1994

Variability of radiologists' mammographic interpretations and recommendations for management

Debra Hope Howard
Yale University

VARIABILITY OF RADIOLOGISTS' MAMMOGRAPHIC INTERPRETATIONS AND RECOMMENDATIONS FOR MANAGEMENT

Debra Hope Howard

Yale University

1994

Permission for photocoping or microfilming of " VARIABILITY OF RADIOLOGISTS' MAMMOGRAPHIC INTERPRETATIONS AND RECOMMENDATIONS FOR MANAGEMENT"

(Title of thesis)

for the purpose of individual scholarly consultation or reference is hereby granted by the author. This permission is not to be interpreted as affecting publication of this work or otherwise placing it in the public domain, and the author reserves all rights of ownership guaranteed under common law protection of unpublished manuscripts.

Debra Hope Howard

Signature of Author

4·1·94

Date

VARIABILITY OF RADIOLOGISTS' MAMMOGRAPHIC INTERPRETATIONS AND RECOMMENDATIONS FOR MANAGEMENT. Debra H. Howard, Carolyn K. Wells, Carol H. Lee, Alvan R. Feinstein, and Joann G. Elmore. Department of Internal Medicine, Yale University School of Medicine, New Haven, CT.

Despite widespread use of mammograms as a screening tool for breast cancer, variability in mammogram interpretation has not been extensively studied. For this project, 10 radiologists were 'blinded' to the research hypothesis. Using a standardized coding form, they interpreted identical sets of 150 mammograms, on two occasions, separated by a five month 'wash-out' period. There was a wide range of variability among the 10 radiologists in their use of diagnostic categories for the 150 patients ('normal', 16%-61%; 'abnormal-probably benign', 13%-47%; 'indeterminate', 8%-33%; and 'abnormal-suspicious for cancer', 9%-25%). The radiologists' recommendation to biopsy also varied, from 9% to 31%. For diagnostic interpretations there was moderate agreement, with a median pairwise weighted kappa of 0.47 and a median weighted percent agreement of 78%. Diagnostic sensitivity ranged from 37% for one radiologist to 85% for another. High sensitivity values were often accompanied by frequent recommendations for immediate work-up in patients who did not have cancer. When noting the most suspicious lesion on a mammogram, the radiologists disagreed on side (right vs. left) in 42% of patients, with variability declining as interpretations became increasingly more suspicious. Major interpretive disagreements, in which a mammogram was called 'normal' by one radiologist and 'abnormal-suspicious for cancer' by another, occurred in 19% of the 150 patients. Major management disagreements, defined as a recommendation for 'routine follow-up only' by one radiologist compared to a biopsy recommendation by another, occurred in 25% of patients. When the same set of 150 mammograms was re-interpreted five months later, intra-observer readings showed better consistency. Later, when assembled in a conference to review disagreements, the radiologists found that the main problems arose from differences in visual perceptions,

characterization of abnormalities and thresholds of concern. Active measures to reduce this variability are warranted and will likely necessitate extensive collaborative efforts for a goal of standardization coordinated with accuracy.

"There are only two lasting bequests we can hope to give our children.
One of these is roots, the other wings."

Hodding Carter

Thank you to my wonderful family for giving me the courage to fly.

This thesis is dedicated to the memory of my grandparents
Dr. Abraham and Sylvia Miller
who exemplified the ideal collaboration in marriage, family and medicine.
Their loving memory will always be a source of inspiration.

ACKNOWLEDGEMENTS

The research for this project was conducted during the years 1991-1994. This thesis is the result of hundreds of long hours of planning and analysis and would have been totally impossible without the cooperation and dedication of many people.

I gratefully acknowledge:

my Thesis Adviser, Dr. Joann G. Elmore, for her clarity of thought, her boundless energy and her completely unselfish guidance.

Dr. Alvan R. Feinstein for his keen intellect and sharp wit and his unique ability to share his enthusiasm for clinical research.

Carolyn K. Wells for her statistical expertise and her continuing support.

Dr. Carol H. Lee for her mammographic insights and her practical input regarding radiologic practice.

my mother, Bette Howard, for her aid and encouragement in this project and always.

the ten participating radiologists who cooperated within many time constraints to contribute their expertise to this project.

Drs. R. I. Horwitz and J. F. Jekel for their thoughtful comments.

the American Cancer Society for the support of a Research Development Grant (#RD-346).

This study was presented, in part, at the Plenary Session of the Annual Meeting of the Association of American Physicians on May 2, 1993 in Washington, D.C. and has been published in the 1994 Transactions of the Association of American Physicians. (Howard DH, Elmore JG, Lee CH, Wells CK, Feinstein AR. Observer Variability in Mammography. Trans Assoc Am Physicians 1993; CVI: 96-100.)

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

INTRODUCTION

In the practice of medicine, results of diagnostic tests are not always definitive. When physicians reach different conclusions from the same information, the variability can lead to important clinical consequences. For example, in the case of blood pressure, fluctuations in measurement may occur. One reading may result in change in a patient's diet, life style, and/or pharmacological regimen; while for the same patient, a different reading may mean no change at all. An evaluation of an 'acute abdomen' may or may not result in surgery; and an interpretation of a mammogram may be the difference between inordinate anxiety and breast biopsy or calming reassurance. This kind of inconsistency is referred to as observer variability.

This thesis investigates observer variability in mammography. In order to better understand the phenomenon, this introduction will first review the importance of mammography in breast cancer screening and, then, observer variability in general. In addition, a review of previous studies of observer variability, specifically in mammography, will aid in placing this investigation in an appropriate context.

BREAST CANCER AND MAMMOGRAPHY

It is currently recommended by the American Cancer Society (ACS),

American College of Radiology (ACR), and National Cancer Institute (NCI) that women begin having regular mammograms every one to two years, at age 40, and every year after age 50 (1). These current guidelines affect millions of adult women annually. Current statistics show that approximately one in every nine women in the United States will develop breast cancer in her lifetime. It was projected, for 1993, that approximately 180,000 women would be diagnosed as having this disease and 45,000 of those women would die from it (2). Until breast cancer can be prevented, the most effective way for women to protect themselves is through early detection and prompt treatment.

Evidence that mammography can reduce mortality from breast cancer is supported by several investigations (3-10). The first study to show a true benefit from screening for breast cancer occurred in 1963, undertaken by the Health Insurance Plan of New York (HIP) (9). The HIP study randomly assigned 62,000 women into two groups: a study group which offered screening on an annual basis for four years and a control group which offered no screening. After seven years of follow-up, the participants in the study group had a reduction in mortality of 23% when compared with the controls. In addition, this study also found that women with cancers detected by mammography alone had a better five year survival rate than those diagnosed by other modalities. The success of the HIP trial inspired the ACS and NCI to implement programs of mammogram screening (10).

There have been numerous technological advances in breast imaging over the past 20 years (11). In addition, since 1987, in an effort to ensure that current technical standards of mammography are maintained, there have been rigorous accreditation requirements set by the ACR (12). For example, equipment used for mammography at accredited sites must be designed specifically for mammography. Film processors must have developer time and temperature settings appropriate for the specific type of mammographic film being used. Procedure manuals and logs must be maintained in compliance with guidelines. Performance standards also must be monitored by assessments of image quality.

There are additional requirements by the ACR for the professional interpreter as well. The physician must be certified in diagnostic radiology by the American Board of Radiology or by the American Osteopathic Board of Radiology, or have received two months of full-time documented formal training in the interpretation of mammograms. It is recommended (not required) that the physician interpret a minimum of 480 mammograms per year. There also must be participation in continuing education, specifically in mammography, with at least 40 hours of documented credits prior to accreditation and at least 15 hours every three years thereafter. These requirements, however, do not evaluate the extent of radiologists' variability or accuracy in interpretation.

OBSERVER VARIABILITY: CLINICAL MEDICINE

The phenomenon of observer variability has been studied in many areas of clinical medicine (e.g., psychiatry, pathology, physical diagnosis and radiology) (13-14). Research in radiology, for example, has demonstrated that experienced radiologists can contradict each other in their interpretation of radiographs (inter-observer variability) and the same radiologists can even be inconsistent in interpreting the same film twice (intra-observer variability) (15-16). The clinical implications of these investigations are apparent in that radiologists failed to identify a substantial proportion of cases with roentgenographic evidence of tuberculosis and frequently changed their readings of positive cases upon later re-reading the same films (15). Similar variation has been reported in the interpretation of hepatic scintigrams (17) and coronary angiograms (18-19).

Observer variability has been well-researched in many radiologic procedures, but it has not been extensively studied in mammography, an area in which it could be of particular importance. In other areas of radiology, results are interpreted in conjunction with additional clinical and diagnostic information; therefore, the radiologic results often are not the sole basis for patient management. In mammographic screening, however, the premise is to detect signs of breast cancer prior to other evidence of disease (e.g., a palpable mass). A recommendation for biopsy may depend heavily on the

mammogram reading; therefore, any interpretive variability can directly affect the patient management plan (20).

OBSERVER VARIABILITY: MAMMOGRAPHY

One of the earliest investigations on observer variability in mammography was done in 1975. Chamberlain and colleagues studied the validity and observer variability of clinical examination and mammography as screening tests for breast cancer (21). The study enrolled 1,215 women over the age of 40 whose individual mammograms were read by two consultant radiologists. The two radiologists agreed on the need for surgical referral in only one-half of the patients they referred.

In a subsequent study on xeromammograms (1982), Boyd and colleagues examined observer variability among nine participating radiologists in the Canadian National Breast Screening Study (NBSS) (20). One hundred xeromammograms were selected for review. Results showed that the radiologists varied substantially in their diagnoses. Agreement on specific diagnostic categories was best for the definitive diagnosis of cancer; least for the diagnosis of benign abnormalities; and intermediate for the diagnosis of normality or suspicion of cancer. Radiologists were also asked to classify films into one of four mammographic patterns (22), each image pattern being associated with varying cancer risks as developed by Wolfe (23). The

radiologists frequently disagreed as to whether or not a particular mammogram could be classified in one of the four patterns, and some radiologists did not classify all mammograms. Of the films which all radiologists did classify, however, general agreement improved.

On review of Boyd's study, it is noteworthy that xeromammography, a technical predecessor to current mammography, is rarely used today. In addition, two of the nine participating radiologists were each responsible for selecting one-half of the study films. Therefore, they were not 'blinded' to the subsequent cancer outcome of each of their selected cases and their interpretations may have been affected. Copies rather than original films were used allowing for some degradation of the image quality and the possibility of affecting interpretation. Furthermore, any xeromammogram considered of inadequate quality by any one of the radiologists was excluded from the analysis. One radiologist designated 18 of the 100 films unsatisfactory. This factor reduced the data pool and may have created a bias in the remaining films.

Baines and colleagues conducted a later investigation involving films from the Canadian National Breast Screeing Study (NBSS) (24). The NBSS, a randomized controlled trial (1981-1988) which recruited 89,835 women in 15 screening centers across Canada, assessed the effect of screening on breast

cancer mortality. One reference radiologist was appointed to continuously audit the study. The random review of films by this reference radiologist was an attempt to monitor the quality of mammograms. Comparisons of interpretations between the reference radiologist and the center radiologists allowed for assessment of inter-observer variability; and by comparing those interpretations with subsequent cancer outcomes, accuracy also was assessed. Results showed that of 5,200 cases known not to have cancer, there was agreement in 75.8% of them. For women with screening-detected cancer (i.e., histologically-proven), agreement occurred in 85.6% of cases. The investigators concluded that for screening mammograms, there was delayed detection of breast cancer in 17% of cases due to observer variability and in 5% of cases due to sub-optimal technical quality.

Variability in mammographic interpretation was also documented in a study by investigators in Turin, Italy (1988) (25). Eight radiologists reviewed 45 cases comprised of nine with histologically-proven cancer; 25 with benign disease, as diagnosed through fine needle aspiration; and 11 with normal breasts, according to two radiologists. The films used were copies obtained from the Canadian National Breast Screening Study. Statistical indices of variability between the radiologists were comparable to those reported by Boyd et al (20).

On review, the Turin investigation was a well-designed study. Supervising radiologists did not participate and the participants were representative of all but one of the local public institutions where mammography was performed. Radiologists were 'blinded' to the subsequent cancer outcome of all cases. Study films, however, were copies taken prior to 1985 and supplied by the NBSS. NBSS films have been criticized for their technical inadequacy in that time period (26). After 1985, an NBSS protocol change (mediolateral oblique positioning replaced straight mediolateral positioning) allowed for significant technical improvement (26).

In a follow-up study by the Turin investigators, films were re-read by the same radiologists two years later (27). As expected, variability was higher among the radiologists than between two readings by the same radiologist. Comparisons between the two studies were diminished because the radiologists had been 'unblinded' to the 'gold standard' distribution of mammograms after the first study; and within the two year interim, the standard of recommended management had shifted from surgical biopsy to fine needle aspiration due to increased availability of low-cost instruments.

