporate patient-level information within each study, when such information is available.

We present our approach using data from a recently published meta-analysis comparing the diagnostic performance of three imaging modalities for the detection of lymph node metastases in women with cervical cancer [14]. In section 2 we survey fixed effects approaches to the problem. The hierarchical regression model is presented in section 3. We take a fully Bayesian approach to model fitting and checking and use Markov Chain Monte Carlo estimation techniques. Technical issues are discussed in section 3 and the analysis of the example is presented in section 4. The final section summarizes our methodological and subject matter conclusions.

2. Meta-analytic models for diagnostic test data

The simplest setting for the methods discussed in this paper involves meta-analyses in which each of m studies contributes a vector z_i of study-level covariates ($i = 1, ..., m$) and a 2x2 table of summary data, showing the agreement between the binary test result and the definitive disease information ("gold standard"). We will use the following notation:

		Test: 0=no	Test: 1=yes	
Truth:	no	y_{i00}	y_{i01}	n_{i0}
	yes	y_{i10}	y_{i11}	n_{i1}

The observed rates of true and false positive test results are then defined as $\widehat{TP}_i = y_{i11}/n_{i1}$ and $\widehat{FP}_i = y_{i01}/n_{i0}$. In some meta-analyses more than one 2x2 table is available from each study. For example, patients may be examined using several tests in a study, leading to correlated binary test results studies.

2.1 Summary ROC (SROC) curve

In the absence of patient and study level covariates, a simple graphical summary of test accuracy is provided by the summary ROC curve (SROC) [4]. The curve is constructed by computing the quantities

$$D_i = \text{logit}(\widehat{TP}_i) - \text{logit}(\widehat{FP}_i) \qquad \text{and} \qquad S_i = \text{logit}(\widehat{TP}_i) + \text{logit}(\widehat{FP}_i)$$

for each study and fitting the linear model

$$D_i = a + bS_i + e_i \tag{1}$$

where e_i is random error. The model can be fitted using ordinary or weighted least squares, or robust regression methods. Weights can be used to account for between-study differences in overall sample size or precision. However, weights cannot simultaneously capture differences in sample size within the disease-positive (n_{i1}) and disease-negative (n_{i0}) groups. These two sample sizes affect the precision of estimated TP and FP rates independently. In practice, weighted and unweighted models can produce very different results, and there is no clear way to choose between these models.

Using the estimates of a and b, a plot of the summary ROC curve can be drawn, with FP on the x-axis and TP on the y-axis. This SROC model corresponds to the assumption that the observed differences across studies result from different thresholds for test positivity. The summary curve is symmetric if $b = 0$, implying constant log-odds across the studies under review. Study level covariates can be incorporated in straightforward manner into equation (1) to provide an exploratory analysis of the effects of study characteristics. Several summaries of the SROC curve have been proposed and can be used to make comparisons between modalities. However, the SROC model does not account for error in S_j and this can bias parameter estimates[15] and summaries that are functions of these estimates. Further exploration is needed to determine the effect of ignoring error in S_j on both point estimation and coverage rates of estimated confidence intervals.

An alternative approach to constructing an SROC curve was proposed by Kardaun and Kardaun[3], who assumed that $\left(\text{logit}(\widehat{TP}_i), \text{logit}(\widehat{FP}_i)\right)$ follows a bivariate normal distribution and postulated a linear relationship between the two components of the bivariate mean. Profile likelihood is used to derive estimates of the slope and intercept in this model, which includes

variability in both rates. Difficulties incorporating covariates into this model limit its usefulness.

2.2 Binomial regression model

A regression model for the meta-analysis of $(\widehat{TP}_i, \widehat{FP}_i)$ pairs was first discussed in [5] and was motivated by the ordinal regression formulation of ROC analysis [16, 17, 18]. In brief, if W denotes the degree of suspicion about the presence of an abnormality, elicited on an ordinal categorical scale with J categories, the parametric ROC model is equivalent to the ordinal regression model $g(P[W \geq j|X]) = (\theta_j + \alpha X)\exp(-\beta X)$, where X is a covariate denoting the (binary) true disease status. The conceptual basis of the model is an assumption that the observed responses W represent a categorization of a latent variable, with distribution corresponding to the link function, $g(\cdot)$. The probit link implies a Gaussian latent variable and is commonly used for single-study receiver operating characteristic analysis [10]. We use the logit link throughout this article because under the logit model, regression parameters estimate log-odds ratios.

As discussed in [5], an ordinal regression model with a logistic link and $J = 2$ can be used in the meta-analysis of studies reporting pairs of $(\widehat{TP}_i, \widehat{FP}_i)$. Under this model, the numbers of positive tests from each study, $y_{ij1}, i = 1, ..., m, j = 0, 1$ are assumed to follow binomial distributions, $y_{ij1} \sim$ Binomial(n_{ij}, π_{ij}), in which the probability of a positive test modelled as:

$$\pi_{ij} = \text{logit}^{-1}((\theta_i + \alpha X_{ij})e^{-\beta X_{ij}}). \tag{2}$$

As in the ROC context, the θ_i's will be called the "positivity criteria" (or "cutpoint parameters"), α the"accuracy parameter" and β the "scale parameter". The tradeoff between TP and FP is modelled through their joint dependence on θ. The binomial regression model is estimated by maximum likelihood and accounts for error in both $\widehat{TP}_i$ and $\widehat{FP}_i$ rates.

It can be readily seen that the binomial regression model (2) implies a linear relationship between logit(TP) and logit(FP). This linearity is a basic assumption in the two SROC models discussed earlier and implies a natural correspondence between SROC and the binary regression analysis. Like the SROC model, the simple binary regression model (2) assumes that observed differences across studies result from different positivity thresholds (θ_i). Tests are assumed to have the same accuracy and scale parameter across studies. In addition, as discussed in [5], more elaborate versions of (2) can be formulated, in which parameters other than θ vary across studies and study level covariates are included. In practice, making choices among such elaborate models is not a straightforward matter. An important consideration in making such choices is to ensure that the resulting model is identifiable.

3. Hierarchical regression analysis

3.1 Model

Hierarchical regression analysis extends the binomial regression model to more fully account for both within and between study variability in TP and FP rate. The model allows the inclusion of patient- and study-level covariates, if such information is available, and has the following structure:

Level I (*Within-study variation*) The number of positive tests from the i-th study, y_{i01} and y_{i11}, are assumed to follow binomial distributions, with the probability of a positive test given by:

$$\pi_{ij} = \text{logit}^{-1}[(\theta_i + \alpha_i X_{ij})\text{e}^{-\beta X_{ij}}] \tag{3}$$

where X_{ij} denotes the true disease status for cases in the ij-th cell. Under this hierarchical SROC model (HSROC), both positivity criteria (θ_i) and accuracy parameters (α_i) are allowed to vary

across studies.

Level II (*Between-study variation*) Study-level parameters in (3), α_i and θ_i, are assumed to be Normally distributed, with mean determined by a linear function of study-level covariates. In the case of a single covariate Z affecting both the cutpoint and accuracy parameters, the model can be written as:

$$\left.\begin{array}{rcl} \theta_i|\Theta,\gamma,Z_i,\sigma_\theta^2 & \sim & N(\Theta+\gamma Z_i,\sigma_\theta^2) \\ \alpha_i|\Lambda,\lambda,Z_i,\sigma_\alpha^2 & \sim & N(\Lambda+\lambda Z_i,\sigma_\alpha^2) \end{array}\right\}\text{conditionally independent}$$

The coefficients γ and λ model systematic differences in positivity criteria and accuracy across studies, due to the covariate Z. However, more general formulations of the model can be considered in which more than one covariate is included and different covariates are used for 'cutpoint' and 'accuracy' regression equations.

Level III The specification of the hierarchical model is completed by the choice of prior distributions for the remaining unknown parameters. In particular, we chose:

$$\Theta \sim \text{Uniform}[\mu_{\theta 1},\mu_{\theta 2}];\quad \gamma \sim \text{Uniform}[\mu_{\gamma 1},\mu_{\gamma 2}];\quad \sigma_\theta^2 \sim \Gamma^{-1}(\xi_{\theta 1},\xi_{\theta 2})$$

$$\Lambda \sim \text{Uniform}[\mu_{\alpha 1},\mu_{\alpha 2}];\quad \lambda \sim \text{Uniform}[\mu_{\lambda 1},\mu_{\lambda 2}];\quad \sigma_\alpha^2 \sim \Gamma^{-1}(\xi_{\alpha 1},\xi_{\alpha 2})$$

$$\beta \sim \text{Uniform}[\mu_{\beta 1},\mu_{\beta 2}]$$

The parameters Θ, Λ, β, γ, λ, σ_θ^2 and σ_α^2 are assumed to be mutually independent. The parameters, $\mu_\theta,\mu_\gamma,\xi_\theta,\mu_\alpha,\mu_\lambda,\xi_\alpha,\mu_\beta$, are assumed to be fixed and are chosen to reflect expected ranges.

Summary ROC (SROC) curves can be derived using the expected values of $\Lambda+\lambda Z$ and β. If true disease state is coded $\frac{1}{2}$ for disease positive cases and $-\frac{1}{2}$ for not diseased cases, then for

a given value of the covariate Z_i, the model-based true positive rate can be expressed as:

$$TP(FP) = \text{logit}^{-1}\left((\text{logit}(FP)e^{E(\beta)/2} + E(\Lambda + \lambda Z_i))e^{E(\beta)/2}\right)$$

The SROC curve is drawn by plotting $(FP, TP(FP))$ for $FP \in [0, 1]$. Extrapolation beyond the available data can be discouraged by plotting curves only over the observed range of FP.

3.2. HSROC Model fitting

Inference from the HSROC model is based on the posterior distributions of model parameters. Because the models we consider are not conjugate, closed form expressions for posterior distributions do not exist. Posterior quantities are estimated by simulating observations from the posterior distribution using Markov Chain Monte Carlo (MCMC) simulation [19]. These simulated values from the full posterior distribution are used to estimate marginal distributions of interest, such as posterior distributions of particular parameters or functions of parameters.

To enable estimation, each covariate was centered at zero. When estimating the fixed effect binomial regression model, covariate centering is required for model identifiability [16]. Under the hierarchical regression model, centering the covariate vector helps to reduce correlation between consecutive draws.

3.2.1 Conditional distributions

The conditional distributions of Level II parameters (Θ, Λ, σ_θ^2, σ_α^2, γ and λ) are standard distributions or truncated versions of standard distributions. For example, the conditional distribution of Λ given $\lambda, z, \sigma_\alpha^2, \alpha$, and μ_α is proportional to a Normal distribution with mean $\sum_{i=1}^{m}(\alpha_i - \lambda Z_i)/m$ and variance σ_α^2/m over the range $[\mu_{\alpha 1}, \mu_{\alpha 2}]$. Variance parameters, σ_α^2 and σ_θ^2, have conjugate

priors, so that

$$(\sigma_\alpha^2|\lambda, z, \alpha, \Lambda, \xi_\alpha) \sim \Gamma^{-1}((\xi_{\alpha 1} + m/2), (\frac{1}{2}\sum_{i=1}^{m}(\alpha_i - \Lambda - \lambda Z_i)^2 + 1/\xi_{\alpha 2})^{-1}). \tag{4}$$

The conditional distributions of Θ and σ_θ^2 are analogous. Finally, the conditional distribution $(\lambda|\Lambda, \sigma_\alpha^2, \alpha, \mu_\lambda)$ is proportional to a normal distribution with mean

$$\frac{\sum_{i=1} Z_i(\alpha_i - \Lambda)}{\sum_{i=1}^{m} Z_i^2}$$

and variance $\sigma_\alpha^2/(\sum_i Z_i^2)$ over the range $[\mu_{\lambda 1}, \mu_{\lambda 2}]$.

The conditional distributions of Level I parameters (θ_i, α_i and β) are not standard. The conditional distribution of study specific accuracies, $(\alpha_i|y_i, n_i, Z_i, \Lambda, \sigma_\alpha^2, \theta_i, \beta)$, is a Binomial–Normal product,

$$\exp\left(-\frac{(\alpha_i - (\Lambda + \lambda Z_i))^2}{2\sigma_\alpha^2}\right) \prod_{j=1}^{2} \binom{n_{ij}}{y_{ij}} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{(n_{ij} - y_{ij})}$$

The conditional distribution of θ_i has similar form. The conditional distribution of the scale parameter, $(\beta|y, n, Z, \theta, \alpha, \mu_\beta)$, is proportional to the product of $2m$ Binomials,

$$\prod_{i=1}^{m} \prod_{j=1}^{2} \binom{n_{ij}}{y_{ij}} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{(n_{ij} - y_{ij})}$$

with positive probability over the range $[\mu_{\beta 1}, \mu_{\beta 2}]$.

3.2.2 Metropolis steps

We simulated draws from the nonstandard conditional distributions of β, θ_i and α_i using an adaptive Metropolis step [20]. The Metropolis algorithm works by simulating candidate draws from a jumping distribution. Candidate draws are accepted if they increase the posterior density, otherwise the chain stays at the current draw. The scale parameter (β) was sampled using a

univariate Normal jumping distribution. Study-level parameters (θ_i,α_i) were jointly sampled using a bivariate Normal jumping distribution. Normal variance and variance-covariance matrices were calculated using a multiple of the inverse information matrix evaluated at the current values of unknown parameters. Because low rejection probabilities can indicate that the jumping distribution is underdispersed relative to the target distribution, covariance matrices were inflated so that rejection rates varied between 20% and 40% for all parameters [21].

3.2.3 Choice of priors

Prior ranges for Θ, Λ and β should be chosen to reflect subject matter knowledge about the diagnostic modalities under review. In general, the interval $[-10, 10]$ covers all reasonable values of Θ. Similarly, the interval $[-5, 5]$ covers all reasonable values of β. Because we expect positive test results, indicating disease, to be more common among patients with disease, the interval $[-2, 20]$ covers all reasonable values of Λ.

Selection of the Inverse Gamma priors for the between-study variance parameters, σ_θ^2 and σ_α^2, is more difficult. The goal in making this choice is to select a relatively diffuse distribution, which nevertheless does not assign much probability to very large values of the variance parameters. For example, although a $\Gamma^{-1}(0.1, 1)$ is quite diffuse, it assigns unduly large weight to large values of σ; the lower quartile of the $\Gamma^{-1}(0.1, 1)$ distribution is 28.35. We chose a $\Gamma^{-1}(1, 2)$ prior for variance parameters because this covers the expected range of variability in these data. Quartiles of the $\Gamma^{-1}(1, 2)$ distribution are 1.44, 2.89, and 6.96. The probability that an $\Gamma^{-1}(1, 2)$ random variable is greater than 9 is 0.20.

3.2.4 Parameter estimation The goal of estimation is description of the posterior distribution of model parameters and summary statistics that are functions of model parameters. Posterior 95%

credible intervals were estimated by empirical 2.5% and 97.5% posterior percentiles of simulated draws. The mode of symmetric posterior distributions was approximated by the mean value across simulated draws (i.e., Θ, Λ, and β). The mode of asymmetric posterior distributions was approximated by the median value across simulated draws (i.e., σ_θ^2 and σ_α^2).

3.2.5 Assessing convergence

Estimation was based on draws from several chains started at extreme points in the parameter space. The CODA program[22] was used to evaluate the convergence to the target distribution. We relied primarily on examination of trace plots and estimates of scale reduction proposed by Gelman and Rubin[21]. The scale reduction statistic is essentially the ratio of the between chain variance to the within chain variance.

3.2.6 Diagnostics

Diagnostic statistics were used to evaluate possible model misspecification, overall goodness of fit, and to identify of outlying and possibly influential data points. Our approach roughly follows the suggestions of Weiss [23].

Checks for model misspecification were restricted to evaluation of the prior distributions for study-specific parameters θ and α. Recall that we assume that both statistics are normally distributed. Under the exchangeable model (i.e. $\gamma = \lambda = 0$), the sums of squares $S_\theta = \sum_i(\theta_i - \Theta)^2/\sigma_\theta^2$ and $S_\alpha = \sum_i(\alpha_i - \Lambda)^2/\sigma_\alpha^2$ should follow a a χ^2 distribution with m degrees of freedom, where m is the number of studies in the meta-analysis. Under the nonexchangeable model sums of squares are of the form $S_\alpha = \sum_i(\alpha_i - \Lambda - \lambda Z_i)^2/\sigma_\alpha^2$. Because large values of tail probabilities suggest misspecification of prior distributions, we estimate $p(\alpha) = P(S_\alpha < \chi^2_{m,0.025}) + P(S_\alpha > \chi^2_{m,0.975})$, and $p(\theta) = P(S_\theta < \chi^2_{m,0.025}) + P(S_\theta > \chi^2_{m,0.975})$, to evaluate these

priors.

Global goodness of fit can be evaluated using two chi-square discrepancy statistic. The first is based on estimated counts:

$$D^2_{\text{count}} = \sum_i \sum_j \frac{(y_{ij1} - E(y_{ij1}|\text{model, data}))^2}{E(y_{ij1}|\text{model, data})}$$

where y_{ij1} is the number of subjects testing positive in not-diseased ($j = 0$) and diseased ($j = 1$) groups. D^2_{count} is compared to a χ^2_{2m} distribution. The second global goodness of fit statistics is based on continuity-corrected log-odds ratios:

$$D^2_{\text{log(or)}} = \sum_i \frac{(\log(OR_{cc})_i - E(\log(OR_{cc})_i|\text{model, data}))^2}{\sqrt{var(\log(OR_{cc})_i|\text{model, data})}}$$

where $\log(OR_{cc})_i$ is the observed continuity corrected log-odds ratio for the i-th study.