GOALS OF THESIS

The goals of this thesis are to determine whether, with the newer technical methods and increased utilization of mammography today, the

amount of variability between and among radiologists has changed. Secondly, this thesis investigates the variability of radiologists' accuracy in mammographic interpretation and, finally, attempts to identify the sources of that variability.

METHODS

CASE SELECTION

Mammograms for this study were selected from those done at Yale-New Haven Hospital (YNHH) in 1987. One hundred and fifty patients were chosen using a stratified random sampling technique. Cases covered a spectrum of three diagnostic interpretations ('normal', 'abnormal-probably benign', and 'abnormal-suspicious for cancer') which are common mammographic categories used in clinical practice. The mammograms also had a concomitant 'gold standard' designation of 'cancer' or 'non-cancer'. The initial sampling year, 1987, was chosen in order to allow adequate follow-up time and to ensure that the designated 'gold standard' cancer outcomes were representative of the true diagnoses.

The 'gold standard' requirements enabled the categorization of films into subsequent 'cancer' and 'non-cancer' cases. The designation of breast cancer

was given to a case if the diagnosis was histopathologically confirmed at YNHH within three years of the 1987 mammogram. A designation of 'non-cancer' was given if the following criteria were met: the patient did not have histologically-confirmed breast cancer at YNHH after the 1987 mammogram and a diagnosis of 'normal' or 'abnormal-probably benign' was given to a follow-up mammogram taken in 1990. At the time of the follow-up mammogram in 1990, the patient completed a personal history form. On that form, a further requirement was the specification of a negative history of breast cancer. This ensured that the patient had not been diagnosed with cancer, in the interim, at a hospital other than YNHH. For example, if a patient had a negative biopsy which was performed as the result of the 1987 mammogram or within the three year interim following, she still required a 'normal' or 'abnormal-probably benign' 1990 mammogram. Also, no cancer could be specified on the 1990 patient history form in order to be designated as 'non-cancer' for the purposes of this study.

Patients were ineligible for this study for the following reasons: previous diagnosis of breast cancer before the 1987 mammogram; no definitive interpretation for the 1987 film (e.g., film labeled as 'abnormal-indeterminate' pending additional studies such as magnification views or ultrasound); past cosmetic breast surgery (reduction or augmentation); large breast size requiring larger film size or multiple films per view; a designated abnormality

visible only on lateral view; presence of a marker placed on patient's skin; unavailable films; and films regarded as having inadequate technical quality.

Some of the above criteria bear explanation. Patients with a history of breast cancer were ineligible because they were already on a different follow-up protocol influenced by their pathology results. By definition, they would not be part of a screening population and an 'inception cohort' was desired for this study. Regarding patients with previous cosmetic surgery, mammograms should be read with full knowledge of the surgery and the patient's clinical history was not provided for all cases in this study. Patients with large breast size were excluded because an increase in the number of study films required per patient would have become logistically difficult. If a critical abnormality was noted only on a lateral film the case was excluded because lateral views were not included in this study. (Only mediolateral-oblique and craniocaudal views were used.) The mammograms for each case were reviewed for technical quality by the coordinating radiologist (C. H. Lee) who was 'blinded' to the patient's identity, 1987 diagnostic interpretation, and 'gold standard' cancer outcome. This quality review was implemented because mammograms with technical problems can obscure cancer detection (24). Once a mammogram was included in the study, a participating radiologist did not have the option of labelling a film 'technically unsatisfactory'.

Mammograms had been done with standard film-screen technique on a Thompson CGR 500T Unit using Kodak Ortho M film and Kodak MIN-R screens. The original mammograms, not copies, were used in the testing to ensure maximum image quality. A mediolateral-oblique and craniocaudal view were available for each breast. To maintain confidentiality for patients, all mammograms were coded with numbers specific for the research. The study protocol was approved by the Human Investigation Committee of the Yale University School of Medicine.

Cases were selected from the 4,000 mammograms taken at YNHH in 1987. Following are definitions and an account of the randomized selection process for the three interpretive categories ('normal', 'abnormal-probably benign', and 'abnormal-suspicious for cancer').

a. 'Normal' Category

The definition of 'normal' was applied to those 1987 mammograms which were interpreted as having no significant abnormality. This category included patients with fibrocystic changes and dense breasts. Mammograms with a recommendation for age-appropriate follow-up were included in this category (i.e., no other follow-up recommendation could have been made). Approximately 3,000 cases fulfilled this definition of 'normal'.

From these 3,000 'normal' cases, the study patients were randomly selected. A listing of all patients who received mammograms in 1987 were recorded in a record book. This record book contained 132 pages and was divided into two groups of 66 pages. Each page contained a maximum of 37 lines. A random number table was used to generate three sets of numbers. The first set determined the group of pages from which the patient would be selected (odd = first group and even = second group). The second set of numbers delineated the page from which the patient would be selected, and the third set specified the page line from which the actual case was chosen. If the patient who met the above criteria did not have a 'normal' mammogram interpretation, candidates in line order below the specified position were considered until one was chosen. If the end of the page was reached without obtaining a case, the process was repeated.

As a list was generated, it was entered into the Yale Decrad computer system (Decrad) which contained a file of all radiologic procedures and diagnostic results at YNHH. Individuals were ineligible for this category if they did not have a 1990 follow-up mammogram which was interpreted as 'normal' or 'abnormal-probably benign'. The patients' medical record numbers were also entered into the Yale pathology computer file and were ineligible for the study if any history of breast cancer, prior to 1987, was indicated.

b. 'Abnormal-probably benign' Category

The definition of 'abnormal-probably benign' applied to mammograms from 1987 in which findings (e.g., a mass, calcification, focal asymmetric density, architectural distortion etc.) were of notable concern. Any mammogram with a recommendation for follow-up in less than one year was also included in this category.

The Yale pathology files were used to determine if any of those patients had a positive biopsy for breast cancer at YNHH within three years of the original mammogram. Decrad was also used to determine which patients with no evidence of cancer had a 1990 mammogram interpreted as 'normal' or 'abnormal-probably benign'. This allowed for further sub-categorization into 'non-cancer' and 'cancer' patients.

c. 'Abnormal-suspicious for cancer' Category

The definition of 'abnormal-suspicious for cancer' applied to 1987 mammograms for which a recommendation of biopsy or needle localizaton was made. A mammogram remained in this category even if the above-mentioned procedures were later deemed unnecessary.

The Yale pathology file was used to determine biopsy results. A positive biopsy for breast cancer placed a patient in the sub-category of 'cancer'. If a

patient had a negative biopsy as well as a follow-up mammogram in 1990 with an interpretation of 'normal' or 'abnormal-probably benign', she was placed in the sub-category of 'non-cancer'. If a patient's 1987 mammogram was interpreted as 'abnormal-suspicious for cancer' and no biopsy or needle localization was performed, yet the patient had a follow-up 1990 mammogram which was 'normal' or 'abnormal-probably benign', the patient was also placed in the category of 'non-cancer'.

PARTICIPATING RADIOLOGISTS

The coordinating radiologist (C.H. Lee), a specialist in mammography, assisted in study design and case selection. This radiologist did not participate as one of the ten study radiologists. Radiologists from community and academic practices in Connecticut and New York were invited to participate. Most participants were recruited through professional contact by members of the research team and one radiologist was contacted by a random telephone call through the Connecticut Yellow Pages. Requirements for participation were board certification in diagnostic radiology and a clinical practice that included reading mammograms. A distinct effort was made to select a broad spectrum of radiologists with regard to their experience in the interpretation of mammograms. The hope was that this range of experience would more accurately represent the reality of practicing radiologists reading

mammograms.

The radiologists were 'blinded' to the research objectives, study design, and number of cases in which cancer was subsequently diagnosed. Each radiologist received an honorarium provided by the American Cancer Society which was modest compensation for the total effort.

RESEARCH DESIGN

The research had two phases. In Phase I, the radiologists independently read the 150 patients' mammograms. This was followed by a 'wash-out' period of five months in order to decrease the possibility that the radiologists might recognize the previously-read films. In Phase II, the radiologists re-reviewed the same 150 patients' films. The radiologists were informed that male patients were excluded from the study as well as females with a history of previous breast cancer; but were not informed that the same films would be shown a second time. The radiologists had no time limit for reviewing the films.

Inter-observer variability was assessed for the 150 cases in both Phase I and II. In both phases, 50 cases were shown with each patient's age only, but no other clinical history. These were used to assess intra-observer variability. The remaining 100 cases in each phase were used to determine whether

knowledge of the patient's clinical history affected the mammographic interpretation. The effects of the presence or absence of clinical histories are not discussed in this thesis, but will be analyzed and presented in a subsequent report. Of these 100 cases, 50 were shown with a detailed clinical history in Phase I and with each patient's age only in Phase II. The sequence was reversed for the remaining 50 cases.

Information on patient history was extracted from a form completed by each patient at the time of the 1987 mammogram and transferred to a patient history form used in this study (Appendix I) . The detailed clinical history included signs and symptoms (e.g., palpable lump, nipple discharge, skin change), location of previous breast biopsy, menopausal status, estrogen use, history of other cancers, and family history of breast cancer. In instances where the patient had an abnormality noted on self-examination prior to the 1987 mammogram, the location was illustrated on the patient history form by a radiology technician and this was included on the history form as well. In Phase II, an additional five cases were shown for a third time with a deliberately leading 'sham' clinical history. Again, the results of the 'sham' history will be presented in a subsequent report.

In both phases of testing, the films were arranged in a random sequence and numbered 1-150 and 151-305. For each mammogram, the radiologists

used a check-list form (Appendix III) to indicate observations, diagnostic interpretations and management recommendations. The check-list contents were based on two sources: a routine form used at YNHH for screening mammograms and a standardized lexicon for mammography developed and recommended by the American College of Radiology. Before Phase I, the check-list form was reviewed with the study radiologists who each received a brief written summary of instructions for its use (Appendix V). The radiologists were asked to try to simulate their own clinical practice in making diagnostic interpretations and follow-up recommendations.

In descriptive observations for each mammogram, the radiologists noted the presence of specific abnormalities (e.g., a mass or calcification) and the location (right or left breast as well as a standard twelve-hour clock demarcation). In the presence of two or more abnormalities in the same patient, the radiologists were asked to note the two which were most suspicious. These were further delineated by noting which of those two was more suspicious (i.e., evoking the most concern about possible cancer).

The diagnostic interpretations could be chosen from one of four categories: 'normal', 'abnormal-probably benign', 'abnormal-indeterminate' (when the radiologist felt uncertain about the diagnosis), or 'abnormal-suspicious for cancer'. The radiologists were asked to avoid the category

'abnormal-indeterminate' whenever possible.

Management recommendations could be chosen from: routine mammographic follow-up (i.e., follow-up screening mammogram after one or two years according to the patient's age); short interval mammographic follow-up (e.g., within six months); or immediate work-up (defined as additional mammographic views, ultrasound and/or biopsy). The radiologists were instructed to note at least one follow-up recommendation and they could recommend as many categories as were applicable.

DATA ANALYSIS

The original data from the patient history forms and the radiologists check list forms were coded numerically (Appendix II and IV), double-entered and verified for electronic coding. Statistical analyses were done either with electronic hand calculators or with programs in the EPI INFO and SAS statistical systems (28-29). The methods of data analysis are reviewed in the next four sections.

a. Radiologists' Observations, Interpretations and Management Recommendations

The range of variability among the ten radiologists was assessed for observations, diagnostic interpretations and management recommendations for

Phases I and II. The range illustrates the lowest and highest percentage of times that one radiologist used a specific category (from the 150 cases) compared to the percentage of times that another radiologist used the same category. The median percentage of times that a category was noted was also assessed.

b. Agreement in Diagnostic Interpretations and Recommendations

Inter-observer agreement on diagnostic interpretations and recommendations for the 150 cases was calculated from the readings among pairs of radiologists for Phases I and II. Intra-observer variability was assessed for the 50 mammograms that appeared in both phases with the age of the patient as the only history. The statistical indices of observer variability in paired comparisons were percentage agreement and the kappa statistic (30). Both indices were weighted because not all disagreements were considered to be of the same magnitude. For example, a situation in which one radiologist diagnosed a case as being 'normal' and another diagnosed the same case as 'abnormal-probably benign' was considered a less important disagreement than if one radiologist diagnosed 'normal' and another, 'abnormal-suspicious for cancer'. By using weighted statistics, these variations in the magnitude of disagreement could be taken into account when determining the overall agreement among the radiologists.

An example of two radiologists' diagnostic interpretations for the same 150 patients is shown in Table I. This can serve as an example of the appraisal of pairwise comparisons of interpretations by two radiologists. The numbers on the shaded diagonal indicate interpretive agreement between Radiologist A and Radiologist B. The two radiologists agreed on the interpretation of 'normal' in 34 of the 150 patients and, at the other extreme, on the interpretation of 'abnormal-suspicious for cancer' in 15 patients. The two radiologists did not agree on the mammograms outside of the shaded diagonal. For example, when Radiologist A interpreted 'normal' in 64 patients, Radiologist B agreed in only 34, interpreting 18 of those 'abnormal-probably benign', 10 'abnormal-indeterminate', and 2 as 'abnormal-suspicious for cancer'.

Percentage agreement is calculated by adding the total number of cases which lie on the shaded diagonal (indicating perfect agreement) and, then, dividing this sum by the total number of cases. Thus, the percentage of perfect agreement in this example is [(34+10+12+15/150]x100=47%. However, to allow for differences in the degree of importance of disagreements, a weighted calculation is used. The weighted agreement is calculated by multiplying the number in each cell by a specified weight (i.e., 1, 0.66, 0.33, or 0) (28) and adding all the products. This sum is divided by the total number of cases. Thus, weighted agreement for Table 1 is

$[\{[(34+10+12+15)x1]+[(18+23+6+3+6+4)x0.66]+[(10+1+1+5)x0.33]+[(2+0)x0]\}/150]x100=78\%$.

Some agreement of interpretations of mammograms can be expected by chance alone. The kappa statistic corrects the amount of agreement observed for the agreement that is expected by chance. The formula for weighted kappa is:

$K_w = PO_w - PC_w / 1 - PC_w$

where PO_w is the proportion of weighted agreement (calculation shown above) and PC_w is the proportion of weighted agreement expected by chance. To calculate a weighted kappa statistic, this weighted agreement expected by chance must be derived. This is calculated by multiplying the row totals (Radiologist A's interpretations) by the column totals (Radiologist B's interpretations) and dividing by the total number of cases. This gives a calculated value for each cell. This value is then multiplied by a cell weight, the products added, and that sum divided by the total number of cases. For Table 1, the percent weighted agreement expected by chance is 59.5%. For the example shown in Table 1, kappa value is $K_w = 0.78 - 0.595 / 1- 0.595 = 0.46$.

To interpret the weighted kappa statistic, a score of 1.0 indicates perfect agreement and a score of ≤ 0.0 indicates agreement which is no better than expected by chance. Landis and Koch have suggested the following ratings of

agreement for values of kappa: <0, poor; 0 - .2, slight; .21 - .40, fair; .41 - .60, moderate; .61 - .80, substantial; and .81 - 1.00, almost perfect (31).

c. Accuracy

The radiologists' accuracy (sensitivity and specificity) for the diagnosis of cancer was determined for the category 'abnormal-suspicious for cancer' vs. a combination of all other diagnostic categories ('normal', 'abnormal-probably benign', and 'abnormal-indeterminate'). The percentage of patients for whom each radiologist recommended immediate work-up, defined as a recommendation for additional mammogram views, ultrasound, and/or biopsy, was evaluated for both the 'cancer' and the 'non-cancer' patients.

Specific attributes of the participants were correlated with accuracy to determine if any relationship existed. Attributes examined include: clinical practice type (academic vs. private); years of radiologic experience; percent of time spent ıeading mammograms; and whether or not the participant considered him/herself an 'expert' in mammography. The statistical index used was the Pearson correlation coefficient (28).

d. Major Disagreements in Locations of Abnormalities, Interpretations, and Recommendations

For any pair of two radiologists reading the same patient's mammogram, major disagreements were defined as follows:

1) Location of abnormalities: The most significant abnormal lesion was said to be located in the right breast by one radiologist and in the left breast by another. A citation of 'bilateral' by one radiologist and a 'specific side' by another radiologist was not counted as a disagreement. Cases in which only one radiologist noted a lesion were excluded from this tabulation.

2) Diagnostic interpretations: The same patient's films were called 'normal' by one radiologist and 'abnormal-suspicious for cancer' by another.

3) Management recommendations: For the same patient, routine mammographic follow-up was proposed by one radiologist and biopsy by another.

The proportions of major clinical disagreements were calculated in two ways: first, for their occurrence within the 150 patients (per-patient comparison); and second, for their occurrence within the maximum number of pairwise comparisons (per-pairwise comparison).

For the per-patient calculations, the numerator was the number of patients in whom at least two radiologists had a major disagreement in location of abnormality, diagnostic interpretation or management recommendation. The denominator for these calculations was 150 patients for the diagnostic

interpretations and management recommendations. For calculations of disagreement in the location of abnormalities, the denominators varied since abnormalities were not noted in all mammograms. The denominators for these calculations were the number of patients (or pairwise comparisons) in whom at least two radiologists had noted an abnormality. The percentage of patients for whom disagreements in location of abnormalities occurred, was calculated when: a) any abnormality was noted, b) an immediate work-up was recommended, and 3) a biopsy was recommended.

For the per-pairwise calculation, the numerator is the number of pairwise disagreements noted and the denominator is the maximum number of possible pairwise comparisons. The maximum number of pairwise comparisons for 10 radiologists is 45 (= 10 x 9/2) pairs for an individual patient, and 6,750 (= 45 x 150) pairs for the entire series of 150 patients.

ASSESSMENT OF SOURCES OF VARIABILITY

After the Phase II testing, and with preliminary results available for the Phase I analysis, the participating radiologists were assembled for a conference at which they were 'unblinded' to the study goals and research design. Preliminary data from Phase I was presented. Six cases were reviewed which were representative of great variability in interpretation. The radiologists received copies of their personal check-list forms for the cases

discussed. The goal of the discussion was to identify causes of the variability which had been noted and to understand better at what level the variability was occurring.

RESULTS

SELECTION OF CASES

Cases were selected from the 4,000 mammograms taken at YNHH in 1987 which had been interpreted as 'normal', 'abnormal-probably benign' and 'abnormal-suspicious for cancer'. Each of the three categories was partitioned further into 'cancer' and 'non-cancer' outcomes based on the 'gold standard' definition. Originally, a goal of 50 mammograms in each of the three interpretive categories was desired. In addition, it was hoped that approximately one-half of the 'abnormal - probably benign' category and one-half of the 'abnormal - suspicious for cancer' category would be 'cancer' cases. This would have allowed for assessment of variability in the more complicated cases while trying to adhere to percentages seen in clinical practice. Because of eligibility and exclusion criteria, especially in the 'abnormal - suspicious for cancer' category, the final study population was different than predicted. The following three paragraphs are an account of the results of the selection process for each of the three diagnostic categories.

There were approximately 3,000 mammogram cases interpreted as 'normal' in 1987 at YNHH of which 275 cases were randomly selected for this study. Of those 275 cases, 89 had appropriate 1990 follow-up (i.e., a mammogram interpreted as 'normal' or 'abnormal-probably benign') and of those 89 cases, 36 did not meet other eligibility criteria (Table 2). One patient with a 'normal' 1987 mammogram which was randomized into the group of 275 cases was found to have subsequent breast cancer within three years of the original mammogram. A final group of 54 'normal' category patients (one with subsequent breast cancer) completed this category.

Criteria for the category 'abnormal-probably benign' applied to 567 patients in 1987. Subsequent breast cancer was found in 17 patients, 10 of whom did not meet eligibility criteria. Of the 550 remaining 'abnormal-probably benign' category cases, 190 had appropriate follow-up and 93 were chosen randomly. Of those 93, eligibility criteria were not met by 39. A final group of 61 patients (seven with subsequent breast cancer) comprised the 'abnormal-probably benign' category.

The criteria for the category 'abnormal-suspicious for cancer' applied to 124 patients in 1987. Biopsy-proven breast cancer was found in 43 patients (before 1990) and 24 of these cases did not meet eligibility criteria. Of the remaining 81 'abnormal-suspicious for cancer' category patients who did not

have histologically-confirmed cancer at YNHH, 29 had appropriate 1990 follow-up. Eligibility criteria were not met by 13 patients. A final group of 16 patients was obtained bringing the 'abnormal-suspicious for cancer' category to a total of 35 cases (19 'cancer' patients).

In summary, within the final study population of 150 patients, the diagnostic interpretations from 1987 (and subsequent breast cancer status) were as follows: 54 'normal' (1 'cancer' patient); 61 'abnormal-probably benign' (7 'cancer' patients); and 35 'abnormal-suspicious for cancer' (19 'cancer' patients).

The 150 patients included 95 (63%) who were at least 50 years of age (range of 33 - 83 years) (Table 3). Twenty-two (15%) patients were symptomatic (e.g., noted a lump or nipple discharge) and 40 (27%) had a past benign breast biopsy. Forty-four (29%) patients had a relative with breast cancer. Breast cancer was diagnosed within three years after the 1987 mammogram in 27 (18%) patients; and of those 27 patients, 22 were diagnosed within the first year.

Of the 27 study patients diagnosed with breast cancer, many histological types were represented. Twenty-three patients had ductal carcinoma (intraductal, 6; infiltrating, 7; intraductal/infiltrating, 1; mucinous, 1; papillary, 1;

and intraductal/papillary, 1). Lobular carcinoma was diagnosed in three patients (infiltrating, 2; infiltrating/in situ, 1). Combined ductal and lobular carcinoma was seen in one patient.

PARTICIPATING RADIOLOGISTS

Of the ten radiologists who participated in this study (Table 4), seven were in private practice in New York or Connecticut and three held full-time academic positions. A total of 15 radiologists, known to members of the research team through previous professional associations, were contacted. Of these, nine agreed to participate in the study. One participant was contacted, on the first attempt, through a random telephone call using the New Haven Yellow Pages. The group's clinical experience, defined as the number of years in practice, had a median of 12 years (range, 5 -30) and the group had a median of 3.5 years (range, 1.5 - 20) of specific experience in reading mammograms. The median estimated number of mammograms interpreted in the year before the study was 1900 (range, 200 - 6,000); and the median percentage of time spent reading mammograms in clinical practice was 25% (range, 8% - 50%).

After Phase II of the study but prior to being 'unblinded', a telephone survey of the participants was conducted to gain additional information

(Appendix VI). Without defining the term 'expert', in the telephone survey, three radiologists considered themselves an 'expert' in mammography. Seven radiologists said they were not aware that the same mammograms had been shown in Phases I and II. Of the remaining three radiologists, two said that they recognized less than 3% of the mammograms in Phase II and one claimed to have recognized about 25%. When asked if they had given the study mammograms the same consideration as in their clinical practices, seven answered "yes"; two answered "I tried"; and one replied "no" (implying that less consideration was given the study films than in his/her clinical practice). Although this radiologist answered "no", results of sensitivity and specificity for this radiologist were on par with the other nine participants. When the radiologists were asked in what percentage of mammogram cases they thought variability became clinically important, responses ranged from 1% to 40% with a median of 15%.

OBSERVER VARIABILITY

a. Radiologists' Observations, Interpretations and Management Recommendations

The median and range of the percentage of patients in whom the radiologists noted the indicated observations, diagnostic interpretations and management recommendations varied among the 10 radiologists (Tables 5,6).

For example, in Phase I (Table 5) among the 150 patients, the median percent of patients with a mass noted was 26%. The range for the 10 radiologists was quite wide, with a mass being noted in only 19% of the cases by one radiologist and, at the other extreme, in 40% by another. The use of the diagnostic interpretation of 'normal' had a median of 31% with a range of 16% - 61%. For management recommendations in Phase I, one radiologist recommended routine mammographic follow-up in only 22% of the cases compared with 71% by another, and the recommendation for biopsy ranged from a low of 9% to a high of 31%. Phase II results were similar (Table 6). It should be noted that these percentages do not necessarily refer to the same patients. For example, two radiologists may have recommended a biopsy in 20% of patients, but this did not mean that the biopsy recommendations were necessarily made on the same patients.

b. Agreement on Diagnostic Interpretations and Recommendations

All 10 radiologists agreed on a patient's specific diagnostic interpretation in only nine (6%) cases. These nine cases were interpreted by all 10 radiologists as either 'normal' or 'abnormal-suspicious for cancer'. At least five of the 10 radiologists agreed in 121 (87%) cases. In most of these cases, the interpretations were also 'normal' or 'abnormal-suspicious for cancer'.

The agreement among the 10 radiologists on diagnostic interpretation

and biopsy recommendations was assessed (Tables 7,8). The statistical indices of concordance were weighted for the four ordinal interpretive categories, but not for the binary (yes/no) biopsy categories. For Phase I (Table 7), the median percentage agreement was 78% (range, 71% - 82%) for inter-observer variability in interpretations, and 85% (range, 65% - 91%) for recommendations to biopsy. The corresponding median kappa values were 0.47 (range, 0.31 - 0.55) for diagnostic interpretations and 0.49 (range, 0.20 - 0.69) for biopsy recommendations. Landis and Koch (31) would assign these kappa values a rating of 'moderate' agreement. Phase II results were essentially similar (Table 8). The corresponding results were higher for intra-observer than for inter-observer agreement.

c. Accuracy

There was considerable variability in accuracy among the 10 radiologists in diagnosing the 27 patients with 'cancer' (Table 9). For Phase I, interpretations ranged from Radiologist A, who called 23/27 patients 'abnormal-suspicious for cancer' and none 'normal', to Radiologist J who called only 10/27 patients 'abnormal-suspicious for cancer' and interpreted six as 'normal'.