Outliers and potentially influential points were identified using plots of sensitivity versus specificity and by examining chi-square residuals, e.g., $(y_{ij} - E(y_{ij}|\text{model, data}))^2/E(y_{ij}|\text{model, data}))$. The sensitivity of the model to potentially outlying and influential points can be examined by removing these points and re-estimating parameters.

4. Example: Evaluation of Lymph Node Metastases

4.1. Data

To demonstrate the hierarchical model, we reanalysed data from a published meta-analysis of diagnostic imaging tests used to detect lymph node metastasis in patients with cervical cancer [14]. This study compared three tests for detection of lymph node metastasis: lymphangiography (LAG), computed tomography (CT), and magnetic resonance imaging (MR). Data were combined from 37 studies, of which 17 examined LAG, 19 examined CT and 10 examined MR. Nine studies

examined more than one test. Observed true positive and false positive rates are shown in Figure 1.

[Figure 1 about here]

The original analysis by Scheidler and colleagues used a fixed effect SROC [4], Q^* statistics, and likelihood ratio statistics to compare tests. The Q^* statistic corresponds to the estimated true positive rate at the point on the SROC curve where the sensitivity is equal to the specificity of the test. SROC curves were estimated separately for each test type. In addition, likelihood ratio statistics were used to describe the post-test change in odds of disease. The likelihood ratio negative (LR^-) estimates the post-test odds of disease given a negative test. The likelihood ratio positive (LR^+) estimates the post-test odds of disease given a positive test. Scheidler et al concluded that the three test had similar diagnostic performance. Although the analysis did not find statistically significant differences among the modalities, the authors noted that MR seemed to perform somewhat better than CT or LAG.

4.2 HSROC Computations

We used HSROC analysis to derive summaries of the diagnostic performance of the three modalities, allowing different expected cutpoint (Θ) and accuracy (Λ) parameters for each modality. Because the shape of the three SROC curves looked different we allowed separate scale parameters (β) for each test. Because the spread of the observed points the three SROC curves seemed to differ across modalities, we allowed separate variance parameters (σ_θ^2, σ_α^2) for each modality.

In these analyses, we coded disease state as $+\frac{1}{2}$ for disease positive cases and $-\frac{1}{2}$ for disease negative cases.

The model was estimated using the combined set of data from all modalities but did not explicitly include correlation terms for data derived from studies that compared two or all three of the modalities. In particular, 2 studies examined CT and LAG, 4 studies examined CT and MR, and 2 studies examined CT twice. Although it is possible to extend the model to cover such correlations, the cross-tabulated data from studies evaluating more than one modality were not available. Because we expect positive correlation between diagnostic test results, we expect that ignoring this correlation could cause a slight conservative bias in comparisons between tests. Results from the combined analysis were compared to those from analyses conducted separately for each modality.

The sampler was run using 8 independent MCMC chains. Experimental runs showed that the sampler was slowly mixing. To ensure coverage of the target distribution, estimation was based on multiple sequences with overdispersed starting points. Because of high between-draw correlation, every 50th iteration was saved from each sequence of 100,000 simulated draws. Metropolis covariance parameters were updated every 1,000 iterations to maintain rejection rates between 20% and 40%. Eight different chains were run, with starting points based on β, Θ and Λ: $\beta^{(0)} \in \{-2.5, 2.5\}$, $\Theta^{(0)} \in \{-5, 5\}$, and $\Lambda^{(0)} \in \{-1, 10\}$. Starting values for the prior variability of study specific cutpoints (σ_θ^2) and accuracies (σ_α^2) were set to 9, nearly half of what we believe is a reasonable range for θ_i and α_i. All parameters had estimated scale reduction statistics[20] that were between 1.00 and 1.09 and most were between 1.00 and 1.01, indicating that the sampler had converged. All saved iterations were used to evaluate convergence, but the

first 5,000 were excluded from point and interval estimation.

Summary statistics were calculated for each draw of the sampler, and these values were used to estimate their posterior modes and 95% credible intervals. The summary statistics we used were the overall likelihood ratio positive, overall likelihood ratio negative, overall TP rate, and overall FP rate. These overall performance estimates were calculated for each test. Statistics describing overall test performance were functions of parameters describing performance across studies: the level I model parameter β and level II model parameters describing expected values of study-level parameters. For example, overall TP and FP rates for CT are estimated from:

$$\begin{aligned} TP_{CT} &= \text{logit}^{-1}[(\Theta_{CT} + \Lambda_{CT}/2)\mathrm{e}^{-\beta_{CT}/2}] \\ FP_{CT} &= \text{logit}^{-1}[(\Theta_{CT} - \Lambda_{CT}/2)\mathrm{e}^{\beta_{CT}/2}] \end{aligned}$$

The (TP_{CT}, FP_{CT}) pair summarizes the overall performance of the CT. Likelihood ratio statistics were estimated using overall TP and FP for each test, for example, $LR^{+}_{CT} = TP_{CT}/FP_{CT}$ and $LR^{-}_{CT} = (1 - TP_{CT})/(1 - FP_{CT})$.

Probability estimates were based on the overall proportion of times a statement was true. These estimates also exclude the first 5,000 iterations.

4.3. Results

Table 1 shows parameter estimates based on MCMC estimation. Although 95% credible intervals for all three scale parameters included zero, there was evidence that the scale parameter for LAG was different than the scale parameters for CT. The estimated mode of $\beta_{LAG} - \beta_{CT}$ was 1.34, with 95% credible interval (0.068,2.71). The positivity criteria across studies of LAG tended to be less variable than the positivity criteria used in studies of both CT (with estimated probability 0.910) and MR (with estimated probability 0.960).

[Table 1 about here]

Figure 2 shows estimated SROC curves based on estimated expected values of $(\Lambda_{LAG}, \beta_{LAG})$, $(\Lambda_{CT}, \beta_{CT})$, and $(\Lambda_{MR}, \beta_{MR})$. To avoid extrapolation beyond the data, curves are plotted over the observed ranges of false positive rates.

[Figure 2 about here]

Comparisons based on overall measures of test performance show that CT and MR tended to have lower expected FP and TP rates than LAG (Table 2). There was also evidence of differences in the likelihood ratio positive and likelihood ratio negative of the three modalities (Table 2). The estimated probability that LAG had a better (lower) LR^- than CT was 0.951. LAG also had a lower LR^- than MR with an estimated probability of 0.761. This is evidence that negative LAG results are more informative than negative CT or MR results. On the other hand, the likelihood ratio positive (LR^+) for LAG was worse (lower) than the LR^+ for CT with probability 0.866, and worse than the LR^+ for MR with probability 0.976. This is evidence that positive CT or MR results are more informative than positive LAG results.

[Table 2 about here]

4.4. Model checks

There was no evidence of significant lack of fit in the HSROC model. Estimated TP and FP rates were close to observed values (χ^2_{74}=23.45, p-value=1.000), as were estimated log-odds ratios (χ^2_{37}=7.71, p-value=1.000). Chi-square degrees of freedom for goodness-of-fit statistics were calculated using the number of independent studies. Normal distributions seemed to reasonably approximate the distribution of cutpoint parameters (observed 5% tail probability=4.41%) and accuracy parameters (observed 5% tail probability=4.44%).

None of the study results were identified as influential based on chi-square residuals. Fitted plots showed two points with outlying FP rates for LAG. Results from analyses that excluded these points were similar to results based on the full data set and therefore these studies were retained in analyses.

5. Discussion

The hierarchical summary ROC (HSROC) model for combining estimated pairs of sensitivity and specificity from multiple studies extends the currently used fixed-effects summary ROC (SROC) model. The HSROC model describes within-study variability using a binomial distribution for the number of positive tests in diseased and not diseased patients. An underlying ROC model that allows variability in both the positivity criteria and accuracy across studies determines the binomial probabilities. Variation in positivity criteria and accuracy is modelled using a Normal distribution, with a linear regression in the mean that allows dependence on study-level covariates. More heavy tailed distributions (such as t or Cauchy) can also be used instead of a Normal in the second level of the hierarchical model. As is commonly the case with hierarchical regression

models, the HSROC model allows more complete accounting of between-study variability than is possible with fixed-effects formulations. In addition, the HSROC model provides more realistic accounting of within-study variability that the original fixed-effects SROC model [4], which used a Normal error distribution and did not account for the measurement error in the primary covariate.

The HSROC approach provides a flexible modelling framework that can be extended when more information is available. For example, when studies report results from more than one modality, the hierarchical model can be appropriately extended to incorporate within-study correlation. This extension requires information about the joint distribution of test results, either from multiple similar pairs across several studies, from cross-tabulation of test results within studies, or from patient-level data within studies. When patient level information is available, the within-study (Level I) model can be extended to incorporate patient-level covariates. This extended model can also be applied to data from a single study when results are clustered within participating institutions and/or readers (see [24] for a hierarchical analysis of ROC data).

The fully Bayesian approach to model fitting, although computationally intensive, leads to simulated values from the posterior distribution of the parameters, on the basis of which the analyst can easily calculate summaries of the posterior distribution of a broad range of functions of the parameters. For example, in our reanalysis of the Scheindler data we derived estimates of functionals of the posterior distribution of likelihood ratio statistics, and differences between likelihood ratio statistics for the three modalities. On the basis of these estimates it appears that LAG provides different clinical information than CT or MR, even though all three tests had similar overall accuracy. Bayesian modelling allowed us to express these findings via probabilistic statements. Such probabilistic estimates may be easier to interpret than classical frequentist

summaries.

The Bayesian model also allows description of sources of variability. The differences we found in the variability of positivity criteria were consistent with the technological development of these three tests. At one extreme, LAG is a widely used standard diagnostic test with well developed positivity criteria. At the other extreme, MR was a new diagnostic approach at the beginning of the meta-analyzed time period, without an accepted positivity criteria. The estimated variability of cutpoint parameters was low for LAG. The variability of CT cutpoints was more than twice the variability of LAG cutpoints, and the variability of MR cutpoints was more than four times the variability of LAG cutpoints. This suggests that MR accuracy could be improved through the definition and adoption of good positivity criteria.

The advantages of the HSROC model come at a price: estimation requires Markov Chain Monte Carlo (MCMC) simulation. MCMC estimation requires programming, simulation, evaluation of convergence and model adequacy, and synthesis of simulation results. Programming the MCMC simulation can be time consuming. Although some versions of the proposed model can be fitted within publicly available software (BUGS[25]) the full analysis is elaborate and, depending on the specific model under consideration, may require extensive programming. Even if the burden of programming task was eliminated, implementation of MCMC simulation will still entail nontrivial analysis tasks including evaluation of convergence and the adequacy of prior distributions and this requires some statistical expertise. However, the increased complexity of the proposed analysis must be measured against the advantages from the approach, including more realistic assumptions, more precise description of the impact of covariates, and greater flexibility in choice of descriptive statistics.

References

[1] Irwig L., Tosteson A.N., Gatsonis C.A., Lau J., Colditz G., Chalmers T.C. and Mosteller F. "Guidelines for meta-analyses evaluating diagnostic tests", *Ann. Int. Med.*, 120, 667–676 (1994).

[2] Irwig L., Macaskill P., Glasziou P. amd Fahey M. Meta-analytic methods for diagnostic test accuracy. *J. Clin. Epi* 48:119-130 (1995)

[3] Kardaun J.W.P.F. and Kardaun O.J.W.F. "Comparative Diagnostic Performance of Three Radiological Procedures for the Detection of Lumbar Disk Herniation", *Methods of Info. in Med.*, 29, 12–22 (1990).

[4] Moses L.E., Shapiro D. and Littenberg B. "Combining Independent Studies of a Diagnostic Test Into a Summary ROC Curve: Data–Analytic Approaches and Some Additional Considerations", *Statistics in Medicine*, 12, 1293–1316 (1993).

[5] Rutter C.M. and Gatsonis, C. "Regression Methods for Meta–analysis of Diagnostic Test Data", *Academic Radiology*, 2, S48-S56 (1995).

[6] Hasselblad V., Mosteller F., Littenberg B., Chalmers T.C., Hunink M.G., Turner J.A., Morton S.C., Diehr P., Wong J.B. and Powe N.R. "A Survey of Current Problems in Meta-Analysis", *Medical Care*, 33, 202–220 (1995).

[7] Hasselblad V. and Hedges L.V. Meta-analysis of screening and diagnostic tests. *Psych Bulletin*, 117:167-178 (1995)

[8] Shapiro D. Issues in combining independent estimates of sensitivity and specificity of a diagnostic test. *Academic Radiology* 2:S37-47 (1995).

[9] de Vries S.O., Hunink M. and Polak J. Summary ROC curves as a technique for meta-analysis of the diagnostic performance of duplex ultrasonography in peripheral arterial disease. *Academic Radiology*, 3:361-369 (1996)

[10] Hanley J.A. "Receiver Operating Characteristic (ROC) Methodology: The State of the Art", *Critical Reviews in Diagnostic Imaging* 29, 307–335 (1989).

[11] DuMouchel W. Bayesian meta-analysis in Statistical Methodology in the Pharmaceutical Sciences, Berry,D. (ed), 1990; pp. 509–529, Dekker, New York.

[12] Morris C. and Normand S-L. Hierarchical models for combining information and for meta-analysis. in *Bayesian Statistics 4*, 1992. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds) Oxford University Press, Oxford

[13] Normand, S-L. Meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine* (in press).

[14] Scheidler J., Hricak H., Yu K.K., Subak L. and Segal M.R. "Radiological Evaluation of Lymph Node Metastases in Patients with Cervical Cancer: A Meta-analysis", *JAMA*, 278, 1096-1101 (1997).

[15] Carroll R.J., Ruppert D. and Stefanski L.A. *Measurement Error in Nonlinear Models,* Chapman and Hall, New York, 1995.

[16] McCullagh P. "Regression Models for Ordinal Data", *Journal of the Royal Statistical Society, series B*, 42, 109–142 (1980).

[17] Tosteson A.N.A. and Begg CB." A general regression methodology for ROC curve estimation", *Medical Decision Making*, 8: 204–215 (1988).

[18] Toledano A. and Gatsonis, C.A. " Ordinal regression methodology for ROC curves derived from correlated data". *Statistics in Medicine*, 15:1807-1826 (1996).

[19] Gelfand A.E. and Smith A.F.M. "Sampling–Based Approaches to Calculating Marginal Densities", *JASA*, 85, 398–409 (1990).

[20] Gelman A., Carlin J.B., Stern H.S. and Rubin D.B. *Bayesian Data Analysis,* Chapman and Hall, New York, 1996.

[21] Gelman A. and Rubin D.B. "Inference from Iterative Simulation Using Multiple Sequences", *Statistical Sciences*, 7, 457–511 (1992).

[22] Best N., Cowles M.K. and Vines K. *CODA: Convergence Diagnostics and Output Analysis Software for Gibbs sampling output, Version 0.20,* MRC Biostatistics Unit, Cambridge, 1995.

[23] Weiss R.E. "Bayesian Model Checking with Applications to Hierarchical Models", *Unpublished Technical Report,* August 13, 1996.

[24] Gatsonis, C.A. Random effects models for diagnostic accuracy data. *Academic Radiology* 2:S14-S21, (1995)

[25] Spiegelhalter D., Thomas A., Best N. and Gilks W. "BUGS 0.6, Bayesian inference Using Gibbs Sampling Manual," MRC Biostatistics Unit, 1997.

Figure 1. Detection of Lymph Node Metastases, using lymphangiography (LAG), computed tomography (CT) or magnetic resonance (MR) imaging: Observed true positive (TP) and false positive (FP) rates are from data reported across 37 studies that were originally meta-analyzed by Scheilder and colleagues[14].

Figure 2. Estimated summary receiver operating characteristic curves for lymphangiography (LAG), computed tomography (CT) and magnetic resonance (MR) imaging, based on hierarchical regression modelling.

Table 1. Hierarchical ROC parameter estimates:

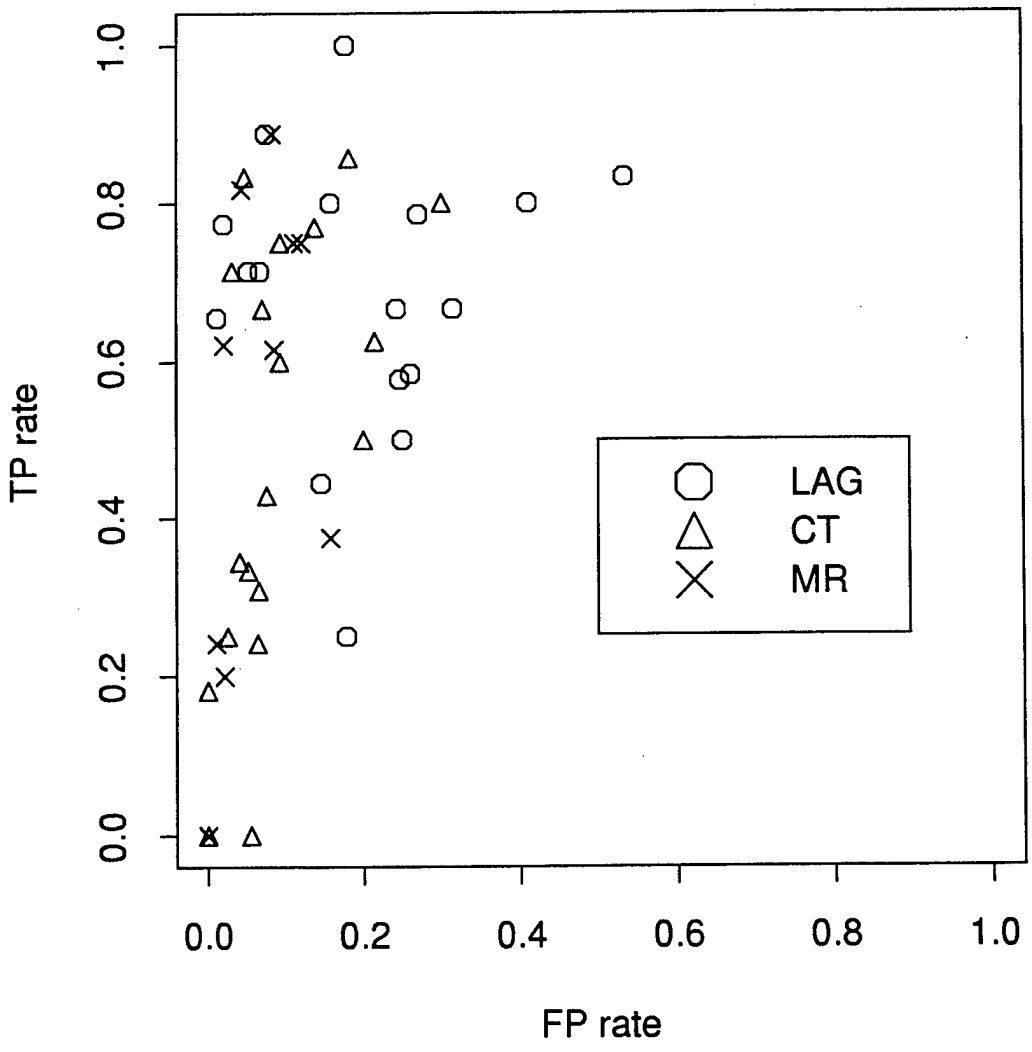
estimated posterior modes with 95% credible intervals in parenthesis.

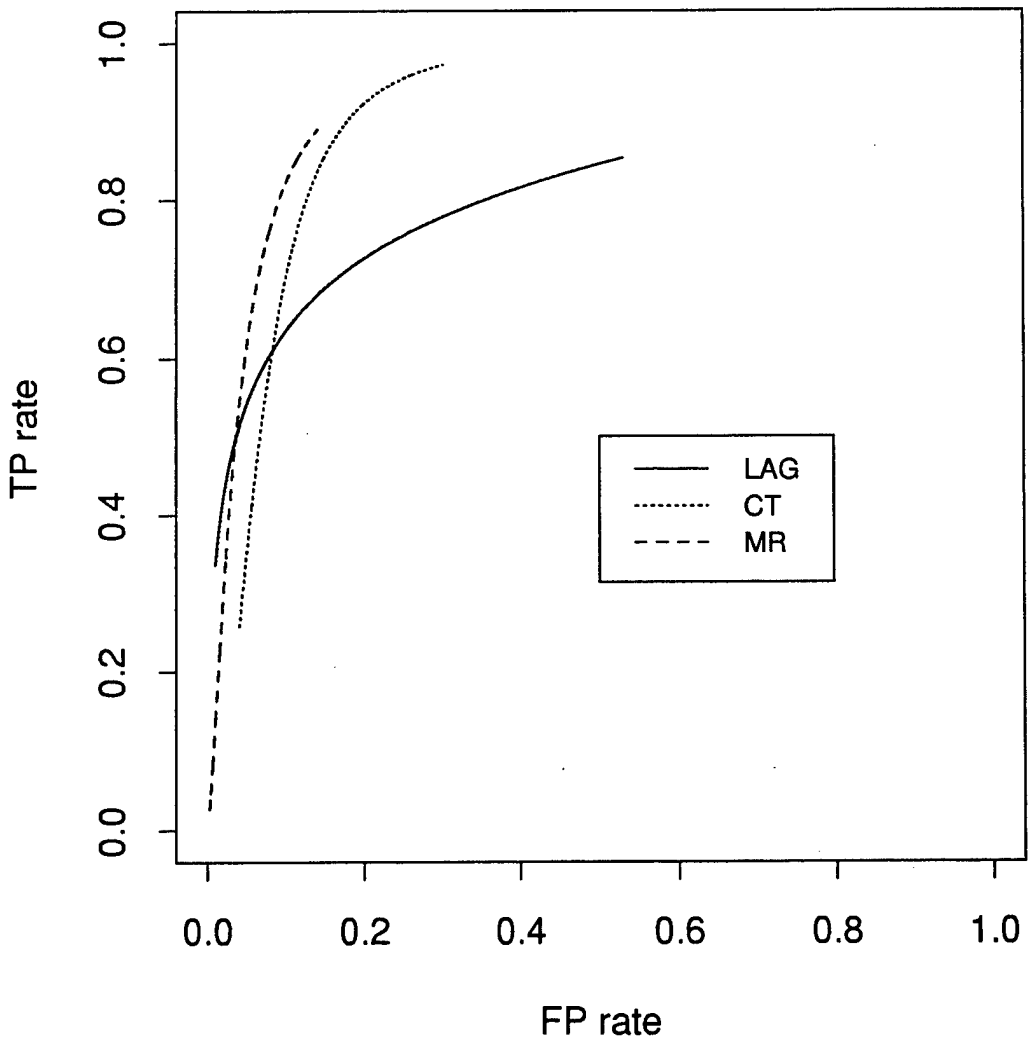
parameter	test type		
	LAG	CT	MR
Θ	-0.178 (-0.783,0.492)	-1.900 (-3.050,-1.090)	-1.800 (-3.210,-0.731)
σ_θ^2	0.248 (0.090,0.587)	0.653 (0.202, 1.570)	1.070 (0.264, 3.160)
Λ	2.360 (1.660,3.120)	3.720 (2.290, 6.030)	3.920 (2.350, 6.330)
σ_α^2	1.080 (0.271,2.780)	0.395 (0.106, 1.150)	0.852 (0.148, 2.900)
β	0.658 (-0.273,1.720)	-0.683 (-1.650, 0.125)	-0.350 (-1.430, 0.587)

Table 2. Overall rates and likelihood ratios:

posterior modes with 95% credible intervals in parenthesis

test type	false positive rate	true positive rate	likelihood ratio negative	likelihood ratio positive
LAG	0.141 (0.080,0.220)	0.666 (0.578,0.750)	0.389 (0.288,0.502)	5.03 (2.96,8.51)
CT	0.070 (0.044,0.102)	0.486 (0.324,0.643)	0.553 (0.390,0.719)	7.15 (4.63,10.60)
MR	0.047 (0.019,0.090)	0.542 (0.298,0.757)	0.480 (0.260,0.729)	12.8 (5.91,24.5)


TP rate
FP rate
0.0
0.2
0.4
0.6
0.8
1.0
LAG
CT
MR


TP rate
0.0
0.2
0.4
0.6
0.8
1.0
FP rate
0.0
0.2
0.4
0.6
0.8
1.0
LAG
CT
MR

Appendix D

Assessing Mammographer Accuracy:
A comparison of clinical and test performance

CM Rutter and S Taplin

submitted to *Journal of Clinical Epidemiology*

Funding number: DAMD17-97-1-7193

June 7, 1999

Assessing Mammographers' Accuracy:
A comparison of clinical and test performance

Carolyn M. Rutter and Stephen Taplin

Group Health Cooperative of Puget Sound, Center for Health Studies

This study was supported, in part, by grants CA63731 from the National Cancer Institute and BC962461 from the U.S. Department of Defense.

We wish to acknowledge the careful work of Kari Rosvik and Deb Seger who made this study possible, and the many mammographers who gave their time to this study. We want to especially thank Mary Kelly, MD and Donna White, MD who provided valuable leadership.

Address correspondence to:

Carolyn Rutter

Group Health Cooperative, Center for Health Studies

1730 Minor Avenue, Suite 1600

Seattle, WA 98101

email: rutter.c@ghc.org

phone: 206.287.2190

fax: 206.287.2871

Assessing Mammographers' Accuracy:

A comparison of clinical and test performance

Abstract

Direct estimation of mammographers' clinical accuracy requires the ability to capture screening assessments and correctly identify which screened women have breast cancer. This clinical information is often unavailable and when it is available its observational nature can cause analytic problems. Problems with clinical data have led some researchers to evaluate mammographers using a single set of films. Research based on these test film sets implicitly assumes a correspondence between mammographers' accuracy in the test setting and their accuracy in a clinical setting. However, there is no evidence supporting this basic assumption. In this article we use hierarchical models and data from 27 mammographers to directly compare accuracy estimated from clinical practice data to accuracy estimated from a test film set. We found no evidence of correlation between clinical and test accuracy. These findings raise important questions about how mammographer accuracy should be measured.

keywords: sensitivity, specificity, hierarchical models, mammography.

running title: Assessing Mammographers' Accuracy: clinical versus test performance

1. Introduction

Screening mammography is an effective method of detecting early stage breast cancer. However, the diagnostic value of a mammogram depends on both the technical quality of the film and a mammographer's ability to interpret that film. In the last decade mammographic technology has been relatively stable, allowing researchers to focus on the subjective interpretation of mammograms (e.g.,[1, 2]).

The Mammography Quality Standards Act recognized the effect of mammographers' interpretations on screening assessments and encouraged medical audits of mammographers' clinical assessments. Evaluating mammographers' performance using clinical assessments is intuitively appealing, because this is 'real life' performance. For many researchers, the medical audit is the gold standard measure of performance.[3] However, our ability to draw conclusions about the performance of particular mammographers from these clinical assessments is limited because each mammographer reviews a different set of films. The difficulty of films varies with characteristics of the women evaluated (e.g., breast density), characteristics of lesions (e.g., size), and characteristics of technical film quality (e.g., positioning). Variability in film difficulty results in chance differences among mammographers. Systematic differences in the difficulty of films reviewed can also occur, for example, when mammographers tend to send difficult cases to a particular colleague. Differences in the number of films reviewed also affects comparisons between mammographers through the variability of estimated performance. Because performance estimates based on fewer patients tend to be more variable, and therefore more extreme, comparisons that ignore differences in variability can be misleading. Statistical models have a limited ability to adjust for differences in the films read by each mammographer.[4, 5]

Estimation of clinical accuracy is further complicated by the influence that clinical assessments have on the probability of detecting breast cancer. Screening accuracy estimation focuses on the correspondence between a mammographer's clinical interpretation and a woman's true disease state. Because most women only undergo biopsy if a mammographer finds an abnormality, undetected breast cancer cases emerge symptomatically or during a second screening exam. Thus, undetected breast cancer can only be identified when follow-up information exists. A one year follow-up is generally used, with women classified as disease positive at the time of a screening

mammogram if breast cancer is diagnosed within one year.[3]

Estimation and comparison of clinical screening performance is also hampered by the relatively low incidence of breast cancer. The one year incidence of invasive breast cancer is approximately 3.5 per 1,000 among American women who are over 49 years old.[6] Low incidence rates make it difficult to precisely estimate a mammographer's rate of cancer detection, since most mammographers will evaluate very few cancers in a single year.

Standardized testing of mammographers is an alternative way to estimate their accuracy. Using standardized film sets removes many of the problems with clinical data. Each mammographer views the same films in the same setting and with the same patient information. Test sets exclude films from women without necessary follow-up information, so that true disease state is known with a high degree of certainty. Test sets can also include more films from women with breast cancer than would be seen in clinical practice, allowing more precise estimation of sensitivity. In summary, use of a test film set controls for film difficulty, film quality and the information presented during film evaluation, offering a relatively simple method of estimating mammographers' accuracy under standardized conditions.

Although estimating accuracy from assessments of standardized film sets avoids many of the problems with clinical data, the artificial conditions introduce other problems. Mammographers know that in the test setting their decisions will not affect patient care. The test itself may be burdensome given time constraints. There is also evidence suggesting that the higher prevalence of disease in test film sets introduces bias. Egglin[7] found that radiologists were more likely to interpret arteriograms as positive for pulmonary emboli when viewed in a higher prevalence film set, regardless of true disease state. When this 'context bias' exists, sensitivity increases with increasing prevalence while specificity decreases.

Studies describing mammographer variability based on test film sets (e.g.,[1, 2]) implicitly assume a strong correlation between mammographers' performance estimated from test sets and mammographers' performance in clinical practice. However, this assumption has never been tested. In this article we directly compare mammographers' clinical and test performance.

2. Data

We analyzed data from 27 mammographers practicing at a large staff model not-for-profit health

maintenance organization (HMO). The mammographers included in this study were voluntary participants, though this group essentially included all of the mammographers practicing with the HMO at the time of the study.

Both clinical and test data sets use films from women who remained enrolled in the HMO for at least two years after their index mammogram. Women with breast cancer were identified using the regional Surveillance Epidemiology and End Result registry.[8] Our reference standard for true disease state called a woman 'disease positive' at the time of her screening mammogram if either invasive cancer or ductal carcinoma in situ were detected within the following two years. We used a two year definition because routine follow-up care included mammographic follow-up at either one year or two year intervals, depending on a woman's particular risk factors for breast cancer.

Clinical Data: Clinical data used mammographers' final interpretations and recommendations based on mammograms from asymptomatic women screened from 1990 through 1994. Mammographers interpretations and recommendations have been collected as part of clinical practice for every mammogram evaluated since 1986, using standardized data collection forms. During the time period we examined, mammographer interpretations could be coded as 'negative', 'inconclusive', or 'positive'. Final interpretations and recommendations were combined and coded into one of five possible clinical assessments: 1) negative mammogram and recommendation for mammographic follow up at 1 year or later; 2) inconclusive mammogram and recommendation for mammographic follow up at 1 year or later; 3) inconclusive mammogram and recommendation for follow up in less than 1 year (short interval follow-up); 4) inconclusive mammogram and recommendation for biopsy or surgical referral; and 5) positive mammogram.

Test Data: Mammographers were evaluated using test film sets during late 1994 and early 1995. As part of an educational intervention, each mammographer assessed the same set of screening mammograms. Test mammograms were drawn from the population of women screened between 1985 and 1991, using stratified random sampling. Most (92.5%, 111/120) films were selected from the 1990/1991 time period. Films were stratified by the woman's true disease state and the original (clinical) mammographer's assessment. We defined recommendations for short interval follow-up, request for additional work-up, referral to biopsy, and positive mammo-

gram interpretations as positive mammographic assessments, corresponding to clinical assessment categories 3, 4 and 5. Based on each screened woman's true state and dichotomous clinical assessment, we created four strata: true positive (TP) films (positive assessment, breast cancer within one year); false negative (FN) films (negative assessment, breast cancer within one year); true negative (TN) films (negative assessment, no breast cancer); and false positive (FP) films (positive assessment, no breast cancer). From these strata, we randomly selected 23 TP films, 9 FN films, 72 TN films, and 16 FP films. Because of the stratified sampling scheme, the test film set was not representative of the mix of films seen in clinical practice: it included an excess of films from women with breast cancer and films that originally lead to incorrect assessments. Out of these 120 films, 7 films (3 TP films and 4 FP films) were excluded from analyses because marks were placed on films during the course of the study. To allow correspondence with the clinical analyses, the reference standard for test films was recalculated, using a 2 year follow-up period. Applying the 2 year follow-up caused one TN film to be recoded as a FN. Within the 113 test mammograms used for analyses, original readers were 67% sensitive and 86% specific. The average age of screened women who contributed films to the test set was 50 years, ranging from 40 to 87 years.

Mammograms were displayed at each participating mammography clinic in a dedicated reading room. Films were displayed in four sets of 30, and each set was displayed for 2 weeks. Mammographers scheduled a time to review films and were given 1 hour to read each set of 30 films. Each 'film' included a two-view mammogram, representing a single screening event, and the woman's most recent prior two-view screening mammogram. Prior mammograms were unavailable for 43 women (38%). No additional clinical information was provided, and mammographers were not provided with the disease prevalence in the test set. Mammographers provided one rating for each breast, using standardized data collection forms. The 5 possible screening assessments were: 1) negative or benign; 2) probably benign (short interval follow-up needed); 3) possibly abnormal (additional views needed); 4) suspicious abnormality (biopsy should be considered); and 5) highly suggestive of malignancy. Each mammographer provided data that was at least 98% complete (222/226 ratings) and 15 of the 27 mammographers provided complete data. There were no apparent patterns of missing data between mammographers. These breast-level ratings

were recoded as woman-level assessments. If the woman was diagnosed with breast cancer within two years of the mammogram, then the rating given to the breast with disease was used in the analyses. If the woman did not develop cancer in the following two years, then the maximum of the two breast ratings was used.

3. Methods

We are primarily interested in the degree of correlation between mammographers' accuracy measured in a clinical setting and accuracy measured in a test setting. The accuracy measures we focused on are sensitivity and specificity. Sensitivity is the proportion of women with breast cancer who had a positive mammogram assessment. Specificity is the proportion of women without breast cancer who had a negative mammogram assessment.

Calculation of sensitivity and specificity requires definition of a positive assessment. For clinical assessments, we defined ratings 3, 4 and 5 as positive mammograms, corresponding to recommendations for short interval follow-up or biopsy. Unfortunately, test assessments do not completely match clinical assessments. This is partly because clinical assessments were based on final recommendations whereas the test scale included a recommendation for additional views. Clinical data did not include recommendations for additional views because this is an intermediate clinical recommendation, with final recommendations based on these additional views. Given the difference in these two measurement scales, we defined positive outcome in the test set as a recommendation for short interval follow-up, additional views, or biopsy in the test data set, corresponding to ratings 2, 3, 4, or 5. Mammographers' ratings of test films were based on an explicitly ordinal scale that defined a recommendation for additional films (possibly abnormal) as more strongly indicative of disease than a recommendation for short interval follow-up (probably benign).