Sensitivity for diagnosing cancer was highest (85%) for Radiologist A and lowest (37%) for Radiologist J (Table 10). The values of sensitivity generally declined as specificity rose. This inverse relationship was most evident in the recommendations for immediate work-up (Table 10). This trend

was also seen in Phase II analysis (Table 11).

An analysis was performed in order to examine any significant correlation between specific physician attributes and accuracy (sensitivity and specificity) of interpretation. Attributes examined were clinical practice (academic vs. private), years of radiologic experience (since completion of residency training), percent time spent reading mammograms (in the year prior to this study), and whether or not the participant considered him/herself an 'expert' in mammography. No significant correlation was found for any of the above. There was a promising correlation between the percent time spent reading mammograms and interpretive sensitivity (i.e., the more time a radiologist spent reading mammograms, the higher his/her sensitivity/specificity values). However, since only 10 radiologists were studied, it was not statistically significant (r = 0.09, 0.33).

d. Disagreements in Locations of Abnormalities

When two or more radiologists noted an abnormality, disagreement on side (right vs. left) occurred in 42% of per patient comparisons and 12% of per pairwise comparisons for Phase I (Table 12). In cases where immediate work-up was recommended, disagreement on the side of the abnormality occurred in 33% of patients and 9% of pairwise comparisons. There was disagreement in biopsy recommendations, as to right or left side, in 9% of patients and in 2%

of pairwise comparisons. For example, there was more agreement on which side to biopsy than which side required immediate work-up. Although this variability was considerable, it declined as interpretations became increasingly more suspicious. Phase II results are similar (Table 13). Surprisingly, intra-observer analysis (the frequency with which a radiologist disagreed with his/her own specification of the side with the most significant abnormality) showed disagreement in 14% of patients and in 3% of per pairwise comparisons for recommendations for biopsy. These values are actually higher than the inter-observer calculations.

e. Major Disagreements in Interpretations and Recommendations

In Phase I, of the 150 patients receiving readings from 10 radiologists, a major interpretive disagreement occurred in 19% of patients and 2% of the pairwise comparisons (Table 14). Major management disagreements occurred in 25% of patients and 3% of the pairwise comparisons. Once again, the intra-observer comparisons showed better agreement. Phase II results are similar (Table 15).

EXPLANATIONS FOR VARIABILITY

At the conclusion of Phase II testing, a conference attended by eight of the ten participating radiologists was convened. Participants were 'unblinded' to the study goals and research design. Data from Phase I was also

presented. Six mammograms which showed considerable variability in interpretation were reviewed. The conference was tape-recorded and a verbatim manuscript was produced (Appendix VII). The following excerpts illustrate the discussion of one case. The radiologists are listed by number in the order in which comments were made..

The 10 radiologists' ranges of diagnostic interpretations and management recommendations for this case (subsequently found to be 'non-cancer') were:

INTERPRETATIONS	# RADIOLOGISTS
Normal	3
Abnormal-probably benign	1
Indeterminate	4
Abnormal-suspicious for cancer	2

MANAGEMENT RECOMMENDATIONS	# RADIOLOGISTS
Age-appropriate follow-up	3
Repeat mammogram$\leq$ 6 months	1
Additional x-ray views now	1
Ultrasound	0
Biopsy	2

Radiologist #1: "What bothered my eye was the asymmetry in the periareola area....The thickening in the left nipple caught my eye...when I look at a mammogram, I look for some degree of symmetry...I was worried...I read it as malignant. I thought it was Paget's, frankly...I didn't see anything in the right breast that bothered me. I thought there were benign calcifications."

Radiologist #6: "I agree." *(calcifications)*

Radiologist #2: "I called this normal and I was fairly shocked because when you put the film up, I said, 'look at that'...I didn't see it...This is the thing that you hope doesn't happen to you in your practice."

Radiologist #3: "There is some range of asymmetry that is tolerable...This passes my threshold for asymmetry."

Radiologist #4: "I made no comment about that *(asymmetry).* I dwelled on the calcifications. Elsewhere in the breast I would consider that significant *(asymmetry)*, but the areola is the one area where there can be both much more thickening and asymmetry...If you think that is abnormal, then I don't think I would follow it...you'd either call it normal or punch biopsy the skin. Those are the only choices."

Radiologist #5: "I thought some of these were irregular *(calcifications)...*I was worried about them"....*(Later)* "It's a few of them, it's not all of them...I don't know how many you have to have before you worry about them. I wouldn't call these benign, I know that."

Discussion Leader #1: "If I said, O.K., picture this...calcifications that are branching, irregularly shaped, everybody would agree that those are suspicious. The question is when you look at this, some people come up with, 'these fit those criteria and therefore are suspicious' and others say, 'these do not fit those criteria and therefore are not suspicious.'"

Radiologist #7: "....there was certainly asymmetry between the appearance of both breasts, and the periareola skin thickening in the left breast

was worrisome, but 'indeterminate' in my eyes....I wanted more films...and plus bring her back in 3 months."

Upon careful review of the conference, three main sources of variability were identified. The first was differences in visual perception. Not surprisingly, subtle findings were sometimes missed (as in the above example by Radiologist 2). In other instances, some radiologists noted masses where others saw only normal parenchyma.

A second source of variability arose from different perceptions about attributes of the same abnormality (as illustrated by the comments of Discussion Leader #1). Although agreeing that the mammogram showed calcifications and also agreeing on the criteria for 'benign' and 'suspicious' calcifications, there was still disagreement as to whether the calcifications in question fit 'benign' or 'suspicious' criteria.

The third main source of variability was different thresholds of concern about perceived abnormalities correlating to variability in management recommendations. For example, Radiologist #1 was "worried" by the asymmetry which (s)he noted in the periareola area and felt that it was malignant. Radiologist #4, while agreeing that there was nipple asymmetry, was not concerned by it, feeling that the areola is one area where thickening

and asymmetry can be tolerated.

A feeling of great interest and spirited cooperation pervaded the conference. Results of the discussions indicated that at least part of this variability may be reducible. This would require improvement of diagnostic criteria and thresholds of concern through extensive collaborative effort. The evening concluded with a presentation of certificates of appreciation to each radiologist for his/her support and assistance in this thesis project (Appendix VIII).

DISCUSSION

This investigation found substantial variability among 10 radiologists in mammographic diagnostic interpretations, management recommendations and clinical accuracy. The design of this study attempted to avoid the problems of previous studies of observer variability. The films used in the study employed relatively current techniques of mammographic imaging. Although there have been great advances in breast imaging over the past twenty years, there have been no major developments since 1987, the year from which the study mammograms were taken. The participating radiologists read original, not duplicate, films; therefore the concern about image degradation was negated. Following the random selection of films, the coordinating mammographer (in a

'blinded' review) screened all mammograms to eliminate those of inadequate technical quality. This minimized the risk of cancer detection being obscured.

The participants were 'blinded' to research goals and study design. They were also unaware of the diagnostic distribution of the films and subsequent outcome of each patient's status for breast cancer. However, one radiologist claimed to recognize approximately 25% of study cases from previous clinical experience at YNHH. Nevertheless, the considerable degree of variability found in the diagnostic interpretations and management recommendations are similar to the findings of previous studies (20-22, 24-27) (Table 16).

Careful attention was paid to the selection of the participating radiologists. Radiologists were sought who were currently reading mammograms in their clinical practices, yet had varied experience in mammography. This was an attempt to represent more accurately the professional community currently reading mammograms. Also, the differences in experience among the radiologists could be analyzed to see if a correlation with accuracy existed. Even though there were ten radiologists, too small a group to test statistical significance, the group was not too small to note trends and a trend was noted in the percent time spent reading mammograms and sensitivity/specificity values.

One of the most interesting challenges of this project was the decision concerning methods for analyzing results of disagreement. Analysis was conducted by two methods: per patient and per pairwise comparisons. Together, these calculations probably present the best overall sense of the boundaries of the disagreement. In the per patient comparisons, the variability may have been over-estimated as comparisons were dependent on the number of participating radiologists. For example, a major disagreement occurred if any two of the radiologists disagreed. It was possible for nine radiologists to be in total agreement, but one dissenting radiologist could cause a particular case to be categorized as a major disagreement. Clearly, the greater the number of participating radiologists, the greater the chance of major disagreement and variability. On review of this analysis, no one radiologist was consistently a minority dissenter. In other words, there was no 'oddball' reader. In the per pairwise comparisons, variability could have been under-estimated. The per pairwise disagreements were determined with a denominator (45 x 150 = 6,750) which is the maximum number of pairwise agreements. This number was used as it is logical and easy to explain; however, it is actually twice the number of possible pairwise disagreements which is 3,750 (25 x 150). Therefore, the presented numbers for per pairwise comparisons are actually one-half of what they should be for a true value. For example, major disagreements in interpretations occurred in 4% of the per pairwise comparisons as opposed to 2% previously stated.

Another reason for the possible underestimation of variability is that when observations, diagnostic interpretations and management recommendations were compared among radiologists (Tables 5,6), there were no means to confirm that they were indeed referring to the same abnormality on the same patient's film. For example, two or more radiologists may have noted a mass in 30% of patients, but the lesion of concern may have been in different breasts. Even if two or more radiologists recommended biopsy of the same breast in the same patient, there could have been disagreement as to the specific location of the lesion in question (i.e., one could be commenting on a lesion at one o'clock and the other, a lesion at seven o'clock.)

The category of 'abnormal-indeterminate' presented a challenge as to where and how it should be included in the analysis. A decision was made to maintain those films within the study so as not to reduce the data pool and cause possible bias. Eliminating those films interpreted as 'abnormal-indeterminate' could have led to greater agreement as evidenced by increased kappa values. For example, if mammograms interpreted as 'abnormal-indeterminate' were excluded, the kappa value of Radiologists A and B (Table I) would increase from 0.46 to 0.59. In the determination of sensitivity and specificity, if the diagnostic categories were partitioned into a combination of 'abnormal-suspicious for cancer' and 'abnormal-indeterminate' vs. 'abnormal-probably benign' and 'normal', there may have been a trend toward increased

sensitivity at the expense of decreased specificity (e.g., sensitivity for Radiologist A, as shown in Table 10, would increase from 85% to 89%, but specificity would decrease from 93% to 70%).

INVESTIGATIVE CONSIDERATIONS

No research can perfectly mimic the 'real world' and this study, too, must be viewed with considerations. Although 'blinded' to research goals, and asked to simulate their own clinical practices, the participants in this study were still aware that they were in a 'test' situation. On the one hand, this may have produced a sense of 'over-reading', resulting in abnormalities noted and work-ups recommended which may have been dismissed in daily clinical practice. On the other hand, knowledge that the study interpretations were not directly related to individual patient care may have created less concern and resulted in 'under-reading'.

In this test situation, the radiologists did not have access to previous mammograms for comparison or, in some situations, a complete patient history. They were also limited to two views per breast per patient. Access to this information may have made a difference in results of variability and accuracy. The results of an immediate work-up recommendation (ultrasound, magnification views, etc.) was also not available. If these results had been

immediately available, for example, an 'abnormal-indeterminate' study mammogram may have been changed to an 'abnormal-probably benign' or 'abnormal-suspicious for cancer' category.

In introductory guidelines before Phase I, the participants were asked to limit the use of the 'abnormal-indeterminate' category whenever possible. There was an awareness that too many 'abnormal-indeterminate' interpretations might yield inconclusive results. Some radiologists adhered to this guideline more than others. This request may have forced a premature categorization of mammograms in some cases. At the outset, one radiologist felt "uncomfortable" in limiting the use of this category and, subsequently, used it quite often.

The study population was 'enriched' with difficult cases in the sense that there were more diagnoses of 'abnormal-probably benign' and 'abnormal-suspicious for cancer' (which subsequently proved to have breast cancer). This spectrum of patients might not be found in a random collection of 150 cases of asymptomatic screening mammograms. If the spectrum of study mammograms had reflected the proportions encountered in actual practice, there would have been too few cases to provide any challenges for the radiologists and disagreement may have been diluted. Agreement would have likely increased as would be reflected by higher kappa values and weighted

percent agreements.