3.1 Statistical Model

We used a hierarchical model to describe mammographers' test and clinical performance measures, and to examine relationships between these measures. Each mammographer contributed data from two 2×2 tables, showing the overall agreement between their assessments and womens' disease state. We use the following notation:

		Mammographic Interpretation: negative	positive	
Breast Cancer:	no	y_{ij00}	y_{ij01}	n_{ij0}
	yes	y_{ij10}	y_{ij11}	n_{ij1}

Where $i = 1, \ldots, m$, indicates mammographer and $j = 1, 2$ indicates the data source (1=test and 2=clinical).

The model we use accounts for within mammographer variability in estimated sensitivity and specificity by modeling the number of positive assessments each mammographer gave to diseased (y_{ij11}) and not-diseased (y_{ij01}) women with Binomial(n_{ij1},π_{ij1}) and Binomial(n_{ij0},π_{ij0}) distributions. By using the observed sample sizes in Binomial distributions for each mammographer and data set, the model accounts for differences in the amount of data available. The binomial probability of a positive test is based on receiver operating characteristic models,[9] and is given by :

$$\pi_{ijk} = \text{logit}^{-1}(\theta_{ij} + \alpha_{ij} D_{ijk})$$

If D_{ijk} was coded 0 for disease negative films and 1 for disease positive films, then under this model the ith mammographer evaluates the jth data set with specificity equal to $1 - \text{logit}^{-1}(\theta_{ij})$ and sensitivity equal to $\text{logit}^{-1}(\theta_{ij} + \alpha_{ij})$. It is simpler to explain the interpretation of θ_{ij} and α_{ii} in terms of false positive rates (equal to 1 - specificity) and true positive rates (equal to the sensitivity). The parameter θ_{ij} captures the ith mammographer's overall tendency to give positive assessments, so that true positive rates increase with increasing false positive rates. The parameter α_{ij} captures the difference between true positive and false positive rates and measures the log-odds ratio of a positive test for films with breast cancer relative to films without breast cancer. As in the ROC context, we call θ_{ij} "cutpoint parameters" and α_{ij} "accuracy parameters".

The parameters θ_{ij} and α_{ij} could be calculated directly from the data. However, they are not estimable when either sensitivity or specificity is 100%, a situation that is more likely when a mammographer evaluates few films. The hierarchical model uses all available information to better estimate these individual parameters. Under the hierarchical model, both cutpoint parameters (θ_{ij}) and accuracy parameters (α_{ij}) are assumed to vary across mammographers and

data sources. We assume θ_{ij} and α_{ij} follow a bivariate normal distribution, implemented as:

$$\left.\begin{array}{rcl} \theta_{i1}|\Theta_1, \sigma_{\theta 1} & \sim & N(\Theta_1, \sigma_{\theta 1}^2) \\ \alpha_{i1}|\Lambda_1, \sigma_{\alpha 1} & \sim & N(\Lambda_1, \sigma_{\alpha 1}^2) \end{array}\right\} \text{conditionally independent}$$

and

$$\left.\begin{array}{rcl} \theta_{i2}|\theta_{11}, \theta_{21}, \ldots, \theta_{m1}, \Theta_2, \tau, \sigma_{\theta 2} & \sim & N(\Theta_2 + \tau(\theta_{i1} - \frac{1}{m}\sum_{i=1}^m \theta_{i1}), \sigma_{\theta 2}^2) \\ \alpha_{i2}|\alpha_{11}, \alpha_{21}, \ldots, \alpha_{m1}, \Lambda_2, \lambda, \sigma_{\alpha 2} & \sim & N(\Lambda_2 + \lambda(\alpha_{i1} - \frac{1}{m}\sum_{i=1}^m \alpha_{i1}), \sigma_{\alpha 2}^2) \end{array}\right\} \text{conditionally independent}$$

Thus, the model assumes that within each data set, mammographers' cutpoint and accuracy parameters are (conditionally) independent. The linear regression models for θ_{i2} and α_{i2} build in correlation between cutpoint parameters and correlation between accuracy parameters, with:

$$\begin{array}{rcccl} \text{corr}(\theta_{i1}, \theta_{i2}) & = & \rho_\theta & = & \dfrac{\tau\sigma_{\theta 1}}{\sqrt{\tau^2\sigma_{\theta 1}^2(\frac{m-1}{m}) + \sigma_{\theta 2}^2}} \\ \text{corr}(\alpha_{i1}, \alpha_{i2}) & = & \rho_\alpha & = & \dfrac{\lambda\sigma_{\alpha 1}}{\sqrt{\lambda^2\sigma_{\alpha 1}^2(\frac{m-1}{m}) + \sigma_{\alpha 2}^2}} \end{array}$$

These correlation parameters are more informative than the between dataset correlation of sensitivity or specificity. Correlation in sensitivity and specificity can be driven by mammographers' overall tendency to provide positive calls. The correlation parameters ρ_θ and ρ_α separate the overall tendency to call a film positive from the ability to distinguish between films from women with and without breast cancer. Under this model, ρ_θ measures the correlation between cutpoint parameters that are associated with overall preponderance to call a film 'positive' while ρ_α measures association between accuracy parameters that are independent of these cutpoint parameters.

Because the regression model is centered, the expected value of θ_{i2} is Θ_2 and the expected value of α_{i2} is Λ_2. Assuming that θ_{ij} and α_{ij} are normally distributed and linked via a regression model allows fuller use of the available data, resulting in better estimation. Mammographer's cutpoint and accuracy parameters are smoothed toward overall expected values Θ_j and Λ_j, with the degree of smoothing determined by the amount of data each contributes to the model. Estimates for mammographers with less data will tend to be nearer to expected values than

estimates for mammographers with more data, while corresponding interval estimates widen to reflect lack of information available for these parameters.

The hierarchical model is completed by specifying prior distributions for the remaining unknown parameters. Priors were chosen to cover the range of plausible values of parameters and were selected to be uninformative. We used a Normal(0,10) prior for Θ_1, Θ_2, Λ_1, and Λ_2, and a Normal(0,100) prior for τ and λ. We used an inverse gamma, $\Gamma^{-1}(0.5, 2)$, for $\sigma_{\theta 1}$, $\sigma_{\theta 2}$, $\sigma_{\alpha 1}$ and $\sigma_{\alpha 2}$. This prior is diffuse, but does not overweight large values. Quartiles of the $\Gamma^{-1}(0.5, 2)$ distribution are 3.03, 8.80, and 39.41. The parameters Θ_1, Θ_2, Λ_1, Λ_2, τ, λ, $\sigma_{\theta 1}$, $\sigma_{\theta 2}$, $\sigma_{\alpha 1}$ and $\sigma_{\alpha 2}$ are assumed to be mutually independent.

This model was estimated using the BUGS program.[10] To improve estimation, the disease state indicator D_{ijk} was centered so that $D_{ijk} = \frac{1}{2}$ for disease positive films and $D_{ijk} = -\frac{1}{2}$ for disease negative films. This transformation does not affect the interpretation of the parameters α_{ijk} and θ_{ijk}. Standard model diagnostics were used to assess convergence of the sampler, as described in the CODA manual.[11] These models resulted in estimated posterior distributions for the model parameters. We present estimated posterior modes and 95% credible intervals based on the 2.5% and 97.5% percentiles. The posterior mode was estimated by the posterior mean for approximately symmetric distributions, and by the posterior median for skewed posterior distributions.

4. Results

There was wide variability in the amount of clinical data available for each mammographer (Table 1). The 27 mammographers clinically evaluated an average of 1890 films during the four year period (range 232 to 3818), and saw an average of 15 mammograms from women with breast cancer (range 1 to 32). The average clinical prevalence rate across mammographers was 8 cancers per 1,000 mammograms.

Plots of the sensitivity and specificity suggest moderate positive correlation between clinical and test performance. Figure 1 shows that overall, mammographers tended to be both more sensitive and more specific in clinical practice. The observed correlation between clinical and test sensitivities was -0.096; correlation between specificities was 0.446.

[Table 1 about here]

[Figure about here]

The hierarchical model accounts for within mammographer variability in sensitivity and specificity and accounts for differences in the number of films read in clinical practice. The model can be used to better estimate each mammographers' clinical and test-based sensitivity and specificity, and thus to better estimate between dataset correlation in sensitivity and specificity. Model based estimates of sensitivity and specificity combine information from the entire sample with each mammographer's information. The degree to which estimates differ from observed values reflects the amount of data available, the values of other parameter estimates (i.e., $\hat{\theta}_{i1}$, $\hat{\theta}_{i2}$, $\hat{\alpha}_{i1}$, $\hat{\alpha}_{i2}$, $\hat{\tau}$ and $\hat{\lambda}$) and underlying distributional assumptions. Estimates of clinical specificity were equal to model estimates because these were based on large numbers of films. In contrast, estimates of clinical sensitivity were more strongly influenced by additional information, especially for mammographers who evaluated very few films. Model-based estimates of between dataset correlation of sensitivity and specificity were similar to observed correlation estimates. Correlation between clinical and test sensitivity was 0.185 with 95% credible interval (-0.269,0.593). Correlation between clinical and test specificity was 0.408 with 95% credible interval (0.161,0.616).

We found little evidence of correlation between clinical and test performance parameters (Table 2). Our point estimate of correlation between clinical and test cutpoints was moderate ($\rho_\theta = 0.220$) although the 95% credible interval was broad and covered zero. The estimated probability that $\rho_\theta > 0$ was 89.3%. Our point estimate of the correlation between clinical and test accuracies was near zero ($\rho_\alpha = -0.026$).

We found expected overall differences in test and clinical accuracy. The test film set was constructed to be more difficult than films seen in usual clinical practice, and as expected the estimated mean clinical accuracy parameter (Λ_2) was greater than the estimated mean test accuracy parameter (Λ_1), indicating that overall readers were more accurate when evaluating clinical data than test data.

Point estimates also demonstrated that mammographers had an overall tendency to give more positive assessments in their clinical practice than in the test setting (mode Θ_2 > mode Θ_1), even though the prevalence of breast cancer was much higher in the test setting.

Estimated between mammographer variability tended to be higher in clinical practice than in the test setting (e.g., $\sigma^2_{\theta 2} > \sigma^2_{\theta 1}$ and $\sigma^2_{\alpha 2} > \sigma^2_{\alpha 1}$), possibly reflecting wider variability in the films each mammographer reads in clinical practice, or the relatively small number of cancer films each mammographer evaluated over the course of four years in clinical practice.

[Table 2 about here]

5. Discussion

These results represent a comprehensive comparison of mammographers' assessments in test and clinical settings. The clinical data was based on automated collection of mammographers' interpretations and recommendations. The data systems also allowed two year follow-up of each woman screened. The test data included a relatively large set of 113 mammograms and included 30 cancers. Finally, our statistical model allowed for differences in the number of films each mammographer assessed during clinical practice.

There was general agreement between observed values and hierarchical model results. Mammographers tended to be less accurate when evaluating the more difficult test film set, and tended to give more positive assessments in their clinical practice. Thus, we found no evidence of context bias as described by Egglin[7]. That is, mammographers did not tend to make more positive assessments in the higher prevalence test film set. However, we cannot conclude from this study that context bias does not exist, because the test context included both a higher disease prevalence and a more difficult set of films.

Model-based estimates of between dataset correlation of sensitivity were stronger than observed correlation, and the estimated between dataset correlation of specificity was statistically different from zero. However, between dataset correlation of sensitivity and specificity appears to be driven by correlation in the mammographers tendency to call tests positive rather than corre-

lation in their accuracy evaluating the two data sets. We found moderate, but not statistically significant, correlation between mammographers' overall preponderance to identify cancer using the two data sources. But there was no apparent correlation between the hierarchical model's accuracy parameters.

We do not believe that the lack of correlation between clinical and test performance resulted from differences in outcome scales. The basic assumption that we are testing is that these two measures are correlated because both are measures of the same underlying construct, mammographer accuracy. We are not interested in the equality of these two measures; we expect these accuracy estimates to differ because of differences in film difficulty, film quality, and the information provided to mammographers.

We do not believe that the lack of correlation between clinical and test performance resulted from dichotomizing the outcome scales. We did not attempt to model the ordinal outcomes directly or via the area under the receiver operating characteristic (ROC) curve because in both clinical and test settings mammographers' maximum false positive rates were relatively low. Because of this, the area under their ROC curves were strongly influenced by false positive rates that were outside of the observed data range especially for clinical data. The sensitivity and specificity pairs we used in analyses contained most of the information available from ROC curves.

There are many possible explanations for the lack of correlation in these data. Our 'gold standard' for true disease state was based on a two-year follow-up interval, and misclassification of diseased and not diseased women may have attenuated observed correlation. Our sample of 27 mammographers may have been too small to detect statistically significant correlation, although point estimates suggest there was not clinically relevant correlation in accuracies. Examining mammographers practicing within the same HMO may have reduced variability so that correlation was not observable. Many of the mammographers in this study worked together and discussed difficult cases with each other on a day-to-day basis. Finally, lack of correlation may have resulted from differences in the type of films included in the two data sets. Clinical data included assessments of exams based on imaging studies, such as ultrasound and magnification views. If evaluation of 2 view mammograms requires different skills than evaluation of additional work-

up images, then the inclusion of these films in the clinical set could attenuate the estimated correlation between clinical and test accuracy. However, excluding these films would drastically reduce the number of cancer cases included in the clinical set and could bias comparisons by reducing the clinical data set to films that the original reader was able to assess without additional work-up. Because these results were unexpected, we must consideration possible explanations. However, these explanations are ultimately conjecture.

The apparent lack of correlation between test and clinical assessments could be interpreted in at least two ways. One interpretation is that evaluations based on clinical assessments and evaluations based on test film sets are measuring two different kinds of accuracy. Because we are interested in clinical performance, concluding that test-based assessments of accuracy are different from clinical accuracy means either throwing out the test data sets as a reasonable means of mammographer evaluation, or seeking out ways to make test evaluations more comparable to clinical evaluations. A second interpretation is that the apparent lack of correlation between clinical and test performance resulted from differences in the clinical case mix of participating mammographers. Clinical data included assessments based on both standard screening mammograms and screening mammograms that included additional work up, such as ultrasound and magnification views. We do not know how these different types of films were distributed across mammographers, or whether there were any informal systems of referral at the mammography centers. Systematic differences between mammographers could also have been introduced through differences in screened populations, for example, differences in the average age of women screened. Concluding that the clinical data are problematic means either throwing out the clinical data as a means of mammographer evaluation, or seeking out ways to make the clinical evaluations more comparable across mammographers. Unfortunately, our analyses cannot guide our conclusions about clinical and test data, though they caution us against extrapolating results from one setting into another.

References

[1] Beam CA, Layde PM, Sullivan DC. Variability in the Interpretation of Screening Mammograms by US Radiologist, **Archives of Internal Medicine,** 1996; 156: 209–213.

[2] Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms, **New England Journal of Medicine,** 1994; 331:1493-1499.

[3] Linver MN, Osuch JR, Brenner RJ, Smith RA. The Mammography Audit: A Primer for the Mammography Quality Standards Act (MQSA), **American Journal of Radiology,** 1995; 165:19–25.

[4] Christiansen CL, Morris CN. Improving the Statistical Approach to Health Care Provider Profiling, **Annals of Internal Medicine,** 1997; 127:764–8.

[5] Delong ER, Peterson ED, DeLong DM, Muhlbaire LH, Hackett S, Mark DB. Comparing Risk-Adjustment Methods for Provider Profiling, **Statistics in Medicine,** 1997; 16:2645–2664.

[6] Reis LAG, Kosary CL, Hankey BF, Miller BA, Edwards BK (eds). **SEER Cancer Statistics Review, 1973–1995.** Bethesda, MD: National Cancer Institute; 1998.

[7] Egglin TK, Feinstein AR. Context bias. A problem in diagnostic radiology, **Journal of the American Medical Association,** 1996; 276:1752–1755.

[8] Miller BA, Reis LAG, Hankey BF. **SEER Cancer Statistics Review 1973-1990.** NIH Publication No. 93-2789; Bethesda, MD: National Cancer Institute; 1993.

[9] Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art, **Critical Reviews in Diagnostic Imaging,** 1989; 29:207-35.

[10] Spiegelhalter D, Thomas A, Best N, Gilks W. **BUGS 0.6, Bayesian inference Using Gibbs Sampling Manual,** MRC Biostatistics Unit, 1997.

[11] Best N, Cowles MK, Vines K. **CODA: Convergence Diagnostics and Output Analysis Software for Gibbs sampling output, Version 0.20,** Cambridge: MRC Biostatistics Unit; 1995.

Table 1. Mammographic Assessments of 27 Mammographers: Rate of positive assessments, indicating disease, with the total number of assessments in parenthesis.

	Test Data				Clinical Data			
	specificity		sensitivity		specificity		sensitivity	
mammographer	(N)		(N)		(N)		(N)	
1	0.880	(83)	0.897	(29)	0.922	(1715)	1.000	(14)
2	0.687	(83)	0.833	(30)	0.816	(1492)	1.000	(14)
3	0.687	(83)	0.833	(30)	0.804	(2341)	0.929	(14)
4	0.880	(83)	0.833	(30)	0.823	(2129)	0.933	(15)
5	0.867	(83)	0.800	(30)	0.896	(2818)	0.880	(25)
6	0.756	(82)	0.733	(30)	0.917	(2221)	0.941	(17)
7	0.867	(83)	0.767	(30)	0.965	(1733)	0.684	(19)
8	0.904	(83)	0.700	(30)	0.911	(2045)	0.917	(12)
9	0.867	(83)	0.833	(30)	0.879	(1742)	0.826	(23)
10	0.831	(83)	0.800	(30)	0.832	(1435)	0.833	(12)
11	0.867	(83)	0.800	(30)	0.915	(3299)	0.935	(31)
12	0.831	(83)	0.724	(29)	0.865	(230)	1.000	(2)
13	0.867	(83)	0.833	(30)	0.870	(971)	0.800	(10)
14	0.904	(83)	0.867	(30)	0.877	(675)	0.500	(2)
15	0.783	(83)	0.833	(30)	0.881	(2546)	0.955	(22)
16	0.880	(83)	0.800	(30)	0.930	(441)	1.000	(1)
17	0.855	(83)	0.867	(30)	0.883	(3167)	0.960	(25)
18	0.854	(82)	0.867	(30)	0.822	(1451)	1.000	(11)
19	0.771	(83)	0.833	(30)	0.901	(3786)	0.875	(32)
20	0.904	(83)	0.767	(30)	0.905	(1276)	0.714	(7)
21	0.855	(83)	0.833	(30)	0.908	(3186)	0.800	(25)
22	0.904	(83)	0.733	(30)	0.880	(2585)	0.947	(19)
23	0.855	(83)	0.793	(29)	0.943	(1643)	0.846	(13)
24	0.807	(83)	0.828	(29)	0.913	(1726)	1.000	(10)
25	0.819	(83)	0.767	(30)	0.864	(1151)	1.000	(4)
26	0.892	(83)	0.900	(30)	0.920	(2169)	0.842	(19)
27	0.759	(83)	0.833	(30)	0.867	(663)	0.833	(6)

Table 2. Hierarchical model estimates from the posterior distribution.

parameter and description	Estimates mode	95% credible region
Θ_1: expected cutpoint parameter, test data	-0.100	(-0.330, 0.125)
Λ_1: expected accuracy parameter, test data	3.322	(2.888, 3.553)
Θ_2: expected cutpoint parameter, clinical data	0.066	(-0.216, 0.352)
Λ_2: expected accuracy parameter, clinical data	4.361	(3.928, 4.798)
$\sigma^2_{\theta 1}$: between-mammographer variance of cutpoints, test data	0.261	(0.153, 0.489)
$\sigma^2_{\alpha 1}$: between-mammographer variance of accuracy, test data	0.409	(0.215, 0.823)
$\sigma^2_{\theta 2}$: between-mammographer variance of cutpoints, clinical data	0.337	(0.190, 0.658)
$\sigma^2_{\alpha 2}$: between-mammographer variance of accuracy, clinical data	0.502	(0.247, 1.096)
τ: regression coefficient, cutpoint parameters	0.560	(-0.341, 1.530)
λ: regression coefficient, accuracy parameters	-0.048	(-1.020, 0.945)
ρ_θ: correlation between clinical and test cutpoints	0.220	(-0.142, 0.486)
ρ_α: correlation between clinical and test accuracy	-0.026	(-0.477, 0.446)

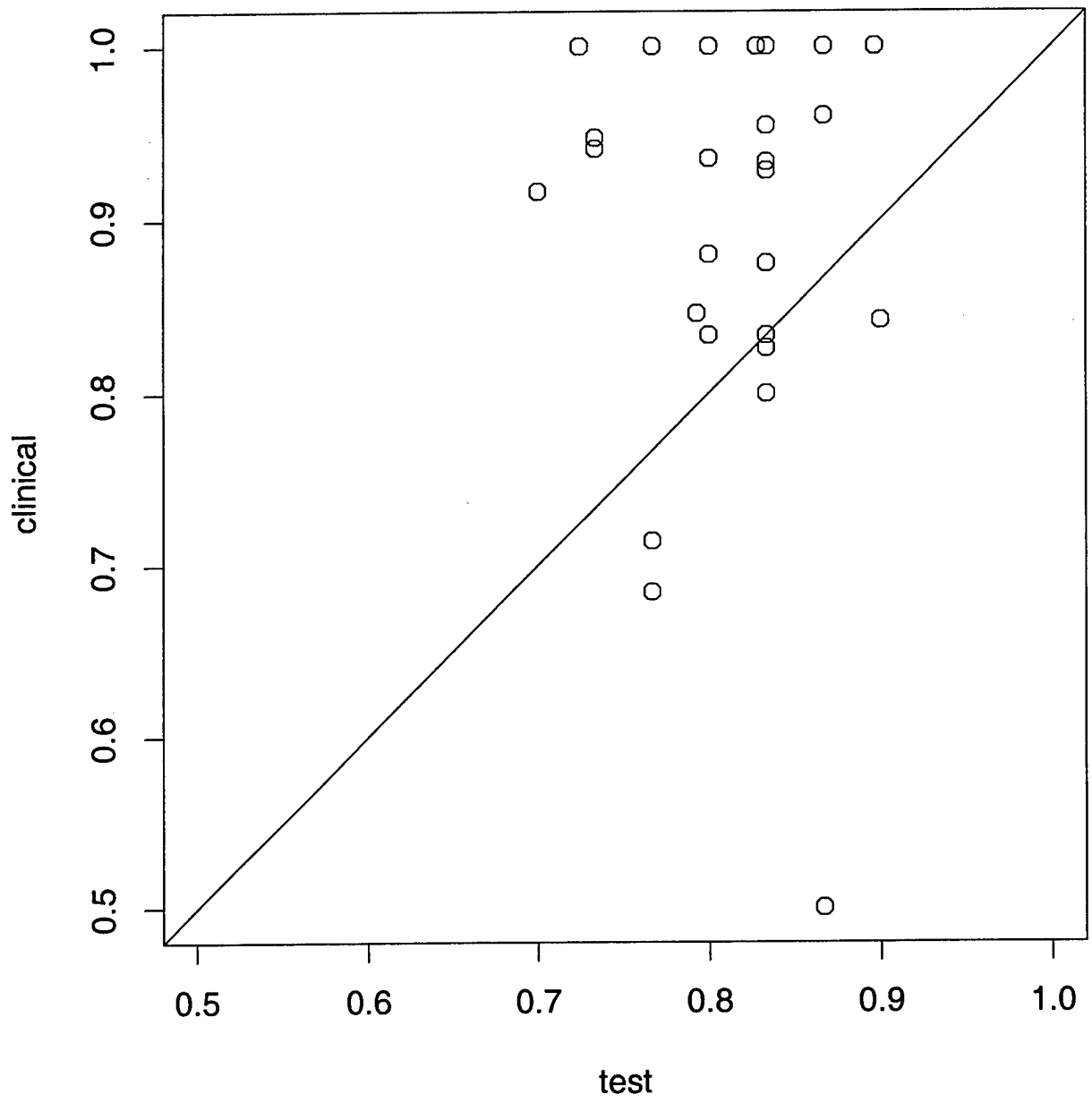


Figure 1A. Sensitivity in clinical practice versus sensitivity in a test setting for 27 mammographers.

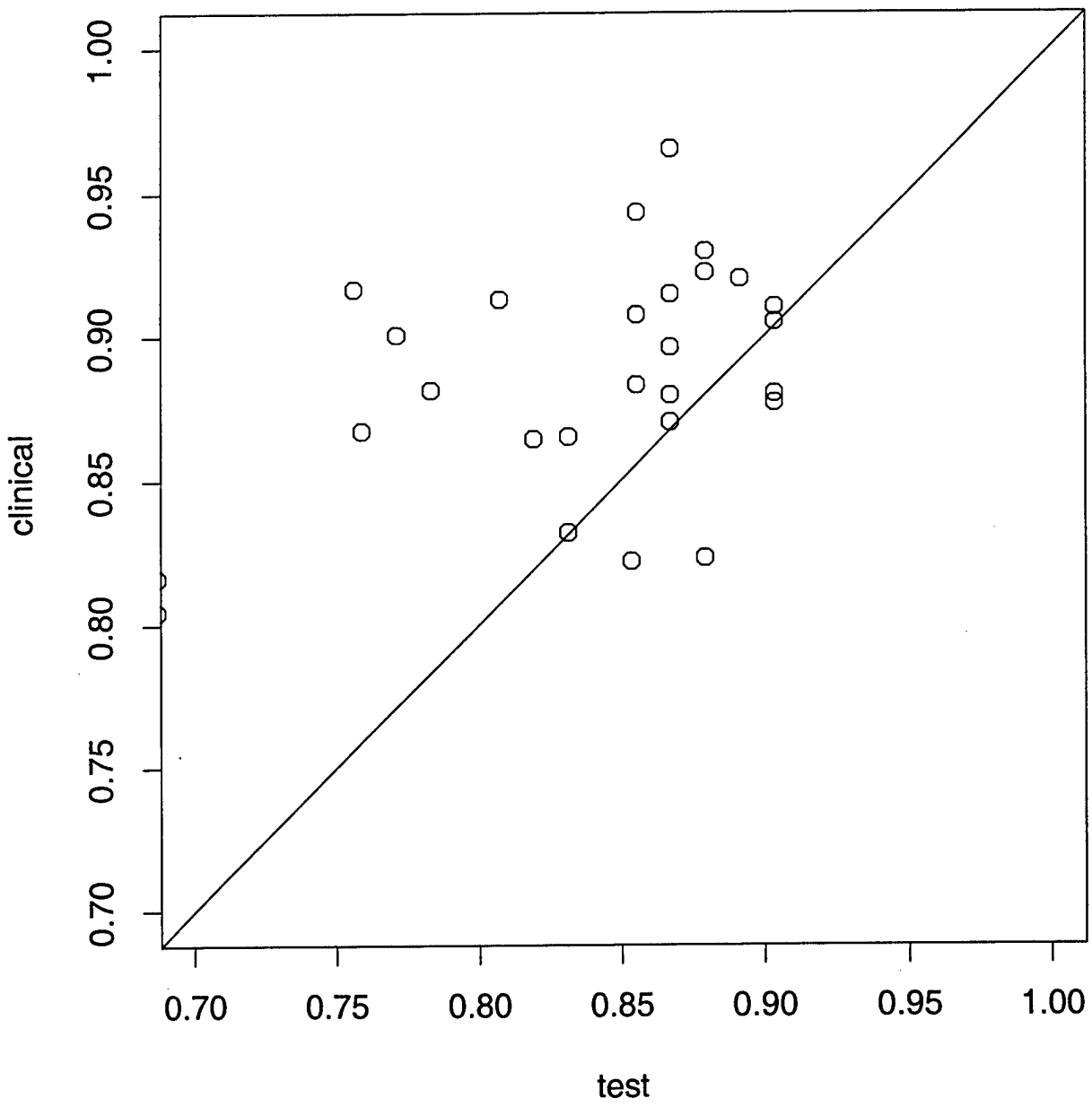


Figure 1B. Specificity in clinical practice versus sensitivity in a test setting for 27 mammographers.

Appendix E

Design of a Study
to Improve Accuracy in Reading Mammograms

MS Pepe, N Urban, C Rutter, G Longton

Journal of Clinical Epidemiology 50: 1327-1338, 1997

J Clin Epidemiol Vol. 50, No. 12, pp. 1327–1338, 1997
Copyright © 1997 Elsevier Science Inc.

ELSEVIER

0895-4356/97/$17.00
PII S0895-4356(97)00204-7

Design of a Study to Improve Accuracy in Reading Mammograms

Margaret Sullivan Pepe,[1,*] *Nicole Urban,*[1] *Carolyn Rutter,*[2] *and Gary Longton*[1]

[1]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington; and [2]Center For Health Studies, Group Health Cooperative, Seattle, Washington

ABSTRACT. This paper is concerned with the design and analysis of mammography reading studies. In particular we consider studies aimed at evaluating interventions to improve the accuracy with which mammograms are read. A simple randomized design is suggested in which a relatively large group of readers read sets of mammograms before and after an intervention phase. We propose solutions to three difficult statistical issues that arise in the context of such studies: (i) the choice of primary outcome measure; (ii) the data analysis technique to be employed; and (iii) the methodology for calculating sample sizes for readers and images to be read.

First, we argue in favor of using sensitivity and specificity as the primary outcome measures rather than receiver operating characteristic (ROC) curves in mammography studies, although the latter are considered state of the art for many types of radiology reading studies. We argue that sensitivity and specificity are more clinically relevant and conceptually more straightforward than ROC curves. Second, we suggest a bivariate approach to data analysis for evaluating intervention effects on sensitivity and specificity. This accommodates the correlations inherent between these measures and allows for estimation of joint effects on them. Finally we propose a method for power calculations that uses computer simulation techniques. Simple formulas for sample size calculations are not available in part because variability in accuracy amongst readers and variation in difficulty among images introduce complexity into power calculations. The simulation method that we propose accommodates such complexity and is easy to implement.

The methodology was motivated by a study funded by the Department of Defense to evaluate the potential efficacy of an educational intervention. In the context of this study we illustrate the steps involved in power calculations and apply the data analytic techniques to the sort of data expected to result from this study. Though the proposed methods were motivated by this particular study, the statistical considerations are relevant more broadly in mammography and indeed in other types of radiologic imaging studies. Standards for the conduct of radiologic reading studies are not yet well developed, as they are for randomized clinical trials and for case-control studies. We hope that the discussion in this paper will add to the dialogue necessary for development of such standards. J CLIN EPIDEMIOL 50;12:1327–1338, 1997. © 1997 Elsevier Science Inc.

KEY WORDS. ROC curves, sensitivity and specificity, computer simulation, diagnostic tests, screening

1. INTRODUCTION

Mammography screening for breast cancer has been shown to be associated with decreased breast cancer mortality, at least in women over the age of 50 years [1]. Major efforts are currently underway to improve participation by women in screening programs [2]. Nevertheless, there is concern about the quality of mammography screening and there is general agreement that improvements in quality may lead to improvements in the performance of mammography as a screening modality. Quality might be improved for example by improving the imaging procedures. Alternatively, improvements in the accuracy with which mammographers interpret mammograms may improve the performance of screening mammography. Recent studies [3,4] have shown that there is considerable variability amongst radiologists in their interpretations of screening mammograms. Elmore *et al.* [3] observed that sensitivities ranged from 74% to 96% and that specificities ranged from 35% to 89% among 10 radiologists reading 150 selected mammograms. Beam *et al.* [4] using a much larger sample of 108 radiologists, each reading 79 mammograms, found sensitivities in the range of 47–100% and specificities in the range of 35–99%. These observations suggest that improvement in interpretation may be possible.

As part of a project called the Mammography Quality Improvement Project (MQIP) funded by the Department of Defense and aimed at improving the quality of mammog-

*Address for correspondence: Margaret Sullivan Pepe, Fred Hutchinson Cancer Research Center, Program in Biostatistics, 1124 Columbia Street, MP-665, Seattle, Washington 98104.
Accepted for publication on 20 August 1997.

raphy screening in rural communities, we are developing an educational program to improve the accuracy with which radiologists interpret mammograms. The educational intervention is composed of a series of five sessions in which mammographers read films and are provided with immediate feedback on the accuracy of their interpretations. Feedback is provided using a laptop personal computer that is mailed to the radiologist prior to his reading session. The computer program emphasizes the particular features of each mammogram that are relevant to determining the disease status of the woman screened. Eventually it may be possible to disseminate this sort of intervention over computer networks thus making it attractive in terms of easy accessibility and low cost.

To evaluate the impact of such an intervention on improvements in diagnostic accuracy it will eventually be necessary to perform a study of radiologists' interpretations of screening mammograms in their actual practices. As a preliminary step to such a large-scale study, we will evaluate the intervention effects in a more controlled setting. Specifically, we will have a number of radiologists read a selected set of mammograms before and after the intervention and evaluate changes in accuracy. The mammograms included in this controlled study will be composed of about 50% from women with disease, a proportion much larger than would be observed in practice but necessarily high to estimate sensitivity rates in a small-scale study. Mammograms will be selected to represent a reasonably broad range of interpretive difficulty.

The purpose of this paper is to elucidate some of the key statistical issues in the design of such a controlled reading study. Standards for the design of such studies are not well developed. This contrasts with therapeutic clinical trials and epidemiologic studies where the basic elements of study design are now fairly well standardized [5]. The question we propose to address in this reading study, namely evaluation of an intervention effect in a controlled setting, is a standard sort of question addressed in diagnostic imaging research. Hence the design issues which are dealt with here will have implications for future studies in mammography and in other diagnostic test settings. These same issues also arise in reading studies designed to compare different imaging modalities. The key issues concern the choice of relevant primary outcome measures, appropriate data analysis strategies, and methodology for power calculations that incorporates variability among radiologists and among images. Broader issues in regards to study designs for evaluating imaging tests have been discussed in a more general sense in the literature [6,7].

In Section 2, we consider two sets of measures that can be used to define accuracy in reading mammograms; first, sensitivity and specificity and second, ROC curves. We argue in favor of the former, in part, because they are more clinically relevant and most easily understood, but also because the latter can provide inappropriate conclusions concerning intervention benefits. In Section 3, we detail the basic elements of the statistical design of our study that could be considered a prototype for evaluating intervention effects in diagnostic radiology. An approach to joint analysis of sensitivity and specificity is outlined in Section 4. In Section 5, we describe methodology for power calculations that are appropriate for the proposed design and analysis. We propose the use of computer simulation methods for calculating power because they allow for complex designs and can easily incorporate variability amongst radiologists and images. Having described the steps involved in calculating power in Section 5, we then apply these procedures to the proposed MQIP study in Section 6, in order to illustrate the methods. Concluding remarks follow in Section 7.