CONCLUSIONS

The American College of Radiology is working toward improving standards of mammography. An accreditation program begun in 1987 has specific criteria which encompass both technical performance and professional practice. This accreditation program serves as the only standardized basis for quality assurance. However, this program does not address the variability among individual readers of mammograms. In an indirect way, the development of an ACR lexicon was an attempt to standardize readings of mammograms. In the analysis of this study and in discussion with the radiologists who attended our conference, it became apparent that vocabulary alone would not solve the problem of variability. Even when the radiologists agreed on nomenclature, they had varying levels of disagreement as to whether the nomenclature applied to a specific visual perception. There was additional variability in thresholds of concern provoked by the nomenclature. Finally, even when there was agreement as to both observations and nomenclature, there were still varied management recommendations. For these reasons, it is apparent that further steps are needed to reduce observer variability in mammography. Standardization coordinated with accuracy (not standardization for its own sake) should be an important goal.

The results of this thesis show that radiologists can and do vary, sometimes substantially, in their observations, diagnostic interpretations and management recommendations for mammograms. To reduce this variability, more active steps will be needed than development of nomenclature and dissemination of information regarding visual criteria. Within this project, further study is ongoing to dissect the radiologists' variability in visual perceptions, perceptual designations and threshholds of concern. It is hoped that very specific causes of this variability will be identified e.g., different breast parenchymal densities and use of specific terminology. Collaborative efforts among radiologists also will be required to examine actual performance and to reduce variability while preserving accuracy. In addition, self-auditing procedures by individual radiologists, specialized education and perhaps even specialized accreditation should be considered. Given that millions of women each year are recommended for mammography and its acceptance as a screening tool is continuously increasing, it is of great importance to actively focus on the reduction of observer variability.

REFERENCES

1. Jenks S. ACS keeps mammography guidelines for women under 50. JNCI 1993; 85(5):348-9.

2. U. S. Bureau of Census, Statistical Abstract of the United States:1993 (113th edition) Washington, D.C. 1993.

3. Eddy DM. Screening for breast cancer. Ann Intern Med 1989; 111(5):389-99.

4. Verbeek ALM, Hendriks JHCL, Holland R, Mravunac M, Sturmans F, Day NE. Reduction of breast cancer mortality through mass screening with modern mammography: first results of the Nujmegen project, 1975-1981. Lancet 1984; 1:1222-24.

5. UK Trial of Early Detection of Breast Cancer Group. First results on mortality reduction in the UK trial of early detection of breast cancer. Lancet 1988; 2:411- 6.

6. Tabar L, Fagerberg CJG, Gad A, et al. Reduction in mortality from breast cancer after mass screening with mammography: randomized trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. Lancet 1985; 1:829-32.

7. Andersson I, Aspergren K, Janzon L, et al. Mammographic screening and mortality from breast cancer: the Malmo mammographic screening trial. BMJ 1988; 297:944-8.

8. Palli D, DelTurco MR, Buiatti LE, et al. A case-control study of the efficacy of a non-randomized breast cancer screening program in Florence (Italy). Int J Cancer 1986; 38:501- 4.

9. Shapiro S, Venet W, Strax P, Venet L, Roeser R. Ten to fourteen year effect of screening on breast cancer mortality. JNCI 1982; 69:349-53.

10. Seidman H, Gelb SK, Silverberg E, et al. Survival experience in the breast cancer detection demonstration project. CA 1987; 37(5):258-90.

11. Haus AG. Technical improvements in screen-film mammography. Radiology 1990; 174(3):628-36.

12. McLelland R, Hendrick RE, Zinninger MD, et al. The American College of Radiology Mammography Accreditation Program. AJR 1991; 157:473-79.

13. Feinstein AR. A bibliography of publications on observer variability. J Chron Dis 1985; 38:619-32.

14. Elmore JG, Feinstein AR. A bibliography of publications on observer variability (final installment). J Chron Dis 1992; 45:567-80.

15. Yerushalmy J, Garland LH, Harkness JT, et al. An evaluation of the role of serial chest roentgenograms in estimating the progress of disease in patients with pulmonary tuberculosis. Am Rev Tuberc 1951; 64:225-48.

16. Cochrane AL, Garland LH. Observer error in interpretation of chest films: international investigation. Lancet 1952; 2:505-9.

17. Nishayama H, Lewis JT, Ashare AB, et al. Interpretation of radionuclide liver images: Do training and experience make a difference? J Nucl Med 1975; 16:11-16.

18. Detre KM, Wright E, Murphy MK, et al. Observer agreement in evaluating coronary angiograms. Circulation 1975; 52:979-86.

19. DeRouen TA, Murray JA, Owen W. Variability in the analysis of coronary arteriograms. Circulation 1977; 55:324-28.

20. Boyd NF, Wolfson C, Moskowitz M, et al. Observer variation in the interpretation of xeromammograms. JNCI 1982; 68:357-63.

21. Chamberlain J, Rogers P, Price JL, et al. Validity of clinical examination and mammography as screening tests for breast cancer. Lancet 1975; 1026-30.

22. Boyd NF, Wolfson C, Moskowitz M. et al. Observer variation in the classification of mammographic parenchymal patterns. J Chronic Dis 1985; 39:465-72.

23. Wolfe JN. Breast patterns as an index of risk for developing breast cancer. Am J Roentgenol 1976; 126:1130-9.

24. Baines CH, McFarlane DV, Miller AB. The role of the reference radiologist. Invest Radiol 1990; 25:971-6.

25. Vineis P, Sinistrero G, Temporelli A, et al. Inter-observer variability in the interpretation of mammograms Tumori 1988; 74:275-9.

26. Baines CJ, Miller AB, Kopans DB. Canadian national breast screening study: Assessment of technical quality by external review. AJR 1990; 155:743-47.

27. Ciccone G, Vineis P, Frigerio A, Segnan N. Inter-observer and intra-observer variability on mammogram interpretations. A field study. Eur J Cancer 1992; 28:1054-8.

28. Dean AG, Dean JA, Burton AH, Dicker RC. Epi Info, Version 5: a work processing database and statistics program for epidemiology on microcomputers. Centers for Disease Control, Atlanta, Georgia, U.S.A., 1990.

29. SAS Version 6.04, SAS Institute Inc, Cary NC, 1987.

30. Cohen J. Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968; 70:213-20.

31. Landis RJ, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159-74.

TABLE 1. EXAMPLE OF PAIRWISE COMPARISON OF DIAGNOSTIC INTERPRETATIONS BY TWO RADIOLOGISTS.

		RADIOLOGIST ’B’				
		Normal	Abnormal-Probably Benign	Abnormal-Indeterminate	Abnormal-Suspicious for Cancer	TOTAL
adio- jist ’A’	Normal	34	18	10	2	64
	Abnormal Benign	3	10	23	1	37
	Abnormal Indeterminate	1	6	12	6	25
	Abnormal-Suspicious for Cancer	0	5	4	15	24
	TOTAL	38	39	49	24	150

Summary Measures of Agreement Between Radiologist ’A’ & ’B’:
% Weighted Agreement = 78%
Weighted Kappa = 0.46

Table 2. Selection of Mammogram Cases

					Reasons for Exclusion						
1987 Diagnostic Interpretation & "Gold Standard" Cancer Outcome	Total 1987	Mammo 90-91 Follow up	Randomized	Mammo 90-91 Follow up	Post Breast Cancer	Post Cosmetic Surgery	Large	Marker	Inadequate Quality	Other	Final Size
I. 'Normal'											
No Cancer Cancer	≈ 3000	n/a[2]	275	89	1 1[3]	3	8	1	4	19	53 1
II. 'Abnormal - Probably Benign'											
No Cancer	550	190	93	n / a	7	2	12	3	4	11	54
Cancer	17	n / a	17	n / a	2	0	3	0	1	4	7
III. 'Abnormal - Suspicious for Cancer'											
No Cancer	81	29	29	n / a	6	2	0	0	2	3	16
Cancer	43	n / a	43	n / a	8	1	2	2	0	11	19
TOTAL	≈ 3691	219	457	89	25	8	25	6	11	48	150

[1] Not in file room (checked out/file lost/films missing from file)
Xeromammogram
Abnormality only seen on lateral view

[2] Not Applicable

[3] Case found to have breast cancer diagnosed after 1987 mammogram.

TABLE 3. CHARACTERISTICS OF THE 150 STUDY PATIENTS.

	No. PTS.
AGE $\geq$ 50 YEARS...............	95 (63%)
POST-MENOPAUSE...............	84 (56%)
SYMPTOMATIC (e.g., breast lump).........	22 (15%)
PAST BREAST BIOPSY...........	40 (27%)
RELATIVE WITH BREAST CANCER..	44 (29%)
PREVIOUS HISTORY OF CANCER...	15 (10%)
CURRENT ESTROGEN USE.........	37 (25%)
BREAST CANCER DIAGNOSED WITHIN 3 YEARS AFTER MMG.....	27 (18%)

TABLE 4. CHARACTERISTICS OF THE 10 PARTICIPATING RADIOLOGISTS.

PRACTICE TYPE: 3 Private, 7 Academic (CT & NY)

CLINICAL EXPERIENCE	MEDIAN	RANGE
No. Years practice:	12	5-30
Estimated No. Mammograms read(1991):	1,900	200-6,000
Estimated % time reading Mammograms in practice:	23%	8%-50%

TABLE 5. MEDIAN AND RANGE FOR PERCENT OF PATIENTS IN WHOM 10 RADIOLOGISTS NOTED THE CITED OBSERVATION, DIAGNOSTIC INTERPRETATION, AND RECOMMENDATION (PHASE I).

	% of patients for whom this result was reported	
	Median	**Range**
OBSERVATION		
Mass	26	19-40
Focal Asymetric Density	16	8-35
Calcification	25	16-33
DIAGNOSTIC INTERPRETATION		
Normal	31	16-61
Abnormal-Probably Benign	25	13-46
Abnormal-Indeterminate	19	8-33
Abnormal-Suspicious for Cancer	18	9-25
RECOMMENDATION		
Routine Follow-up Mammogram	28	21-73
Repeat Mammogram ≤ 6 Months	12	5-36
Immediate Follow-up		
-Additional Mammogram Views	37	11-57
-Ultrasound	10	5-23
-Biopsy	18	9-31

TABLE 6. MEDIAN AND RANGE FOR PERCENT OF PATIENTS IN WHOM 10 RADIOLOGISTS NOTED THE CITED OBSERVATION, DIAGNOSTIC INTERPRETATION, AND RECOMMENDATION (PHASE II).

	% of patients for whom this result was reported.	
	Median	**Range**
OBSERVATION		
Mass	29	19-33
Focal Asymetric Density	19	6-46
Calcification	24	19-36
DIAGNOSTIC INTERPRETATION		
Normal	30	15-61
Abnormal-Probably Benign	23	7-47
Abnormal-Indeterminate	21	12-39
Abnormal-Suspicious for Cancer	18	6-31
RECOMMENDATION		
Routine Follow-up Mammogram	31	11-77
Repeat Mammogram ≤ 6 Months	14	0-13
Immediate Follow-up		
-Additional Mammographic Views	38	11-71
-Ultrasound	10	6-23
-Biopsy	20	10-34

TABLE 7. AGREEMENT ON DIAGNOSTIC INTERPRETATIONS AND BIOPSY RECOMMENDATIONS (PHASE I).

	INTER-OBSERVER VARIABILITY (N=150 patients)	INTRA-OBSERVER VARIABILITY (N=50 patients)[1]
	Median (range) for 10 Radiologists	
	Phase I	Phase I vs II Reading
DIAGNOSTIC INTERPRETATION		
% Weighted Agreement	78% (71%-82%)	84% (72%-89%)
Weighted Kappa	0.47 (0.31-0.55)	0.57 (0.37-0.71)
BIOPSY		
% Agreement	85% (65%-91%)	91% (82%-98%)
Kappa	0.49 (0.20-0.69)	0.71 (0.46-0.91)

[1] In Phase II, one radiologist did not give a diagnostic interpretation for one case.

TABLE 8. AGREEMENT ON DIAGNOSTIC INTERPRETATIONS AND BIOPSY RECOMMENDATIONS (PHASE II).

	INTER-OBSERVER VARIABILITY (N=150 patients)[1]
	Median (range) for 10 Radiologists
	Phase II
DIAGNOSTIC INTERPRETATION	
% Weighted Agreement	77% (65%-82%)
Weighted Kappa	0.41 (0.24-0.58)
BIOPSY	
% Agreement	83% (73%-95%)
Kappa	0.49 (0.28-0.80)

[1] In Phase II, one radiologist did not give a diagnostic interpretation for one case.

TABLE 9. ACCURACY OF INTERPRETATIONS IN THE 27 'CANCER' PATIENTS (PHASE I).

	Radiologist									
	A	B	C	D	E	F	G	H	I	J
Normal	0	1	2	1	2	2	3	1	4	6
Abnormal-Benign/Indeterminate	4	6	5	6	6	6	8	11	11	11
Abnormal-Suspicious for Cancer	23	20	20	20	19	19	16	15	12	10

TABLE 10. ACCURACY OF DIAGNOSTIC INTERPRETATION AND PERTINENCE OF MANAGEMENT RECOMMENDATIONS [PHASE I].