2. MEASURES OF ACCURACY

2.1 Definitions

A radiologist reading a set of mammograms for a woman in our study will classify each breast according to his or her suspicion of its showing malignancy. The ACR lexicon for rating a breast [8] which we will employ, defines a 5-point scale with category 1 indicating "normal, routine follow-up recommended," 2 indicating "benign, routine follow-up," 3 indicating "probably benign, early recall recommended," 4 indicating "suspicious for cancer, consider biopsy," and 5 indicating "highly suspicious for cancer, biopsy recommended." A common definition of a screen positive mammogram is one that receives a rating of 4 or greater. These are mammograms that are sufficiently suspicious for cancer that biopsy is recommended and hence they have an impact on clinical practice. Sometimes a rating of a 3 or greater is considered positive. Because of the clinical implications of ratings 4 and 5, we will focus on the positivity criterion of category ≥ 4 here.

Given a definition for screen positivity, since there is a rating for each breast, one can calculate sensitivities and specificities with either "woman" or "breast" as the unit of analysis. The latter includes all non-diseased breasts (including non-diseased breasts from women with cancer), as the denominator for specificity and all diseased breasts as the denominator for sensitivity. However, since the consequences of false positive and false negative errors relate to the woman (rather than the breast), it seems more clinically relevant to use woman rather than breast as the unit of analysis. Thus, for example, we count the proportion of women with disease who have it detected as the sensitivity, rather than defining the sensitivity to be the proportion of diseased breasts which are detected. This accords with previous literature [3]. One could use the maximum of the ratings for the left and right sides as the woman level rating for calculation of sensitivity and specificity. Occasionally, however, a woman with unilateral disease may not have it detected in the affected side but will have a positive mammogram on the unaffected side. In this case, using the maximum rating

will inappropriately inflate the sensitivity. We define sensitivity instead as the proportion of women with disease who have it detected (a rating of ≥4) on the affected side. The specificity is the proportion of women without disease who have a maximum rating of less than 4.

ROC analysis is a statistical technique used to describe accuracy of diagnostic tests when the test outcome is either ordinal or continuous as opposed to binary. The rating data generated in radiology reading studies are ordinal and ROC analysis is often considered optimal for the analysis of such studies as is evidenced, for example, in a recent issue of *Academic Radiology* [9]. An ROC curve is constructed by varying the criterion used for defining a positive mammogram from "rating ≥2" to "rating ≥5," plotting the associated sensitivity and 1-specificity values against each other, and finally fitting a curve to the points so that the curve is anchored at (0,0) and (1,1). Various algorithms exist for fitting a curve, the most notable being the Dorfman-Alf algorithm based on the binormal model [10] and the empirical nonparametric method that simply connects observed ROC points linearly. The area under the ROC curve is usually used to summarize accuracy. Again we suggest that woman rather than breast should be the unit of analysis in defining the ROC curve. That is, in calculating the sensitivity corresponding to the criterion "rating ≥ K," it should be defined as the proportion of women with cancer who have a rating of ≥ K on an affected side.

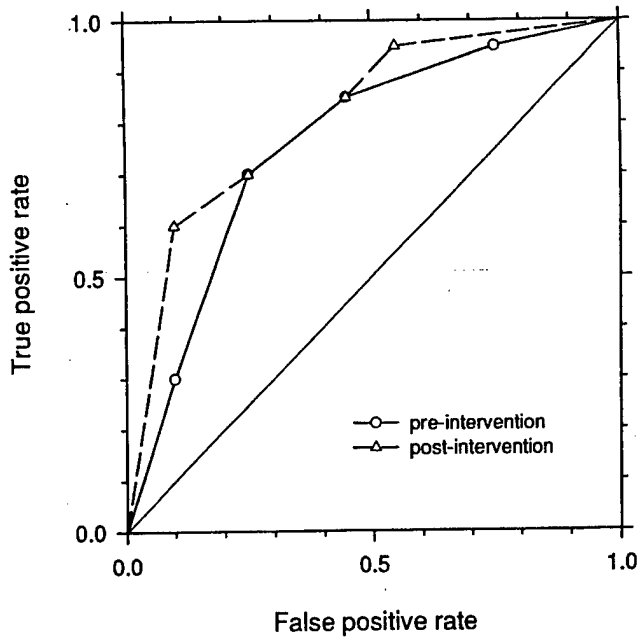


FIGURE 1. An hypothetical setting where the sensitivity and specificity associated with the clinically relevant criteria are unchanged but the empirical ROC curves indicate a benefit of intervention. The (false positive, true positive) points associated with categories 5, 4, 3, and 2 are (0.10, 0.30), (0.25, 0.70), (0.45, 0.85), and (0.75, 0.95) respectively, pre-intervention; and (0.10, 0.60), (0.25, 0.70), (0.45, 0.85), and (0.55, 0.95), respectively, post-intervention.

2.2 *ROC Analysis Versus Sensitivity and Specificity*

ROC analysis was developed originally for diagnostic tests with results on some arbitrary scale. Its primary advantage is that it allows one to assess the inherent capacity of the test to distinguish between diseased and non-diseased subjects without linking the test to some particular threshold for defining screen positive [11,12]. This seems appropriate in radiology experiments when image ratings are arbitrary numbers with no specific clinical meaning attached to them. In that case, shifts in the distributions of ratings are of no consequence as long as they are equally shifted for diseased and non-diseased subjects. In mammography, however, mammogram ratings have very specific clinical meanings and consequent clinical implications. Uniform shifts in the frequencies with which rating categories are chosen can have major clinical implications.

Moreover, in contrast to the prototype setting for ROC analysis, shifts between certain diagnostic categories are of more importance than others. For example, as noted by Kopans [13], whether an image is rated in category 4 versus category 5 has no clinical impact. Similarly classifications in category 1 versus category 2 are clinically irrelevant. However, shifts between categories 4 or 5 and between 1 or 2 can have a big impact on the ROC analysis. To illustrate this consider the setting shown in Fig. 1. The effect of intervention in this setting is to shift classifications of diseased observations from category 4 to category 5 and classification of non-diseased patients from category 2 to category 1. Though these changes are of no clinical import, the ROC type analysis indicates a benefit for the intervention. Thus an ROC analysis can indicate a benefit of intervention even though a clinically relevant benefit does not exist.

Of even more concern is the fact that a clinically relevant benefit of intervention can occur even when the ROC curves pre- and post-intervention are the same. Consider the ROC curve depicted in Fig. 2 for such a situation. The location on the ROC curve of the points associated with the criterion "rating ≥ category 4" indicate that sensitivity was significantly increased without decreasing specificity. This clinically relevant improvement in test accuracy does not manifest itself in an improvement in the ROC curves since the pre- and post-intervention curves are the same. (Interestingly, classic binormal ROC curves do not fit the situation depicted in Fig. 2 and a binormal ROC analysis in this setting may incorrectly indicate that the ROC curve post-intervention is improved over that pre-intervention).

The fact that ROC analysis can yield inappropriate conclusions regarding the clinically relevant effects of intervention argues against its use for the primary analysis of mammography reading study data. Another valid argument for not using an ROC analysis is that it is complicated and not easily understood by clinicians. Moreover, the so-called

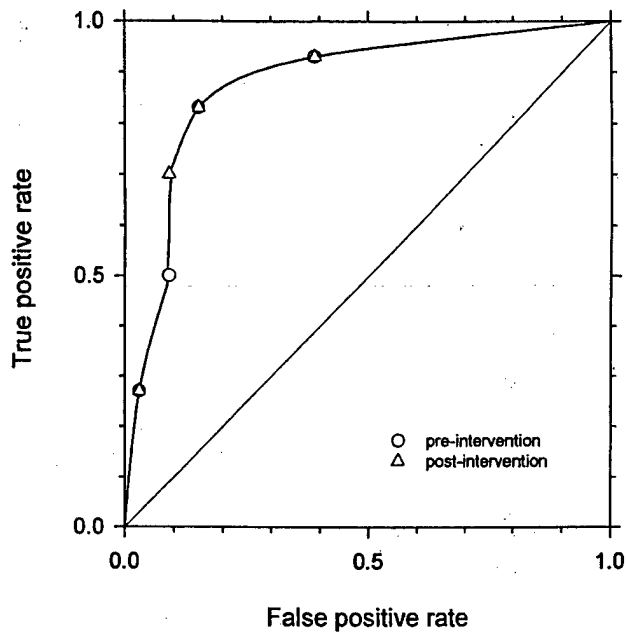


FIGURE 2. An hypothetical setting where ROC curve is unchanged by the intervention but there is a clinically relevant benefit. The sensitivity associated with the clinically relevant criterion is improved from 0.50 to 0.70 while the associated false positive rate remains unchanged at 0.09. The (false positive, true positive) points associated with categories 5, 4, 3, and 2 are (0.03, 0.27), (0.09, 0.50), (0.15, 0.83), and (0.39, 0.93) pre-intervention and (0.03, 0.27), (0.09, 0.70), (0.15, 0.83), and (0.39, 0.93) post-intervention. These points before intervention are labeled with circles and after intervention are labeled with triangles.

"area under the curve" that summarizes the ROC curve in a single number has an interpretation that is not well known or easily understood. It can be interpreted as the probability that a radiologist will have a greater suspicion of cancer from a mammogram from a woman with disease than from a woman without [14]. This probability, however, seems to be of more theoretical than practical relevance.

We propose using the more clinically meaningful quantities of sensitivity and specificity for the primary data analysis and employing ROC analysis as a secondary descriptive device. Though ROC analysis may be statistically more powerful in some settings, statistical power is of secondary importance relative to clinical relevance. Any study should be designed so that it has adequate power to detect changes in the quantities that are of practical relevance. Hence, we suggest that power calculations for a mammography reading study should be based on the ability to detect changes in sensitivity and specificity rather than on the basis of detecting changes in ROC curves.

3. STUDY DESIGN

We now describe the basic elements of the design that we propose for studies evaluating intervention effects on reading accuracy in mammography. In this prototype design, radiologists are randomly assigned to intervention and control groups, with the number in the former being denoted by R_T and the number in the latter denoted by R_C. Two image sets are constructed with M images in each set $S = 1,2$. In set S, a number M_D^S are from women with disease and this number may differ between the two sets. Each reader reads one set of images before the intervention period and one set after. It is important that the sets before and after intervention be different since readers may remember, to some degree, images that they have previously read. Half of the readers chosen at random in each of the intervention and control groups read set 1 before intervention and set 2 after intervention. The other half read them in the opposite order: set 2 followed by set 1. This cross-over of film sets eliminates the possibility of systematic bias due to film sets. The design is balanced in the sense that set 1 is read equally often before and after the intervention phase in both the intervention and control groups, and similarly for set 2. Readers are told the approximate prevalence of diseased images, i.e., $(M_D^1 + M_D^2)/2M$ and that this varies between the two sets. The rationale for telling the readers the approximate prevalence is that it will become apparent in any case after reading the first set of images and that *a priori* knowledge of it should reduce the potential impact as much as possible on the observed improvement in accuracy. Readers will use the ACR lexicon to classify mammograms and for each reading it will be determined if it is screen positive or negative according to whether the rating is at least 4 or less than 4.

Images for inclusion in the study need to be selected so that average sensitivity and specificity at the baseline assessment are relatively low. That is, improvements in accuracy should be possible with the sets of images chosen. If, in the absence of intervention all images from women with disease were easily identified as such, the observed sensitivities pre- and post-intervention would be close to 1 and a change in sensitivity would not be identifiable regardless of the actual effect of intervention. Thus at least some of the diseased images should be difficult but not impossible to identify as being from women with disease. Analogous considerations apply to specificity and the choice of non-diseased images included in the study.

4. DATA ANALYSIS

Having described the basic elements of the design and the choice of primary outcomes, we turn now to the strategy for data analysis. There are two components to the analysis. The first concerns a comparison of post- versus pre-intervention reading accuracy among the R_T readers in the intervention group. The second is the comparison of changes from pre- to post-intervention between the intervention and control groups. We first consider the former analysis, in part because it allows us to define notation most easily.

The purpose of this data analysis is to compare the overall sensitivity pre-intervention with that post-intervention

and to compare the overall specificity pre-intervention with that post-intervention. If $\hat{S}_{r,pre}$ and $\hat{S}_{r,post}$ denote the observed pre- and post-intervention sensitivities for radiologist r, then the observed change in the overall sensitivity $\hat{\Delta}_T$(sensitivity) is the average change in sensitivities across radiologists in the intervention group:

$$\hat{\Delta}_T(\text{sensitivity}) = \frac{1}{R_T}\sum_{r=1}^{R_T}(\hat{S}_{r,\text{post}} - \hat{S}_{r,\text{pre}}).$$

Similarly the observed change in the overall specificity in the intervention group is

$$\hat{\Delta}_T(\text{specificity}) = \frac{1}{R_T}\sum_{r=1}^{R_T}(\hat{F}_{r,\text{post}} - \hat{F}_{r,\text{pre}})$$

where $\hat{F}_{r,pre}$ and $\hat{F}_{r,post}$ denote the observed pre- and post-intervention specificities for radiologist r. Variance estimators for $\hat{\Delta}_T$(sensitivity) and $\hat{\Delta}_T$(specificity) are provided in the appendix. Although $\hat{\Delta}_T$(sensitivity) and $\hat{\Delta}_T$(specificity) are sample means of changes in sensitivities and specificities, their variances are not given by the usual variance formulae for sample means. Indeed such sample variances would overestimate the variability. Rather the correct variance estimators rely on acknowledging that there are in essence two strata of radiologists in the design, which are defined by the ordering of the two image sets which are rated. The variances of $\hat{\Delta}_T$(sensitivity) and $\hat{\Delta}_T$(specificity) are averages of stratum-specific variances, as shown in Appendix A.

Sensitivity and specificity are highly correlated parameters. Radiologists with high sensitivities tend to have low specificities. This will happen for example if they have a low threshold for classifying images as diseased. Similarly, changes in sensitivities and specificities induced by the intervention may be highly correlated. In particular, if the intervention simply changes the implicit threshold a radiologist has for classifying a mammogram as diseased then the sensitivity and specificity will both be changed but in opposite directions. Thus it is important to assess joint effects of intervention on sensitivity and specificity and to account for correlations between them in making inference. This can be accomplished by employing a bivariate analysis approach which is a special case of multivariate analysis, and for which there is a large statistical literature [15]. Using this approach to test the hypotheses that the true average sensitivity and specificity are unchanged by the intervention, H_0: Δ_T(sensitivity) = Δ_T(specificity) = 0, a chi-square test statistic is calculated. This statistic is a function of the observed average changes, $\hat{\Delta}_T$(sensitivity) and $\hat{\Delta}_T$(specificity), their variances and also their correlation. An expression for the chi-squared statistic is provided in the Appendix.

In addition to simply testing the hypothesis of no intervention effect, it will be important to provide a confidence region for the intervention effects on sensitivity and specificity based on the observed data. That is, a range of intervention effects, {Δ_T(sensitivity), Δ_T(specificity)}, which are consistent with the observed data. Such a joint 95% confidence region is defined formally as the set of values (x,y) for which the hypothesis H_0: {Δ_T(sensitivity) = x, Δ_T(specificity) = y} is not rejected at the 5% significance level. This region is an ellipse, centered at the observed intervention effect ($\hat{\Delta}_T$(sensitivity), $\hat{\Delta}_T$(specificity)). We refer the interested reader to the text [15] by Johnson and Wichern (1988, section 5.2) for technical details regarding its calculation. Code for calculating such regions has been written by Murdoch and Chow for the S-PLUS statistical software package and can be obtained from the S-archive on the Statlib computer site (http://lib.stat.cmu.edu). In a similar fashion a joint confidence region for the overall average sensitivity and specificity pre- or post-intervention can be calculated. It is calculated using the observed radiologist specific sensitivities and specificities pre- and post-intervention, and requires only calculation of the means, variances and correlations for these parameters. To illustrate these analyses, Fig. 3 displays joint confidence regions based on a simulated data set. In our opinion these confidence regions provide a simple summary of the information contained in study data regarding intervention effects on reading accuracy. In the simulated data, the analyses show that sensitivity was increased by the intervention whereas there is no evidence of change in specificity.

So far we have considered the comparison of post- versus pre-intervention reading accuracy within the intervention group. To attribute changes in accuracy to the intervention it will be necessary to compare the changes in the intervention group with those in the control group. Without the control group comparison, observed changes might be attributed to other factors, such as the increased reading practice or increased awareness of reader fallibility induced by participation in the study. Thus, turning now to the comparison of intervention and control groups, the main hypothesis to be tested is that the changes in sensitivity and specificity in the intervention group are the same as those in the control group. Using a subscript T to denote the intervention group and subscript C to denote the control group, the null hypothesis is H_0, Δ_C(sensitivity) = Δ_T(sensitivity), Δ_C(specificity) = Δ_T(specificity). A test statistic that has a chi-square distribution with 2 degrees of freedom is described in the appendix for testing this hypothesis. Joint confidence regions for the differences in changes between the groups, namely Δ_T(sensitivity) − Δ_C(sensitivity) and Δ_T(specificity) − Δ_C(specificity), can be calculated using methods analogous to those described earlier for the pre-versus-post-intervention comparison.

5. METHODOLOGY FOR POWER CALCULATIONS

Power calculations for the reading study are somewhat complicated. They must accommodate the facts that readers vary in their accuracy parameters of sensitivity and specificity, that their sensitivities and specificities are likely negatively correlated, that images vary in difficulty and that a

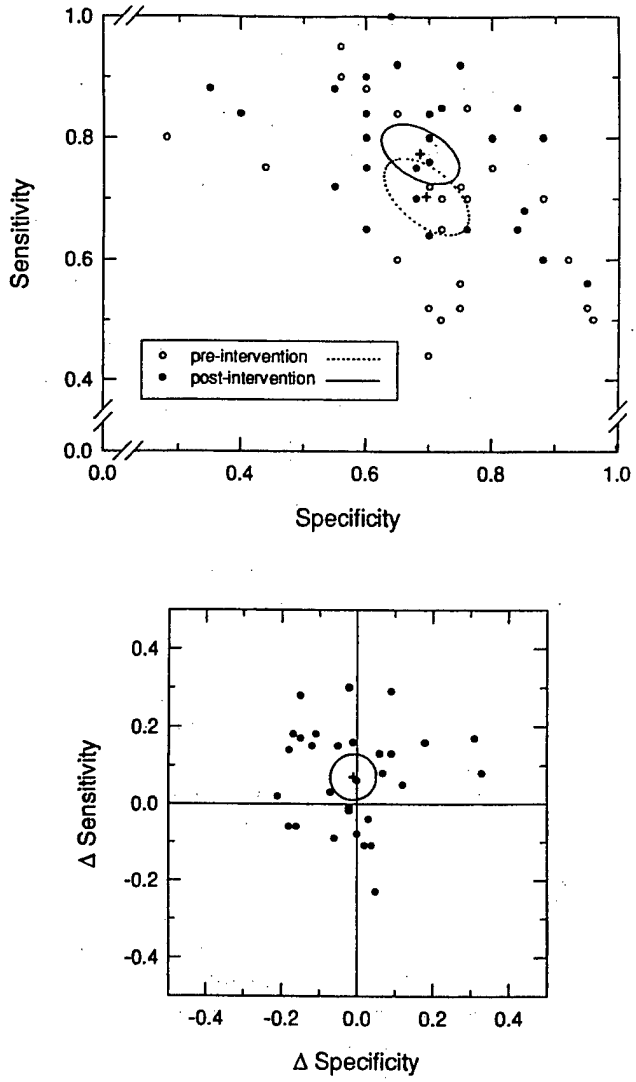


FIGURE 3. Joint confidence regions for sensitivity and specificity both pre and post intervention (*upper panel*) along with a joint confidence region (*lower panel*) for the changes in these parameters. Data used in this illustration were generated using computer simulation methods described in sections 5 and 6. Points correspond to observed data for individual radiologists.

bivariate analysis approach will be employed. These factors together make analytic expressions for sample size intractable. We instead take a computer simulation approach to power calculations. The simulation approach to power calculation is a general and standard method and indeed software has been developed for certain types of applications [16]. The basic idea is to repeatedly simulate data as it is expected or hoped to arise in the course of the study, and determine how often the null hypothesis is rejected. By definition the statistical power of the study is the proportion of simulated studies in which the null hypothesis is rejected. One calculates the power in this fashion using various sample sizes until a sample size is found that provides adequate power. This indirect computer intensive approach to sample size calculation is easily accomplished with modern computers.

5.1 Models for Pre- and Post-intervention Accuracy

To simulate study data we need to define precisely the mechanisms giving rise to the data. We therefore need to make assumptions about the reading accuracies before and after intervention. For this purpose we suppose that before intervention a reader correctly assesses a woman with tumor as being diseased with probability $P^D_{r,i}$. The probability $P^D_{r,i}$ depends on the image denoted by i and on the reader, denoted by r. The probabilities $P^D_{r,i}$ will presumably be higher if the tumor is clearly visible in image i than if it is not. The probabilities will also be higher if the radiologist is conservative and is inclined to recommend biopsy for borderline cases. We let S^D be the sensitivity of the average radiologist to the average film from a woman with tumor. The variability among films in terms of the difficulty that readers have in assessing them, is captured by specifying a distribution for the sensitivities that the average reader has in assessing the films. Here we assume that the average reader's sensitivity to films varies uniformly in an interval $(S^D - a^D, S^D + a^D)$ across different films. Thus for the average radiologist, easier films are read with sensitivity closer to $S^D + a^D$ and more difficult films are read with sensitivity closer to $S^D - a^D$. In a similar fashion, on the average film from a diseased woman, the sensitivity of different readers is assumed to vary uniformly in an interval $(S^D - b^D, S^D + b^D)$ across radiologists. Thus radiologists with high sensitivity to the average film will have sensitivity closer to $S^D + b^D$. In the appendix we detail a logistic model with random effects (also called a mixed model) for the probabilities $P^D_{r,i}$ that give rise to inter-image and inter-reader variability as postulated here. It is assumed that on the logistic scale there are no interactions between reader and image specific effects on the sensitivity.

Observe that for the purposes of simulating data, by specifying S^D and a^D we can now generate a random image effect by choosing a random number in $(S^D \pm a^D)$ that corresponds to the sensitivity an average radiologist has for detecting it. Similarly, having a specified S^D and b^D we are in a position to generate a random reader effect by choosing a random number in $(S^D - b^D, S^D + b^D)$ that corresponds to his sensitivity to the average film. The logistic model displayed in the appendix then yields the probability $P_{i,r}$ that that reader has of correctly assessing that image as diseased.

Analogous considerations apply to the determination of randomly generated specificities which vary across radiologists and across images from women without disease. Values for parameters $F^{\bar{D}}$, $b^{\bar{D}}$ and $a^{\bar{D}}$ need to be specified in order to define the data generating process. Here, $F^{\bar{D}}$ is the probability that the average radiologist will correctly assess the average non-diseased image as such, radiologists vary uni-

formly in $(F^{\bar{D}} - b^{\bar{D}}, F^{\bar{D}} + b^{\bar{D}})$ in their specificities to the average non-diseased film, and images from women without disease vary uniformly in $(F^{\bar{D}} - a^{\bar{D}}, F^{\bar{D}} + a^{\bar{D}})$ in the probabilities of the average reader correctly classifying them. The sensitivities and specificities from single radiologists should be correlated. In the Appendix we describe how negative correlation between sensitivities and specificities within radiologists can be built into the data simulation mechanism.

In summary, for each study radiologist we simulate his/her sensitivity and specificity to the average diseased and non-diseased films, respectively, by randomly sampling correlated numbers from $(S^D - b^D, S^D + b^D)$ and $(F^{\bar{D}} - b^{\bar{D}}, F^{\bar{D}} + b^{\bar{D}})$, respectively. For each study film we determine the sensitivity or specificity that an average radiologist has for it by randomly sampling a number from $(S^D - a^D, S^D + a^D)$ or $(F^{\bar{D}} - a^{\bar{D}}, F^{\bar{D}} + a^{\bar{D}})$. Finally, for each combination of film i and radiologist r, we can calculate $P^D_{i,r}$ or $P^{\bar{D}}_{i,r}$, which is the probability that the radiologist will assess that image correctly.

The $P^D_{i,r}$ and $P^{\bar{D}}_{i,r}$ pertain to probabilities before intervention in the treatment and control groups. One also needs to specify treatment effects in order that corresponding probabilities after intervention can be calculated. We postulate that after intervention the quantities S^D and $F^{\bar{D}}$ are changed to new values but that the variations among readers and among images remain the same. In the Appendix we define in a mathematically precise way a logistic model that incorporates such intervention effects.

5.2 *Simulated Study Data Generation*

Having specified statistical models for pre- and post-intervention rating probabilities that incorporate variation among radiologists and among images, we now turn to the simulation of study data in accordance with the study design that we proposed in section 3. The first step is to generate images and image sets. This entails generating M diseased images (i.e., M image-specific parameters, one for each image), generating M non-diseased images, and finally from the 2M films choosing M at random without replacement to form film set 1. The remaining M films constitute film set 2. The next step is to generate R_T intervention readers and R_C control readers and assign them film sets. That is, for each of $R_T + R_C$ readers we generate pairs of pre- and post-intervention sensitivities and specificities to average diseased and non-diseased films according to the models described in section 5.1. Of the total $R_T + R_C$ readers, R_T are assigned at random to the intervention group and the remaining R_C to the control group. Finally film set orderings are assigned to the readers with half of the intervention readers selected at random being assigned set 1 first and the other half assigned set 2 first. Similarly, $R_C/2$ control readers are assigned set 1 followed by set 2 and the other $R_C/2$ readers are assigned film sets in the opposite order.

The final step in generating data for a simulated study is to actually generate the readings for each reader and image combination. That is, for each reader and for each of the M films in his/her pre-intervention set, a binary random variable is generated which is his/her assessment of whether or not that image shows disease using the probability $P^D_{r,i,\text{pre}}$ if the image is diseased and $1 - P^{\bar{D}}_{r,i,\text{pre}}$ if the image is not diseased. Similarly, for each of the M films in his/her post-intervention set a similar binary random variable is generated using $P^D_{r,i,\text{post}}$ or $1 - P^{\bar{D}}_{r,i,\text{post}}$ noting that the pre- and post-probabilities differ by different amounts for intervention-versus-control radiologists.

Having generated the simulated study data the test statistics of interest can now be calculated. Data are simulated (first the probabilities, then the ratings) and results calculated under the same assumptions and study design many times, with 1000 or 5000 simulated datasets being typical numbers used for power calculations. The proportion of simulated studies in which the null hypothesis is rejected is the calculated study power for that design and under those assumptions.

6. POWER CALCULATIONS: RESULTS FOR THE MQIP STUDY

To fix ideas, we now illustrate the computer simulation method for power calculations in the MQIP study. This illustration also identifies some sources of data to guide assumptions for power calculations.

We need to choose assumed parameters for the baseline sensitivities and specificities, for the variations among radiologists and among images and for intervention effects of interest. We assume that the median sensitivity pre-intervention, S^D, in our study will be in the range of 0.70 to 0.80. This accords with previous studies that found median sensitivities of 0.70 and 0.80 [3,4]. Median pre-intervention specificity will also be assumed to lie in the range of 0.70 to 0.80. Beam *et al.* [4] found a median specificity of 0.94 for mammograms from women with normal mammograms and a median specificity of 0.60 for mammograms from women with benign disease. Elmore *et al.* [3] found a median specificity of 0.94. In contrast to these studies, we will inform the radiologists of the average prevalence that is higher than that expected in a practical screening setting. Because of this and the fact that the films in our study will be somewhat difficult, we anticipate an initial specificity lower than observed in those studies. The variation amongst radiologists in sensitivities and specificities will be assumed such that $b^D = 0.20$ and $b^{\bar{D}} = 0.20$, which is in agreement with the range of approximately 40% in sensitivities (and specificities) among radiologists observed in Beam's study. We could find no data on inter-image variability to suggest appropriate values for a^D and $a^{\bar{D}}$. We assume that they are of the same order of magnitude as the inter-rater variability parameters, $a^D = a^{\bar{D}} = 0.20$. With regard to intervention effects of interest, we consider that changes of 10 percentage

TABLE 1. Power to detect a 10% increase in sensitivity and no effect on specificity in the intervention group

Readers per group (R_T)	Films per set (M)	Pre-intervention sensitivity	Pre-intervention specificity	Power	
				Within intervention group	Comparison with control group
20	30	0.70	0.70	0.70	0.38
20	30	0.70	0.80	0.66	0.34
20	30	0.80	0.70	0.79	0.45
20	30	0.80	0.80	0.77	0.44
20	45	0.70	0.70	0.81	0.48
20	45	0.70	0.80	0.82	0.53
20	45	0.80	0.70	0.91	0.61
20	45	0.80	0.80	0.92	0.64
30	30	0.70	0.70	0.81	0.48
30	30	0.70	0.80	0.83	0.52
30	30	0.80	0.70	0.93	0.60
30	30	0.80	0.80	0.91	0.61
30	45	0.70	0.70	0.94	0.66
30	45	0.70	0.80	0.95	0.66
30	45	0.80	0.70	0.99	0.80
30	45	0.80	0.80	0.99	0.79
40	30	0.70	0.70	0.92	0.61
40	30	0.70	0.80	0.94	0.60
40	30	0.80	0.70	0.97	0.73
40	30	0.80	0.80	0.98	0.75
40	45	0.70	0.70	0.98	0.79
40	45	0.70	0.80	0.99	0.80
40	45	0.80	0.70	0.99	0.88
40	45	0.80	0.80	0.99	0.89

All tests are two sided and are tested at a significance level of 0.05.

points in either sensitivity or specificity are of interest. However, we calculated power for a variety of intervention effects.

Practical considerations concerning time and cost dictate the range of sample sizes that are feasible and therefore, for which power calculations are performed. We anticipate that no more than approximately 80 radiologists are available for the reading study in the rural communities in which our mammography quality improvement study is being conducted. To maximize power, equal numbers of radiologists are assigned to control and intervention groups. Therefore the number of radiologists per group to be considered for power calculation purposes will be in the range of 20–40. Experience suggests that readers can comfortably read no more than 45 films per session. We therefore calculated power for experiments in which the number of films per set, M, was either 30 or 45.

Estimates of power based on computer simulations are shown in Table 1. Though results are shown only for intervention effects on sensitivity with no effect on specificity, because of the symmetry inherent in the design, the same power calculations hold for a 10% change in specificity with no change in the sensitivity. Observe that the power is far larger for the within intervention group assessment of change than for the between group comparison of change. This is to be expected since the variability involved in comparing two random changes is greater than the variability involved in comparing a single change with the null hypothesis of no change. We also observe from Table 1 that the power is less when the baseline sensitivity is 0.70 than when it is 0.80. This is due to the relatively larger binomial variance for the lower baseline rate. To be conservative we focus on this lower rate. Interestingly, the baseline specificity had little impact on the power to detect an intervention effect on the sensitivity.

The target power for our study design is 90%, which allows a 10% chance of an inconclusive result when the intervention increases sensitivity from 0.70 to 0.80. For the within intervention group comparison this cannot be achieved with 20 readers, but it can be achieved with 30 readers if 45 images are included in each image set. The between group comparison, however, has a power of only 66% in this case. Even with use of our maximum resources, i.e., 40 readers per group and 45 images per reading set, the power is only 80%. This allows for a 20% chance of an inconclusive result even when there is a clinically important intervention effect on diagnostic accuracy.

For the MQIP study we chose not to include a control

TABLE 2. Study power to detect various configurations of changes in the intervention group using a study design with 30 readers and 45 films per set

Pre-intervention sensitivity	Δ_T(sens)	Δ_T(spec)	Power
0.60	+0.10	0.00	0.90
0.70	+0.10	0.00	0.95
0.80	+0.10	0.00	0.98
0.60	+0.05	0.00	0.35
0.70	+0.05	0.00	0.39
0.80	+0.05	0.00	0.50
0.60	+0.05	+0.05	0.66
0.70	+0.05	+0.05	0.68
0.80	+0.05	+0.05	0.71

The pre-intervention specificity is assumed to be 0.70 in all cases. The intervention induced change in sensitivity as denoted Δ_T(sens) and in specificity is denoted Δ_T(spec).

group in the reading study component, but instead to focus the study on the within group comparison. The power calculations were an important contribution to this decision but other considerations also played a role. Radiologists would have little motivation to participate in the control arm whereas they would receive continuing medical education (CME) credit for participation in the intervention arm. The possibility that those in the control arm would learn from the baseline assessment was also a concern and thus we were concerned that it might not even be feasible to construct a true control group. Finally, it was felt that if we found a definite positive change in the intervention group, then this would provide sufficient motivation to proceed with more comprehensive controlled studies in the future. Thus we chose to study only the intervention effects in the intervention group and to use sample sizes of 30 radiologists each reading sets of mammograms from 45 women before and after intervention.

The simulation program allowed us the flexibility to explore the performance of this study design in a variety of settings other than that assumed for the primary sample size calculation. First we calculated the probability of rejecting the null hypothesis for settings where there was no intervention effect. Recall that inference for the test statistic is based on a chi-square statistic and is theoretically valid with large samples. However, this study entails relatively small samples. We used the simulations to check the adequacy of the large sample theory in our study. To do this we generated data under the null hypothesis. The rejection probability was approximately 0.06 in the settings we studied, indicating that the true significance level of the test is slightly higher than the target of 0.05 but adequate for our purposes.

We next explored the power of this study design and sample sizes to detect an array of intervention effects. Results are shown in Table 2. Although the study has adequate power to detect a change in sensitivity (or specificity) of 0.10 even when the pre-intervention sensitivity is as low as 0.60, it has little chance of detecting a smaller change of 0.05. On the other hand, if small changes of the order of 0.05 occur in both the average sensitivity and in the average specificity there is a good chance that the simultaneous effects will be detected.

7. DISCUSSION

Diagnostic imaging technology is already a basic component of medical care and continues to develop at a rapid pace. It is clearly important to assess the accuracy with which readers can diagnose disease using such technologies, to evaluate the effects of training strategies and to compare methods. Implications for public health can be enormous. Unfortunately, statistical methodology for evaluating and comparing imaging methods has not received much attention by biostatisticians and epidemiologists involved in public health research. Rather the literature is concentrated in radiology research journals, has generally focused on small scale studies involving only a few readers and has ignored clinical implications associated with different diagnostic categories. We believe that it is time to bring the discussion about study design and analysis for evaluating imaging technology to the broader community of epidemiologists and statisticians involved in public health. This is particularly important as interest increases in the accuracies and costs of these imaging methods. By presenting our thoughts on the design and analysis of a study to evaluate an educational intervention on the interpretation of mammograms, we hope to stimulate such discussion.