	DIAGNOSTIC INTERPRETATION[1]		**MANAGEMENT RECOMMENDATIONS** **% of Patients with Immediate Work-up Recommended**[2]	
RADIOLOGIST	Sensitivity(%) (N=27)	Specificity(%) (N=123)	Cancer Patients (N=27)	Non-Cancer Patients (N=123)
A	85	93	93	57
B	74	94	93	65
C	74	94	96	40
D	74	85	96	64
E	70	87	89	41
F	70	93	85	44
G	59	94	78	30
H	56	93	81	38
I	44	99	85	45
J	37	97	74	11

1. For calculating sensitivity and specificity, partitions were 'Abnormal-Suspicious for Cancer' category vs 'Abnormal-Indeterminate', 'Abnormal-Probably Benign', and 'Normal' categories.
2. 'Immediate work-up' defined as a recommendation to obtain additional mammogram views, ultrasound, and /or biopsy.

TABLE 11. ACCURACY OF DIAGNOSTIC INTERPRETATION AND PERTINENCE OF MANAGEMENT RECOMMENDATIONS (PHASE II).

RADIOL-OGIST	**DIAGNOSTIC INTERPRETATION**[1]		**MANAGEMENT RECOMMENDATIONS % of Patients with Immediate Work-up Recommended**[2]	
	Sensitivity(%) (N=27)	Specificity(%) (N=123)	Cancer Patients (N=27)	Non-Cancer Patients (N=123)
A	74	93	100	72
B	70	98	96	48
C	81	87	85	38
D	85	81	96	53
E	67	85	78	36
F	70	94	85	36
G	67	100	89	33
H	59	89	85	38
I	56	97	85	41
J	30	99	70	10

1. For calculating sensitivity and specificity, partitions were 'Abnormal-Suspicious for Cancer' category vs 'Abnormal-Indeterminate', Abnormal-Probably Benign', and 'Normal' categories.
2. 'Immediate work-up' defined as a recommendation to obtain additional mammogram views, ultrasound, and /or biopsy.

TABLE 12. DISAGREEMENTS BETWEEN 10 RADIOLOGISTS IN STATED LOCATION (RIGHT vs LEFT BREAST) OF MOST SIGNIFICANT MAMMOGRAPHIC ABNORMALITY (PHASE I).

	PER-PATIENT COMPARISON[1]		PER-PAIRWISE COMPARISON[2]	
	Inter-Observer	Intra-Observer	Inter-Observer	Intra-Observer
Total cases in which two or more radiologists reported an abnormality	42% (59/141)	38% (18/47)	12% (384/3306)	10% (27/273)
●● Cases where immediate work-up was recommended (additional views, ultrasound, or biopsy)	33% (41/123)	36% (16/45)	9% (209/2246)	10% (24/247)
●● Cases where biopsy was recommended	9% (4/46)	14% (4/28)	2% (18/733)	3% (4/122)

[1] Numerator is number of patients with a disagreement in stated location of abnormality. Denominator is total number of patients who had an abnormality specified by at least 2 radiologists.

[2] Numerator is number of pairwise comparisons with a disagreement in the stated location of an abnormality. Denominator is the maximum number of pairwise comparisons given that a side was specified by at least 2 radiologists.

TABLE 13. DISAGREEMENTS BETWEEN 10 RADIOLOGISTS IN STATED LOCATION (RIGHT vs LEFT BREAST) OF MOST SIGNIFICANT MAMMOGRAPHIC ABNORMALITY (PHASE II).

	PER-PATIENT COMPARISON[1]	PER-PAIRWISE COMPARISON[2]
	Inter-Observer	Inter-Observer
Total cases in which two or more radiologists reported an abnormality	46% (63/136)	12% (426/3519)
•• Cases where immediate work-up was recommended (additional views, ultrasound, or biopsy)	40% (53/129)	11% (207/2616)
•• Cases where biopsy was recommended	10% (5/52)	3% (21/796)

[1] Numerator is number of patients with a disagreement in stated location of abnormality. Denominator is total number of patients who had an abnormality specified by at least 2 radiologists.

[2] Numerator is number of pairwise comparisons with a disagreement in the stated location of an abnormality. Denominator is the maximum number of pairwise comparisons given that a side was specified by at least 2 radiologists.

TABLE 14. MAJOR DISAGREEMENTS IN DIAGNOSTIC INTERPRETATIONS AND MANAGEMENT RECOMMENDATIONS (PHASE I).

	EXAMPLE OF DISAGREEMENT FOR READING ON THE SAME PATIENT		% WITH MAJOR CLINICAL DISAGREEMENTS			
			PER-PATIENT COMPARISON		PER-PAIRWISE COMPARISON	
	READING 1	READING 2	Inter-Observer	Intra-Observer	Inter-Observer	Intra-Observer
DIAGNOSTIC INTERPRETATION	'Normal'	'Abnormal-Suspicious for Cancer'	19% (28/150)	8% (4/50)	2% (130/6750)	1% (5/499)
MANAGEMENT RECOMMENDATION	'Routine Follow-Up'	'Biopsy'	25% (37/150)	8% (4/50)	3% (234/6750)	1% (5/500)

TABLE 15. MAJOR DISAGREEMENTS IN DIAGNOSTIC INTERPRETATIONS AND MANAGEMENT RECOMMENDATIONS (PHASE II).

	EXAMPLE OF DISAGREEMENT FOR READING ON THE SAME PATIENT		% WITH MAJOR CLINICAL DISAGREEMENTS	
			PER-PATIENT COMPARISON	PER-PAIRWISE COMPARISON
	READING 1	READING 2	Inter-Observer	Inter-Observer
DIAGNOSTIC INTERPRETATION	'Normal'	'Abnormal-Suspicious for Cancer'	23% (35/150)	5% (370/6741)[1]
MANAGEMENT RECOMMENDATION	'Routine Follow-Up'	'Biopsy'	28% (42/150)	3% (213/6741)

[1] In Phase II, one radiologist did not give a diagnostic interpretation for one case.

Table 16. Comparisions of Studies on Observer Variability in Mammography

					Median Weighted Kappa Values (Range)/Weighted Percent Agreement			
					Inter-Observer Variability		Intra-Observer Variability	
Investigators	Year	# of Pts.	# of Rads.	% of Pts. with 'Cancer'	Diagnostic Interpretations	Biopsy Recommendations	Diagnostic Interpretations	Biopsy Recommendations
Chamberlain et al.	1975	1215	2	1.6%	n/a[1]	50%	n/a	n/a
Boyd et al.	1982	100	9	10%	0.42 (0.23–0.59)	0.54 (0.44–0.63)	n/a	n/a
Baines et al.	1981-88	5975	2	11%	74.5%	n/a	n/a	n/a
Vineis et al.	1988	45	8	20%	0.48 (0.33–0.67)	n/c[2]	n/a	n/a
Ciccone et al.	1992	45	6	20%	n/c	0.42 (0.29–0.45)	n/c	0.55 (0.36–0.65)
Howard et al.	1994	150	10	18%	0.47 (0.31–0.55)/78$	0.49 (0.20–0.69)/85%	0.57 (0.37–0.71)/84%	0.71 (0.46–0.91)/91%

[1] Not Applicable

[2] Not Calculated

For office use only

APPENDIX I

PATIENT HISTORY FORM

AMMOGRAM: ________ ________ (1-3) ___ ___ ___

GE: ____ (4-5) ___ ___

ENOPAUSAL STATUS: Pre ___ Post ___ (6) ___

ISTORY OF BREAST BIOPSY: Yes ____ No ____ (7) ___

Years since biopsy ___ ___ (8-9) ___ ___

(10) ___

(11-12) ___ ___

PREVIOUS HISTORY OF CANCER Yes____________ No____ (13) ___

FAMILY HISTORY OF BREAST CANCER: Yes ____ No____ (14) ___

Relative: ______________________________ (15) ___

Age of individual when cancer was found ___ ___ (16-17) ___ ___

(18) ___

(19-20) ___ ___

CURRENT ESTROGEN USE: Yes ____ No____ Uncertain____ (21) ___

Years of estrogen ___ ___ (22-23) ___ ___

SIGNS AND SYMPTOMS:

Lump: Right___ Left___ (24) ___

Nipple change: Right___ Left___ (25) ___

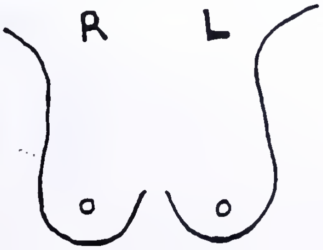
Nipple discharge: Right___ Left___ (26) ___

Pain: Right___ Left___ (27) ___

Skin change: Right___ Left___ (28) ___

(29) ___

(30) ___



APPENDIX II

PATIENT HISTORY FORM: DEFINITIONS AND CODING

The definitions below clarify the terms used on the patient history form. Each definition is followed by numbers (in parentheses) and an abbreviated term used for computer coding.

MAMMOGRAM: mammograms given randomly assigned number, 1 - 150, for Phase I and 151 - 305 for Phase II

(1-3) MAMMO

AGE: age of patient at 1987 mammogram

(4,5) AGE

MENOPAUSAL STATUS: (Pre) - pre-menopausal in 1987 or ≤ 50 years of age with history of hysterectomy and one or both ovaries remaining

(Post) - post-menopausal in 1987 or > 50 years of age with a history of hysterectomy, regardless of whether one or both ovaries remained

(6) MENOPA

0: Pre
1: Post

HISTORY OF BREAST BIOPSY: (Yes/No) - history of breast biopsy prior to 1987 mammogram (positive response included location information such as right, left or bilateral); with more than one biopsy, location information provided for each procedure

(7) HXBX1

0: No
1: Right
2: Left
3: Bilat

YEARS SINCE BIOPSY: number of years, prior to 1987, that biopsy was performed

(8,9) BXYRS1

00: Not Applicable
99: Not Certain

HISTORY OF BREAST BIOPSY: (Yes/No) - repeated for second biopsy, if applicable

(10) HXBX2

0: No
1: Right
2: Left
3: Bilat

YEARS SINCE BIOPSY: information repeated for second biopsy, if applicable

(11,12) BXYRS2

00: Not Applicable
99: Not Certain

PREVIOUS HISTORY OF CANCER: (Yes/No) - history of cancer (other than breast cancer

(13) HXCA

0: No
1: Cervical
2: Parotid Tumor
3: Squamous Skin
4: Uterine
5: Thyroid
6: Basal Cell Lip
7: Melanoma

FAMILY HISTORY OF BREAST CANCER: (Yes/No) - history of breast cancer in patient's family

(14) FMH

0: No
1: Yes

FIRST DEGREE RELATIVE: specific relative afflicted; any other relative categorized as 'other'; if just 'aunt', relative was coded as 'maternal aunt'

(15) REL1

1: Mother
2: Sister
3: Maternal Grandmother
4: Paternal Grandmother
5: Maternal Aunt
6: Paternal Aunt
7: Daughter
8: Other
0: Not Applicable

AGE OF INDIVIDUAL WHEN CANCER WAS FOUND: age of relative when cancer was diagnosed

(16,17) RELAGE1

00: Not Applicable
99: Uncertain

(18) REL2 (Information repeated for second relative)

(19,20) RELAGE2 (Information repeated for second relative)

CURRENT ESTROGEN USE: (Yes/No/Uncertain) - includes use of estrogen, premarin, progesterone, and birth control medication

(21) ESTROGEN
0: No
1: Yes
9: Uncertain

YEARS OF ESTROGEN: length of time, in years, of estrogen use prior to 1987

(22,23) ESTYRS
00: Not Applicable
99: Uncertain

SIGNS AND SYMPTOMS: positive findings, by patient report, for breast lump, nipple change, nipple discharge, pain, and skin change; category was further defined by left or right breast

(24) LUMP
1: Right
2: Left
3: Bilat
0: Not Applicable

(25) NIPPLE CHANGE
1: Right
2: Left
3: Bilat
0: Not Applicable

(26) NIPPLE DISCHARGE
1: Right
2: Left
3: Bilat
0: Not Applicable

(27) PAIN
1: Right
2: Left
3: Bilat
0: Not Applicable

(28) SKIN CHANGE
1: Right
2: Left
3: Bilat
0: Not Applicable

PICTURE: graphic picture and/or comments used, in some cases, to denote specific location of biopsy scar(s), lump(s), skin change(s), etc.; picture completed by mammography technologist at time of 1987 mammogram; additional comments which were written by the technologist were copied verbatim