The choice of primary outcome measure is the most basic element of any study design. We chose to consider the sensitivity and specificity as the basis for evaluating intervention effects. This conflicts with initial statistical reviewers of our study design who were of the opinion that ROC analysis was the only appropriate and indeed the state-of-the-art basis for evaluating an intervention effect. We now argue that in mammography where specific clinical actions are associated with diagnostic rating categories, sensitivity, and specificity provide a more clinically relevant and conceptually straightforward basis for comparison than does ROC analysis. Moreover this approach allows us to evaluate effects on false positive as well as true positive rates. In contrast ROC analysis does not quantify the false positive rates directly but in a sense only uses it to standardize the true positive rate. We do not dismiss ROC analysis entirely but rather we regard the analysis of the specific rating categories of secondary importance and focus the design on sensitivity and specificity. Thus the MQIP study was designed to ensure adequate power to detect changes in the most clinically relevant quantities.

We also needed to decide upon the analysis techniques for making statistical inference about sensitivity and specificity. We propose to simultaneously estimate sensitivity and specificity using multivariate methods. Sensitivity and specificity as we have defined them are average sensitivities

and average specificities of radiologists in our study. They can also be interpreted as marginal or population average quantities, in the sense of being the probability that a diseased (or non-diseased) image will be correctly interpreted as such in the study. The distinction between the population average and average radiologist-specific interpretations has to do with whether one considers the accuracy parameters to be based on data pooled across radiologists (population average) or to be based on calculation of the accuracy parameter for each radiologist and then averaging the results. In our study these quantities coincide because all radiologists expect to read the same numbers of films. In studies where this is not the case, the distinction should be considered and a decision should be made regarding which of the two entities is most relevant.

The approach we propose for statistical inference is relatively straightforward, being based on methods for inference about sample means. Confidence intervals are based on the variance-covariance matrix of the estimated (sensitivity, specificity) parameters or their changes amongst radiologists. Possible non-normality of the average estimates may be an issue in our study, though for the settings considered in the power calculation this did not appear to be the case. An alternative approach to inference which might be more robust would follow the marginal regression modeling approach described by Leisenring, Pepe, and Longton [17]. One could formulate logistic regression models for the population average sensitivity and 1-specificity as

$$\text{logit}\ \{\text{Prob[screen positive} \mid \text{image diseased]}\} = \gamma_0 + \gamma_1 b$$

$$\text{logit}\ \{\text{Prob[screen positive} \mid \text{image non-diseased]}\} = \eta_0 + \eta_1 b$$

where the logit function is logit $\{x\} = ln\ \{x/(1 - x)\}$ and b is 0 if the image was read before the intervention and 1 if it was read after the intervention. The changes in the true and false positive rates are now quantified in the odds ratio parameters γ_1 and η_1, respectively, and joint confidence intervals can be calculated. By adding an interaction term between b and I, where I is an indicator of the radiologist being in the control or intervention groups:

$$\text{logit}\ \{\text{Prob[screen positive} \mid \text{image diseased]}\} = \gamma_0 + \gamma_1 b + \gamma_2 bI$$

$$\text{logit}\ \{\text{Prob[screen positive} \mid \text{image non-diseased]}\} = \eta_0 + \eta_1 b + \eta_2 bI$$

a comparison of the changes in the intervention and control groups can be made by testing if the parameters γ_2 or η_2 are 0. Though this logistic regression modeling approach may provide more robust confidence intervals, we felt that the simpler approach described earlier was adequate for power calculations.

The prototype reading study we have described concerns evaluating the effect of an intervention on the change in accuracy parameters. We note, however, that most of our discussion is also relevant to the comparison of accuracies associated with different imaging modalities. Suppose for example, that there are two sets of women (denoted by set 1 and set 2) from which images have been made using two modalities. A natural study design to compare the modalities would entail readers assigned to read one set of films produced with one modality and the other set of films produced with the other modality. Using the notation $1(A)$ to denote set 1 produced with modality A and similarly for the other combination, readers read either $\{1(A)$ and $2(B)\}$ or $\{2(A)$ and $1(B)\}$. Considering that the ordering may also influence accuracy parameters, this yields four groups of readings, $\{1(A), 2(B)\}$, $\{2(B), 1(A)\}$, $\{2(A), 1(B)\}$ and $\{1(B), 2(A)\}$. A balanced cross-over design would assign radiologists randomly to these four reading assignments. The difference in the sensitivity and specificity between modality A and B can be calculated by simply pooling all relevant readings for modality A and similarly for modality 2. Inference for the difference follows in the same fashion as that described for the change induced by intervention in the intervention group of our study but that now there are 4 rather than 2 strata of radiologists defined by the image reading set assignments.

Power calculations for reading studies are not straightforward due in part to correlations induced by images and readers. That is, for each image there are multiple readings. Moreover, each reader provides multiple readings and radiologist specific sensitivities and specificities are correlated. We propose simple analyses for dealing with these factors but power calculations required a computer simulation approach. We found the process of developing the computer simulation study to be a useful exercise. It compels one to think through the processes generating study data. It also allows one to experiment with the assumptions and design easily. For example, we considered designs that included a larger number of film sets to be read in the study and found that the study power was decreased slightly due to the extra variation introduced. Computer simulations also allow one to check how test statistics perform under the null hypothesis with sample sizes proposed in the study. Hence one can check if inference based on large sample theory is valid in the setting where it is to be applied. We suggest that simulation studies are a useful approach to power calculations in any setting, though given the complexities in radiology reading studies, the case for the technique in this setting is particularly strong.

We appreciate the support of grants GM54438 and CA63146 awarded by the National Institutes of Health, and grant DAMD17-96-1-6288 awarded by the Department of Defense. We thank Molly Edmonds for her excellent technical help in preparing the manuscript.

References

1. Miller AB, Chamberlain J, Day NE, Hakama M, Prorok PC. Report on a workshop of the UICC Project on Evaluation of Screening for Cancer. **Int J Cancer** 1990; 46: 761–769.
2. Rakowski W, Andersen MR, Stoddard AM, Urban N, Rimer BK, Lane DS, Fox SA, Costanza ME for the NCI Breast Cancer Screening Consortium. A confirmatory analysis of the pros and cons of mammography. **Health Psychol** 1997; 16: 433–441.
3. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologist's interpretation of mammograms. **N Engl J Med** 1994; 331: 1493–1499.
4. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists: Findings from a national sample. **Arch Int Med** 1996; 156: 209–213.
5. Begg CB. Advances in statistical methodology for diagnostic medicine in the 1980's. **Stat Med** 1991; 10: 1887–1895.
6. Begg CB, McNeil BJ. Assessment of radiologic tests: Control of bias and other design considerations. **Radiology** 1988; 167: 565–569.
7. Gatsonis C, McNeil BJ. Collaborative evaluations of diagnostic tests: Experience of the Radiology Diagnostic Oncology Group. **Radiology** 1990; 175: 571–575.
8. American College of Radiology. **Breast Imaging Reporting and Data System**. Second Edition. Reston, Virginia: American College of Radiology; 1995.
9. Advances in Statistical Methods for Diagnostic Radiology: A Symposium. **Academic Radiology** 1995; 2: S1.
10. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals: Rating method data. **J Mathematical Psychol** 1969; 6: 487–496.
11. Swets JA, Pickett RM. **Evaluation of Diagnostic Systems: Methods from Signal Detection Theory**. New York: Academic Press; 1982.
12. Metz CE. ROC methodology in radiologic imaging. **Invest Radiol** 1986; 21: 720–733.
13. Kopans DB. The accuracy of mammographic interpretations. **N Engl J Med** 1994; 331: 1521–1522.
14. Hanley JA, McNeil BJ. The meaning and use of the area under operating characteristics curve. **Radiology** 1982; 143: 29–36.
15. Johnson RA, Wichern DW. **Applied Multivariate Statistical Analysis**. New Jersey: Prentice Hall; 1988.
16. SER Corporation. **EGRET Siz Module**. Cambridge, Massachusetts: Cytel Software Corporation; 1997.
17. Leisenring W, Pepe MS, Longton GL. A marginal regression modelling framework for evaluating medical diagnostic tests. **Stat Med** 1997; 16: 1263–1281.

APPENDIX A

1. VARIANCE ESTIMATORS FOR CHANGE IN OVERALL SENSITIVITY AND SPECIFICITY

The change in the overall sensitivity defined in Section 4 can be written formally mathematically as

$$\hat{\Delta}_T(\text{sensitivity}) = \frac{1}{R_T}\left\{\sum_{r:(\text{order}=1,2)} (\hat{S}_{r,\text{post}} - \hat{S}_{r,\text{pre}}) + \sum_{r:(\text{order}=2,1)} (\hat{S}_{r,\text{post}} - (\hat{S}_{r,\text{pre}})\right\}$$

where $\hat{S}_{r,\text{pre}}$ is the observed sensitivity for radiologist r with his pre-intervention film set and $\hat{S}_{r,\text{post}}$ is the corresponding quantity post-intervention. Observe that the order of film sets essentially defines two strata in this setting and the notation (order = 1,2) (or [order = 2,1]) used to denote the stratum in the summation indicates that it includes only radiologists assigned sets in the order set 1 first and set 2 second (or set 2 first and set 1 second). The variance of $\hat{\Delta}_T$(sensitivity) can be estimated using the variance of a stratified sample mean $\hat{V} = 0.5(\hat{V}_{(1,2)} + \hat{V}_{(2,1)})/R_T$, where $\hat{V}_{(1,2)}$ is the sample variance of the quantities $(\hat{S}_{r,\text{pre}} - \hat{S}_{r,\text{post}})$ in the stratum (order = 1,2), and $\hat{V}_2$ is the analogous quantity in the other stratum. The ratio $\hat{\Delta}_T(\text{sensitivity})/\sqrt{\hat{V}}$ can be compared with a standard normal distribution to test for a change in the sensitivity which is statistically significantly different from 0.

2. *Chi-Square Test Statistics for Bivariate Analyses*

To simultaneously test the null hypotheses that both the sensitivity and specificity are unchanged in the intervention group, H_0: $\Delta_T(\text{sensitivity}) = 0 = \Delta_T(\text{specificity})$, the following test statistic can be used

$$[\hat{\Delta}_T(\text{sensitivity})\ \hat{\Delta}_T(\text{specificity})]\ \hat{\sum}_T^{-1} \begin{bmatrix} \hat{\Delta}_T(\text{sensitivity}) \\ \hat{\Delta}_T(\text{specificity}) \end{bmatrix}$$

where the square bracket notation is used to denote vectors and $\hat{\Sigma}_T^{-1}$ is the inverse of a square matrix $\hat{\Sigma}_T$. This matrix $\hat{\Sigma}_T$ is a variance-covariance matrix for the two-dimensional statistic $[\hat{\Delta}_T(\text{sensitivity})\ \hat{\Delta}_T(\text{specificity})]$, and is the analogue of the variance $\hat{V}$ defined above in relation to the one-dimensional quantity $\hat{\Delta}_T$(sensitivity). Formally we write

$$\hat{\sum}_T = 0.5\left\{\hat{\sum}_T^{(1,2)} + \hat{\sum}_T^{(2,1)}\right\}/(R_T - 1)$$

where $\hat{\Sigma}_T^{(1,2)}$ is the sample variance-covariance matrix for the quantities $\{\hat{S}_{r,\text{post}} - \hat{S}_{r,\text{pre}}\ \hat{F}_{r,\text{post}} - \hat{F}_{r,\text{pre}}\}$ in the stratum (order = 1,2), and $\hat{\Sigma}_T^{(2,1)}$ is the analogous quantity calculated for the other stratum. The test statistic is compared with a standard chi-square distribution with 2 degrees of freedom in order to test the null hypothesis concerning changes in sensitivities and specificities.

Consider now the component of the data analysis concerning the comparison of changes between intervention and control groups. Using a subscript C to denote the control group in analogy with our use of the subscript T to denote the intervention group, we define the statistics $\hat{\Delta}_C$(sensitivity), $\hat{\Delta}_C$(specificity) and $\hat{\Sigma}_C$. The estimated differences between the groups in changes of sensitivities and specificities can be written as $\hat{\Delta}_T(\text{sensitivity}) - \hat{\Delta}_C(\text{sensitivity})$ and $\hat{\Delta}_T(\text{specificity}) - \hat{\Delta}_C(\text{specificity})$, respectively. The hypothesis that the changes are the same for intervention and control groups can be tested by comparing the statistic

$$[\hat{\Delta}_T(\text{sens}) - \hat{\Delta}_C(\text{sens})\ \hat{\Delta}_T(\text{spec}) - \hat{\Delta}_C(\text{spec})] \times \left[\sum_T + \sum_C\right]^{-1} \begin{bmatrix} \hat{\Delta}_T(\text{sens}) - \hat{\Delta}_C(\text{sens}) \\ \hat{\Delta}_T(\text{spec}) - \hat{\Delta}_C(\text{spec}) \end{bmatrix}$$

with the quantiles of a chi-square distribution with 2 degrees of freedom, where we use the abbreviations "sens" and "spec" to denote "sensitivity" and "specificity" in the above expressions.

3. Mixed Models for Reading Accuracies

Section 5 outlines a statistical model for sensitivity and specificity parameters which vary with reader and image. Here we present a more formal and precise definition of this model. For radiologist r on diseased film i, we write the chance of correctly identifying it as diseased pre-intervention using a logistic model as

$$P^D_{r,i} = \exp\{\mu^D + \gamma^D_i + \beta^D_r\}/(1 + \exp\{\mu^D + \gamma^D_i + \beta^D_r\})$$

where γ^D_i and β^D_r are random variables specific to this film and radiologist, respectively. For the average radiologist $\beta^D_r = 0$, and for the average film $\gamma^D_i = 0$. Thus for the average radiologist on the average film the sensitivity is $S^D = \exp\{\mu^D\}/(1 + \exp\{\mu^D\})$. The films vary in difficulty in the sense that the average radiologist has a lower sensitivity on some films and a higher sensitivity on others. Mathematically this translates into allowing γ^D_i to vary. We choose it as a random variable so that the average radiologist's sensitivity to different films varies uniformly in an interval $(S^D - a^D, S^D + a^D)$. Technically this is achieved by letting $\gamma^D_i = \ln\{U^D_i/(1 - U^D_i)\} - \mu^D$ where U^D_i is a random variable with a uniform distribution in $(S^D - a^D, S^D + a^D)$. The radiologists also vary amongst themselves in their sensitivities to the same film and this inter-rater variation translates into allowing β^D_r to vary. We simulated data so that on the average diseased film (i.e., $\gamma^D_i = 0$) the sensitivities of radiologists varied uniformly in $(S^D - b^D, S^D + b^D)$. Again, technically we let $\beta^D_r = \ln\{U^D_r/(1 - U^D_r)\} - \mu^D$ where U^D_r is a random variable with a uniform distribution on the interval $(S^D - b^D, S^D + b^D)$.

Turning now to specificities, we write the specificity for radiologist r on non-diseased film j pre-intervention as

$$P^{\bar{D}}_{r,j} = \exp\{\mu^{\bar{D}} + \gamma^{\bar{D}}_j + \beta^{\bar{D}}_r\}/(1 + \exp\{\mu^{\bar{D}} + \gamma^{\bar{D}}_j + \beta^{\bar{D}}_r\})$$

where in analogy with the above notation for diseased films, the average radiologist on the average film has specificity $F^{\bar{D}} = \exp\{\mu^{\bar{D}}\}/(1 + \exp\{\mu^{\bar{D}})$ and parameters $a^{\bar{D}}$ and $b^{\bar{D}}$ indicate variation in the specificity with film and radiologist. As argued in section 5, data should be generated so that the β^D_r and $\beta^{\bar{D}}_r$ are negatively correlated. We incorporated this into the simulation by first generating the sensitivity radiologist-specific random effect parameter, β^D_r, (i.e., his/her sensitivity to the average film) which is based on the random variable U^D_r, and then letting the corresponding random variable for the specificity random effect be defined as

$$U^{\bar{D}}_r = \left\{\left(F^{\bar{D}} - (U^D_r - S^D)\frac{b^{\bar{D}}}{b^D}\right)\right\}.$$

Thus if the radiologist's sensitivity is $x \times b^D$ above the average radiologist's sensitivity to the average film, S^D, his/her specificity will be $x \times b^{\bar{D}}$ below the average specificity to the average film.

Our model postulates that after intervention the quantities $F^{\bar{D}}$ and S^D are changed to new values but that the radiologist and image-specific parameters remain unchanged. Thus, suppose that after intervention the sensitivity of the average radiologist to the average film is $\exp(\mu^D + \alpha^D\}/(1 + \exp\{\mu^D + \alpha^D\})$. Then the chances that radiologist r will correctly classify film i pre- and post-intervention are

$$P^D_{r,i,\text{pre}} = \exp\{\mu^D + \gamma^D_i + \beta^D_r\}/(1 + \exp\{\mu^D + \gamma^D_i + \beta^D_r\})$$

and

$$P^D_{r,i,\text{post}} = \exp\{\mu^D + \alpha^D + \gamma^D_i + \beta^D_r\} /(1 + \exp\{\mu^D + \alpha^D + \gamma^D_i + \beta^D_r\}),$$

respectively. Similarly the postulated change in $F^{\bar{D}}$ specifies a parameter $\alpha^{\bar{D}}$ (analogous to α^D) which facilitates calculation of post-intervention specificities. Having chosen values for the various parameters $(\mu^D, \alpha^D, a^D, b^D)$ and $(\mu^{\bar{D}}, \alpha^{\bar{D}}, a^{\bar{D}}, b^{\bar{D}})$, this completes the first step of the simulation power calculation method, namely specification of accuracy parameter distributions pre-intervention and intervention effects.