(29) PICTURE

0: No
1: Yes

CANCER PRESENT: used for coding purposes only

(30) CANCER

0: No
1: Yes

APPENDIX III

INTERPRETATION CHECK LIST

Mammogram ____________

Radiologist ____________

__________ NO SIGNIFICANT ABNORMALITY

ABNORMAL FINDING

I. Mass:

_______ Probably benign
_______ Intermediate concern
_______ Suspicious for malignancy

II. Calcifications:

_______ Probably benign
_______ Intermediate concern
_______ Suspicious for malignancy

III. Focal Asymmetric Density:

_______ Probably benign
_______ Intermediate concern
_______ Suspicious for malignancy

IV. Architectural Distortion:

_______ Probably benign
_______ Intermediate concern
_______ Suspicious for malignancy

V. Other Findings:

_______ Skin retraction
_______ Nipple retraction
_______ Skin thickening
_______ Trabecular thickening

_______ Skin lesion
_______ Axiliary adenopathy
_______ Solitary dilated duct/ tubular density

VI. Other ____

____

VII. Location of Abnormality

____Left ____Right ____ O'Clock
____Left ____Right ____ O'Clock

__________ ABNORMAL FINDING - PROBABLY BENIGN

__________ ABNORMAL FINDING - INDETERMINATE

__________ ABNORMAL FINDING - SUSPICIOUS FOR MALIGNANCY

RECOMMENDATIONS FOR FOLLOW-UP

_______ Physical exam
_______ Ultrasound exam
_______ Additional X-ray views at this time
_______ Biopsy

_______ Age appropriate follow-up
_______ Repeat Mammogram ___ mths

For Office Use Only

(1) ___
(2-4) ___ ___ ___
(5-6) ___ ___
(7) ___
(8) ___
(9) ___
(10) ___
(11) ___
(12) ___
(13) ___
(14) ___
(15) ___
(16) ___
(17) ___
(18) ___
(19) ___
(20) ___
(21, 22) ___ ___
(23) ___
(24, 25) ___ ___
(26) ___
(27) ___
(28) ___
(29) ___
(30) ___
(31) ___
(32,33) ___ ___

APPENDIX IV

INTERPRETATION CHECK LIST: DEFINITIONS AND CODING

The terms below correspond to the Interpretation Check List Form. The numbers in parentheses, as well as the abbreviated term, were used for computer coding.

(1) PHASE-Phase #

0: 1987 Reading
1: Phase I
2: Phase 2

(2-4) MAMMO-Mammogram

#001 to #150 or #151 to #305

(5-6) MD-Radiologist #

A - J: 10 study radiologists
Z: a 1987 radiologist

(7) HIST-Clinical History

0: Non-Present
1: Present
2: False History Present

(8) FIND-Finding

0: No - No Significant Abnormality
1: Yes - Abnormal Finding - Probably Benign
2: Yes - Abnormal Finding - Indeterminate
3: Yes - Abnormal Finding - Suspicious For Malignancy
9: Unknown - Poor Quality Film

(9) MASS-Mass

0: No
1: Probably Benign
2: Intermediate Concern
3: Suspicious For Malignancy

(10) CALC-Calcifications

0: No
1: Probably Benign
2: Intermediate Concern
3: Suspicious For Malignancy

(11) FOCAD Focal Asymmetric Density

0: No
1: Probably Benign
2: Intermediate Concern
3: Suspicious For Malignancy

(12) ARCHDI-Architectural Distortion

0:	No
1:	Probably Benign
2:	Intermediate Concern
3:	Suspicious For Malignancy

(13) SKINRE-Skin Retraction

0:	No
1:	Yes

(14) NIPPRE-Nipple Retraction

0:	No
1:	Yes

(15) SKINTH-Skin Thickening

0:	No
1:	Yes

(16) TRABTH-Trabecular Thickening

0:	No
1:	Yes

(17) SKINLE-Skin Lesion

0:	No
1:	Yes

(18) AXILAD-Axillary Adenopathy

0:	No
1:	Yes

(19) DUCT-Solitary Dilated Duct/Tubular Density

0:	No
1:	Yes

(20) SIDE1-Side of Abnormality

0:	No Abnormality
1:	Right
2:	Left
3:	Bilateral

(21,22) LOCAT1-Location of Abnormality

00:	No Abnormality
1-12	_____ o'clock
13:	Retroareolar
14:	Central
15:	Axilla
16:	Diffuse/Scattered
17:	Other (10=UOQ, 8=LOQ, 4=LIQ, 2=UIQ)

(23) SIDE2-Side of Second Abnormality

0: No
1: Right
2: Left
3: Bilateral

(24,25) LOCAT2-Location of Second Abnormality

00: No Abnormality
1-12 _____o'clock
13: Retroareolar
14: Central
15: Axilla
16: Diffuse/Scattered
17: Other (10=UOQ, 8=LOQ, 4=LIQ, 2=UIQ)

(26) PHYSEX-Physical Exam Suggested

0: No
1: Yes

(27) ULTRAS-Ultrasound Exam Suggested

0: No
1: Yes

(28) ADDVWS-Additional Views Suggested At This Time

0: No
1: Yes

(29) BIOPSY-Biopsy

0: No
1: Yes
2: If Other Recommendations are Positive

(30) AGEFU-Age Appropriate Follow-up

0: No
1: Yes

(31) REPEAT-Repeat Mammogram

0: No
1: Yes

(32,33) MONTHS-Months Until Repeat Mammogram

00: No Repeat Mammogram

INTERPRETATION CHECK LIST CODING INSTRUCTIONS

The Interpretation Check List was coded according to notations made

by the radiologists. Clarifications of some of the coding complexities follow:

1. Within any sub-category, if two abnormalities were noted, only the most suspicious interpretation was coded. If two abnormalities were noted in different sub-categories, both were coded.

2. In coding the exact location of the abnormality(ies), reference to a clock face was used. If the location was not designated by '___o'clock', but was designated by quadrant, the quadrant was translated to the appropriate 'o'clock' location (2, 4, 8, or 10 o'clock) for each breast.

3. If an 'o'clock' notation was designated (e.g., 7:30 or 2 - 3 o'clock), this was always coded by the lower o'clock number.

4. If the location of the abnormality was designated by a narrative, the following categories were used in coding: retroareolar, central, axillary, diffuse/scattered. Any other narrative was coded as 'other'.

APPENDIX V

RADIOLOGISTS' INSTRUCTIONS FOR INTERPRETATION CHECK LIST

Thank you for participating in this study of mammography.

Below, please find instructions for the interpretation check list which will be used to document your findings. Please complete one check list for each case. Only four views will be available for this study. A clinical history is available for only some of the cases. When a breast biopsy is noted in the clinical history, the results were negative.

Please interpret these mammogram cases as you would in your regular clinical setting.

1. The information in the top right corner of the checklist has been completed by us and should include:
   i. The identification number of the case being interpreted.
   ii. The personal identification letter assigned to you.

2. Only four major categories are available. They are as follows:
   'No Significant Abnormality'
   'Abnormal Finding - Probably Benign'
   'Abnormal Finding - Indeterminate'
   'Abnormal Finding - Suspicious for Malignancy'

   Because these four categories are mutually exclusive, please check only one category for each use. For this study, fibrocystic disease falls under the category of 'No Significant Abnormality'. The category listed as 'Abnormal Finding - Indeterminate' should only be used if it is impossible for you to assign any of the other categories without additional information. The category 'Abnormal Finding - Suspicious for Malignancy' should only be used when recommendations for a biopsy are in order.

3. To ensure consistency within this study, the following brief definitions for 'Abnormal Findings' are provided below:

I. MASS: A space-occupying lesion seen in two different projections. If a potential mass is seen in only a single projection, it should be called a 'DENSITY' until its three-dimensionality is confirmed.

II. CALCIFICATIONS: Any calcium deposit that could potentially represent a malignancy should be noted.

III. FOCAL ASYMMETRIC DENSITY: This is a density that cannot be accurately described using the other shapes. It is visible as asymmetry of tissue density. It could represent an island of normal breast, but its lack of specific benign characteristics may warrant further evaluation.

IV. ARCHITECTURAL DISTORTION: The normal architecture is distorted with no definite mass visible. This includes spiculation radiating from a point and focal retraction or distortion of the edge of the parenchyma.

V. OTHER FINDINGS:

SKIN RETRACTION: The skin is pulled in abnormally.

NIPPLE RETRACTION: The nipple is pulled in or inverted.

SKIN THICKENING: This may be focal or diffuse.

TRABECULAR THICKENING: This is a thickening of the fibrous septae of the breast.

SKIN LESION: Noted when it projects over the breast in two views.

AXILLARY ADENOPATHY: Enlarged axillary lymph nodes, apparently replaced by non-fatty tissue.

SOLITARY DILATED DUCT/TUBULAR DENSITY: A tubular or branching structure that likely represents a dilated or otherwise enlarged duct.

VI. OTHER: this category does not have to be filled out. It may be used at your discretion, if you would like to make note of anything. We would be happy to return these forms to you at the end of the study and you may find them helpful when reviewing our final study results.

4. If your interpretation includes more than one abnormal finding, please indicate all on the check list. When marking the presence of the most suspicious abnormal finding(s). mark a '1' next to the finding(s) and state the corresponding location of the most suspicious finding under the first part of 'VII. Location of Abnormality.' Mark a '2' next to any finding(s) that is of lesser suspicion and state the corresponding location under the second part of 'VII. Location of Abnormality'. Place an 'X' next to any additional abnormal finding(s) noted. The location of the (X) finding(s) can not be designated unless you choose to do so under the 'VI. Other' section. (Although more than one abnormal finding may be noted on this check list form, please choose only one major category for your final interpretation.)

5. At the bottom of the checklist, place a mark next to any 'Recommendations for Follow-Up' you wish to make. Place a mark next to the stated recommendation only if you want to recommend it for specific follow-up. Please recommend at least one form of follow-up. You may recommend more than one type of follow-up if you feel it is warranted (e.g. both a physical exam and an ultrasound). Mark only one choice for timing of mammographic follow-up ('Age Appropriate Follow-up' or 'Repeat Mammogram in ___ Months'). List the time, in months, suggested for repeated mammogram if you choose that recommendation.

APPENDIX VI

TELEPHONE SURVEY OF PARTICIPATING RADIOLOGISTS

1. Do you consider yourself an expert in mammography?

   Yes___ No___

2. Did you recognize any of the films in the second phase?

   Yes___ No___

   If yes to above: A) How many?

   1%-20% / 21%-40% / 41%-60% / 61%-80% / 81%-100%

   B) In how many films did you remember what your initial reading was in Phase I ?

   0% / 1%-20% / 21%-40% / 41%-60% / 61%-80% / 81%-100%

3. Since it would be expected that radiologists may vary in their interpretation of mammograms, in what percentage of cases is this variability clinically important?

   0% / 1%-20% / 21%-40% / 41%-60% / 61%-80% / 81%-100%

4. Did you give these mammograms the same consideration as in your clinical practice?

   Yes___ No___ Other___

APPENDIX VII

TRANSCRIPT: CONFERENCE OF PARTICIPATING RADIOLOGISTS

INTRODUCTION

The Consensus Conference was attended by eight of the ten participating radiologists and all five investigators at the conclusion of Phase II testing. A brief review of the research design and preliminary data was presented. Six mammograms which showed great variability in interpretation were discussed. The radiologists' comments regarding the variability can be summarized in four categories:

1) Abnormalities were simply missed by some radiologists
2) Disagreement was often present as to whether a visual appearance fit criteria for certain diagnoses
3) Radiologists were asked to categorize diagnoses without immediate follow-up information
4) Different thresholds were used when making the diagnoses and recommendations.

The six mammograms discussed by Conference participants are noted below. Following each mammogram 'heading', in parentheses, is the subsequent cancer outcome of each case. (The radiologists were not informed of this result until the end of each discussion.) For each case, the variability among the ten radiologists in interpretations and management recommendations is also shown. For purposes of confidentiality, the radiologists are noted by number in the order that they spoke.

Mammogram I ('NO CANCER')

INTERPRETATIONS	# RADIOLOGISTS
Normal	3
Abnormal-probably benign	1
Indeterminate	4
Abnormal-suspicious for cancer	2

MANAGEMENT RECOMMENDATIONS	#RADIOLOGISTS
Age-appropriate follow-up	3
Repeat Mammogram $\leq$ 6 mos.	1
Add'l X-ray views now	1
Ultrasound	0
Biopsy	2

RAD. #1

"what bothered my eye was the asymmetry in the periareola area...The thickening in the left nipple caught my eye...when I look at a mammogram I look for some degree of symmetry...In my practice I would have asked my technologist about the left nipple...Maybe it is technical but, when I read these mammos I didn't worry about technique. I figured if you presented them to me you already dealt with the technique part...I was worried...I read it as malignant. I thought it was Paget's, frankly...I didn't see anything in the right breast that bothered me. I thought there were benign calcifications."

RAD. #2

"I called this normal and I was fairly shocked because when you put the film up I said, 'Gee, look at that'...I didn't see it...This is the thing that you hope doesn't happen to you in your practice.' "One thing has changed since 1987...some people say that if you can see the skin on a mammogram, then it is an improperly exposed mammogram. I actually have stopped looking at the skin or nipple anymore. In this case, I would certainly examine the nipple, bring her back for some compressions...and mention to her physician, 'What is going on with this nipple?'"

RAD. #3

"There is some range of asymmetry that is tolerable...This passes my threshold for asymmetry."

RAD. #4

"I made no comment about that as well. I dwelled on the calcifications...I can't tell you whether I didn't notice it or I dismissed it as within the range of variability that is O.K. around the areola. Elsewhere in the breast I would consider that significant, but the areola is the one area where there can be both much more thickening and asymmetry. I'm not sure what you do about it other than making a punch biopsy anyway. If you think that is abnormal, then I don't think I would follow it...you'd either call it normal or punch biopsy the skin. Those are the only choices."

RAD. #5 *(Discussing breast calcifications)*

"I thought some of these were irregular. They branch, are...fat in one part and skinny in another...I was worried about them. I must say I am a Homer advocate. He really believes that calcifications cannot be categorized. We've seen in our practice, more than once, calcifications that are punctate, should be benign, and are malignant....It's a few of them, it's not all of them...I don't know how many you have to have before you worry about them. I wouldn't call these benign, I know that. I wouldn't dismiss them...The other problem I had with many of these cases, and when I did the first set I wasn't really sure what to do...I would almost always get mag views...except for a small number of these which are grossly malignant, my next step would have been to get mag views; I wouldn't have biopsied many at all."

DISCUSSION LEADER #2

"What about someone who thought these calcifications were grossly benign?"

RAD. #1

"Now that I look at them, I am a little bit more concerned about them, but I think they didn't really bother me that much because I saw similar calcifications in the right breast...I thought they were all sort of secretory calcifications...I wrote them off. I didn't even mention them."

RAD. #6

"I agree."

RAD. #4

"The ones that are branching are two overlapping secretory calcifications."

DISCUSSION LEADER # 1

"If I said 'O.K., picture a mammogram that has this,' and describe just what you said, calcifications that are branching, irregularly shaped, everybody would agree that those are suspicious. The question is when you look at this, some people come up with, 'Gee, these fit those criteria and therefore are suspicious,' and others say, 'these do not fit those criteria and are therefore not suspicious.'"

DISCUSSION LEADER #2

"The disagreement here is not on the criterion...but whether that particular visual appearance fits the designation of pleomorphic, etc..."

RAD. #9

"I pushed for additional views and to bring her back in three months...I pushed because there was certainly asymmetry between the appearance of both breasts, and the periareola skin thickening in the left breast was worrisome, but indeterminate in my eyes; that is why...I wanted more films right now and plus bring her back in three months."

MAMMOGRAM II ('CANCER')

INTERPRETATIONS	# RADIOLOGISTS
Normal	0
Abnormal-probably benign	2
Indeterminate	5
Abnormal-suspicious for cancer	3

MANAGEMENT RECOMMENDATIONS	# RADIOLOGISTS
Age-appropriate follow-up	0
Repeat mammograms $\leq$ 6 mos.	1
Add'l x-rays views now	6
Ultrasound	7
Biopsy	2

DISCUSSION LEADER #1

"This is a case we picked because everyone agreed on the finding; unlike the last case where some didn't mention it, other people thought calcifications, others density. Everybody agreed what was the abnormality in this case."

RAD. #1

"I think this is a good example of the kind of case in which we are being artificially forced to make checks on a form, which is not necessarily the way most, if not all of us, would practice. I think we all see what it is, probably all of us would want to get spot compression views of it, then probably ultrasound it, and then make a recommendation...so, I checked ultrasound...additional views, and I made an arrow with a question mark to biopsy." *(Group Agreement)*

DISCUSSION LEADER #2

"Is there anything else in the film that anyone wants to comment on?"

RAD. #2

"When I took the magnifying glass, I was really impressed that when you mag that thing it is irregular. It bothered me very much...I read it as suspicious....If

that thing was 5 mm or less, I might sit on it...but this thing is big, a centimeter or bigger. Its got hazy margins...slight increase in trabecular pattern in that area. It caught my eye and everything pointed toward a malignant process...Just because it is round doesn't write off malignant."

RAD. #1

"I don't see smooth margins, but I've seen countless cases of things that I've needled and when the specimen comes out it looks 10x smoother than on the mag view...so, I won't be shocked if it does have smooth margins."

RAD. #3

"I thought from your instructions that you wanted me to say this is malignant or benign. That is not the way I conduct my mammography practice. In my practice, if something is wrong here, I have to get more information. I was extremely uncomfortable the first 50 cases thinking...on the basis of four films they want me to say this is probably benign or...so, after the first 50 cases, I was so uncomfortable...I said...'I'm going to do it my way.'" *(Group Agreement)*

RAD #3

"I wanted to put to the referring physician a big something. So, I put suspicious. I checked physical exam, ultrasound, additional x-rays and biopsy. I have seen too many round shadows that may look benign, but just don't turn out to be...A solitary round mass in a woman's breast; you do everything you can to prove it one way or the other."

RAD. #4

"The important thing about this case is that...I think everybody said to do something more."

DISCUSSION LEADER #1

"None of the radiologists in this room disagree...the problem is with the artificial categorization. The reason why I wanted to show this case is because...I think these are basically the same readings, but if you look at the numbers, it looks different."

MAMMOGRAM III ('CANCER')

INTERPRETATIONS	# RADIOLOGISTS
Normal	1
Abnormal-probably benign	7
Indeterminate	2
Abnormal-suspicious for cancer	0

MANAGEMENT RECOMMENDATIONS	# RADIOLOGISTS
Age-appropriate follow-up	1
Repeat mammograms $\leq$ 6 mos.	5
Add'l x-rays views now	4
Ultrasound	6
Biopsy	0

RAD.#1

"You've got some calcifications, a nodule there, a nodule there. You've got multiple findings and this is the kind of thing where I usually recommend asymptomatic follow-up in six months...I don't know what she was like six months before...let the garden grow, if one of them is a cancer it is going to get bigger...and maybe we might find it in six months. If you say it is abnormal, then are you going to take that one out and that one out? I usually put this in a follow-up category."

RAD. #2

"I wrote the left one off as benign. I didn't want to run up a big bill on this lady...I thought I could write that one off as very well circumscribed, but

recommended follow-up in six months. I was more aggressive on the right, primarily because of the scalloped margins and...the density was certainly larger than 1 cm. The calcifications I wrote off. I recommended ultrasound and said if that ultrasound turns out to be solid then I wanted that removed...Anything more than 8 mm and still well-circumscribed I still may remove it..that is something I've just developed through the years. Why I choose 8 mm? Well, I don't know..I think Kopans uses it."

RAD. #3

"I felt fairly clear about the left breast being reasonably benign...but on the right breast, I felt that there were some margins...I couldn't clear them completely. There was some superposition of tissue and I definitely needed some additional views...I go with anything greater than 8-10 mm in diameter is definitely going to need either mag views or an ultrasound...sometimes I really don't care what the margins look like at all, I don't care if they are blurred, I don't care if they are circumscribed...I'm not using a clear size criteria either. If I can see a complete peripheral halo that is the only time I feel uncomfortable saying that it is probably benign. Sometimes you've got to do something more and I think doing an ultrasound or additional views are going to help you with that...I didn't write down follow-up, but now, in retrospect, I would get some closer follow-up."

DISCUSSION LEADER #1

"I think what is interesting about this case is that the readings were pretty uniform. There were seven 'abnormal-probably benign'; but people described completely different things."

RAD. #2

"There were times I could have given you three or four masses on this thing..but I said, 'O.K., I'll talk about two here.'" *(Group agreement....after being*

told that the patient was subsequently found to have 'cancer' in one mass)

RAD.#4

"I bet you even the people who said probably benign, but wanted to do a work-up are probably not going to be so shocked. I think probably benign means 5-10% of them are going to be cancers, but still we want to look for that percentage...I get the sense that we are all on the same wavelength. If you thought there was a 30-40% chance of cancer you might have it 'indeterminate' or 'suspicious for malignancy'. But, a 10% chance still is enough to do something._

RAD. #3

"I'm prepared to be wrong if I recommend biopsy a lot of the time. It doesn't bother me if something comes back normal. But, there are some things I don't biopsy at all because they are not indeterminate."

DISCUSSION LEADER #1

"We have a 25-30% biopsy rate *(at Yale)*. If you have a 10% positive biopsy rate, your surgeons aren't going to be too happy with your readings."

MAMMOGRAM IV ('CANCER')

INTERPRETATIONS	# RADIOLOGISTS
Normal	5
Abnormal-probably benign	2
Indeterminate	1
Abnormal-suspicious for cancer	2

MANAGEMENT RECOMMENDATIONS	# RADIOLOGISTS
Age-appropriate follow-up	5
Repeat mammograms ≤6 mos.	0
Add'l x-rays views now	5
Ultrasound	0
Biopsy	2

DISCUSSION LEADER #1

"Some people thought calcifications, some thought asymmetric density in the right breast. One person thought thickening of the parenchyma...so this really was a case that had a lot of variability."

RAD. #1

"I read this as normal and as I sit in the back of the room I am a little bothered by the retroglandular tissue of that left breast."

RAD. #2

"I thought this was suspicious because of the calcifications."

DISCUSSION LEADER #2

"Someone who called this 'normal', do those calcifications turn you on?"

RAD. #3

"They are not clustered to my eye, they have varying sizes...they don't maintain a strictly segmental distribution. They didn't get my threshold. We all have different thresholds."

[Part of discussion on Mammogram 'D' and 'E' was missed when the tape was being turned over]

MAMMOGRAM V ('NO CANCER')

INTERPRETATIONS	# RADIOLOGISTS
Normal	1
Abnormal-probably benign	3
Indeterminate	4
Abnormal-suspicious for cancer	2

MANAGEMENT RECOMMENDATIONS	# RADIOLOGISTS
Age-appropriate follow-up	1
Repeat mammograms $\leq$ 6 mos.	4
Add'l x-rays views now	7
Ultrasound	0
Biopsy	0

RAD. #1

"I saw a mass on the right. I saw it on one view, but I saw it on one view fairly well...maybe I didn't have my coffee that day...Now, I'm not so concerned...I think it was right in here. It looked like I saw something that was marginated."

DISCUSSION LEADER #1

"Why is it that some people see masses and others don't?"

RAD. #1

"It is purely perceptual. I can't teach you to perceive something the way I perceive it...that is one thing that you can't be taught."

RAD.#2

"Some of us may have sat down and read these 150 mammograms in three batches of 50, sat down, read 50, got up, sat down...some of us only read two to three at a time."

DISCUSSION LEADER #2

"What is on that film is on that film...it is then a matter of how we perceive that. Now, there are sometimes when somebody will say to me, 'it's a rorschach test and it shows a dancing ballerina on Hollywood Bowl' and somebody else says, 'I don't see it!' That is an instance where you are seeing the same thing but you just don't see the certain forms."

RAD. #3

"This is one of the problems I had with the way the study was set up...I agree there are calcifications on both sides and I want other views. Then somebody asks me what category to put it into. It doesn't really matter in my mind. I don't care if you call it 'suspicious for malignancy'. I want additional views before I suggest a biopsy."

MAMMOGRAM VI ('NO CANCER')

INTERPRETATIONS	# RADIOLOGISTS
Normal	0
Abnormal-probably benign	3
Indeterminate	3
Abnormal-suspicious for cancer	4

MANAGEMENT RECOMMENDATIONS	# RADIOLOGISTS
Age-appropriate follow-up	1
Repeat mammograms $\leq$ 6 mos.	0
Add'l x-rays views now	0
Ultrasound	0
Biopsy	3

DISCUSSION LEADER #1

"This is a cute case because everyone saw the same finding which was skin retraction and this ugly looking area. Phase I readings...with no clinical history was 'abnormal-probably benign', three people. Three people said it was 'indeterminate' and four said it was 'suspicious'. It turns out she had a biopsy...and when you got the history the second time around, everyone read it as either 'abnormal-probably benign', related to the biopsy...or completely 'normal'.

APPENDIX VIII

To all whom these letters shall come, greeting:

Be it known that

has earned grateful appreciation for constructive participation in the Yale University Study of Improvements in Mammography, for enabling the field of Clinical Epidemiology to keep abreast of new research, and for helping the completion of requirements for both a Yale medical student to graduate and a clinical scholar fellow to publish.

Debra Howard
Yale Medical Student

Joann Elmore, M.D., MPH
Robert Wood Johnson Clinical Scholar

LUX ET VERITAS

Alvan R. Feinstein, M.D.
Sterling Professor of Internal
Medicine and Epidemiology

Carol Lee, M.D.
Assistant Professor of
Diagnostic Radiology