AD_____

Award Number: DAMD17-03-1-0606

TITLE: Time-Series Analysis of Human Interpretation Data in Mammography

PRINCIPAL INVESTIGATOR: Craig A. Beam, Ph.D.
Emily F. Conant
Harold L. Kundel
Ji-Hyun Lee
Patricia A. Romily
Edward A. Sickles

CONTRACTING ORGANIZATION: Moffit Cancer Center
Tampa, Florida 33612

REPORT DATE: January 2005

TYPE OF REPORT: Final

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 2005	Final (30 Sep 03 - 29 Dec 04)

4. TITLE AND SUBTITLE
Time-Series Analysis of Human Interpretation Data in Mammography

5. FUNDING NUMBERS
DAMD17-03-1-0606

6. AUTHOR(S)
Craig A. Beam, Ph.D., Emily F. Conant, Harold L. Kundel, Ji-Hyun Lee, Patricia A. Romily, Edward A. Sickles

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Moffit Cancer Center
Tampa, Florida 33612

E-Mail: cbeam@mcw.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

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

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

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

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

See attached.

14. SUBJECT TERMS
No subject terms provided.

15. NUMBER OF PAGES
14

16. PRICE CODE

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

Table of Contents

Cover.....1

SF 298.....2

Table of Contents.....3

Abstract.....4

Background.....5

Procedures and Results.....6

Discussion.....13

References.....14

Appendices.....None

SCIENTIFIC REPORT OF DAMD17-03-1-0606: TIME-SERIES ANALYSIS OF HUMAN INTERPRETATION DATA IN MAMMOGRAPHY

Craig A. Beam, University of South Florida, Moffitt Cancer Research Center & Institute; **Emily F. Conant**, University Of Pennsylvania; **Harold L. Kundel**, University of Pennsylvania; **Ji-Hyun Lee**, University of South Florida, Moffitt Cancer Research Center & Institute; **Patricia A. Romily**, University of South Florida; **Edward A. Sickles**, UCSF.
Primary author's email: beamca@moffitt.usf.edu

ABSTRACT

INTRODUCTION: Recent research has documented that the human observer is a significant source of interpretation errors in mammography in the U.S. However, it has yet to be determined whether or not the rate or likelihood of radiologist-specific error changes across the length of time the radiologist has been reading during a single session, or across the cumulative time the radiologist reads in a year. The purpose of this study was to apply basic methods from the statistical analysis of time series in order to gain novel insights into the characteristics of the human interpretation of mammograms.

PROCEDURES: We applied exploratory statistical time-series analysis methods to describe radiologist performance data across time in three data sets from: (a) a visual scanning study, (b) an interpretation performance study and, (c) from audit data collected at a large screening program over a five-year period.

RESULTS:
Perception Data: Initial analysis of visual scanning data across time revealed clearly defined "epochs" of visual "sampling" between two views of the standard mammogram. A final "epoch" was observed to be characterized by rapid sampling across the two views in proximity to the target.
Interpretation Data: We obtained data from 110 radiologists reading 148 screening mammograms in a reading experiment. The mammograms were presented on 8 mammoviewers. We computed the true positive fraction (tp) (the proportion of breast cancer cases given a recommendation for recall) and the false positive fraction (fp) (the proportion of women without breast cancer who were given a recommendation for recall) for the interpretations given at each mammoviewer. Our data show variability in temporal patterns among the radiologists. For example, the time series graph of one radiologist showed a declining trend in both tp and fp, suggesting an increase in the threshold used by the radiologist to recommend callback-i.e., the radiologist appears to become more stringent with their callbacks. Another radiologist was very constant in having nearly perfect true positive proportion (tp=1.0) across the 8 mammoviewers. Interestingly, this reader's fp was low, indicating the reader had high skill that was consistent throughout the reading experiment.
Audit Data: Presently, radiologists who interpret mammograms are required by federal law to track the outcome of the cases recalled at screening for further work-up. Visual examination of the trends of two measures of the performance of a large screening program over a five year period suggests that whereas the proportion of women recalled at screening might have been stable, the "yield" of the screening program (i.e., the proportion of those called back from screening who were determined to have breast cancer-often referred to as the "Positive Predictive Value" or "PPV") was much more variable.

CONCLUSIONS: Our study found time-related patterns to the interpretation of mammograms. This research is important to breast cancer research and to the breast cancer advocate since it opens new opportunities for improving the early detection of breast cancer by delineating basic trends that heretofore have not been known.

1. Background

It is estimated that at least half of the errors made in clinical medicine are perceptual. Recent research has documented that the human observer is a significant source of errors in mammography in the US. In recognition of the ability of humans to alter decision thresholds in experimental and clinical settings, Receiver Operating Characteristic (ROC) analysis has become the preferred methodology for evaluating sources of interpretive error and comparing performance. However, current ROC methodology is very limited for the analysis of images having multiple responses and/or multiple targets, a situation often reflective of clinical reality. The limitations of current methodology have undoubtedly limited scientific efforts aimed at reducing human error in mammogram interpretation.

An important, but apparently largely under-appreciated, fact is that radiological interpretation of mammograms is an activity that occurs across time. It is estimated that at least half of the errors made in clinical medicine are perceptual[1]. The phenomenon of "Satisfaction of Search", in which readers cease searching the image once a finding is encountered, has been established experimentally as a source of error[2]. Other time-course studies of scanning have correlated expertise and error with search time[3]. However none of these studies have investigated whether the time-series covariance structure of the scanned image, measurable by the autocorrelation function, might provide a novel signature of expertise, with experts more able to organize the image via scanning into truly statistically independent components. Recent research has documented that the human observer is a significant source of interpretation errors in mammography in the US[4]. However, it has yet to be determined whether or not the rate or likelihood of radiologist-specific error changes across the length of time the radiologist has been reading during a single session, or across the cumulative time the radiologist reads in a year. Although radiologists interpreting mammograms are now required by Federal Law to audit the outcome of their positive calls, we know of no comprehensive analysis of the statistical patterns in this data, which may or may not yield unexpected structure when viewed against time. We hope that the "don't buy a car made on a Monday or a Friday" admonition does not hold for mammogram interpretation-but we will not know until we consider the profile of audit data across time.

A statistical time-related characterization of human interpretation error in mammography is largely absent. The purpose of this project is to apply basic methods from the statistical analysis of time series in order to gain novel insights into the characteristics of the human interpretation of mammograms. In doing so, we anticipate hypotheses will be discovered which could then lead to new avenues to improve mammographic screening.

2. Procedures and Results

Throughout the next sections, we report the results of applying graphical methods from the analysis of time series to mammography perception an interpretation data. Each dataset is described in the relevant section below.

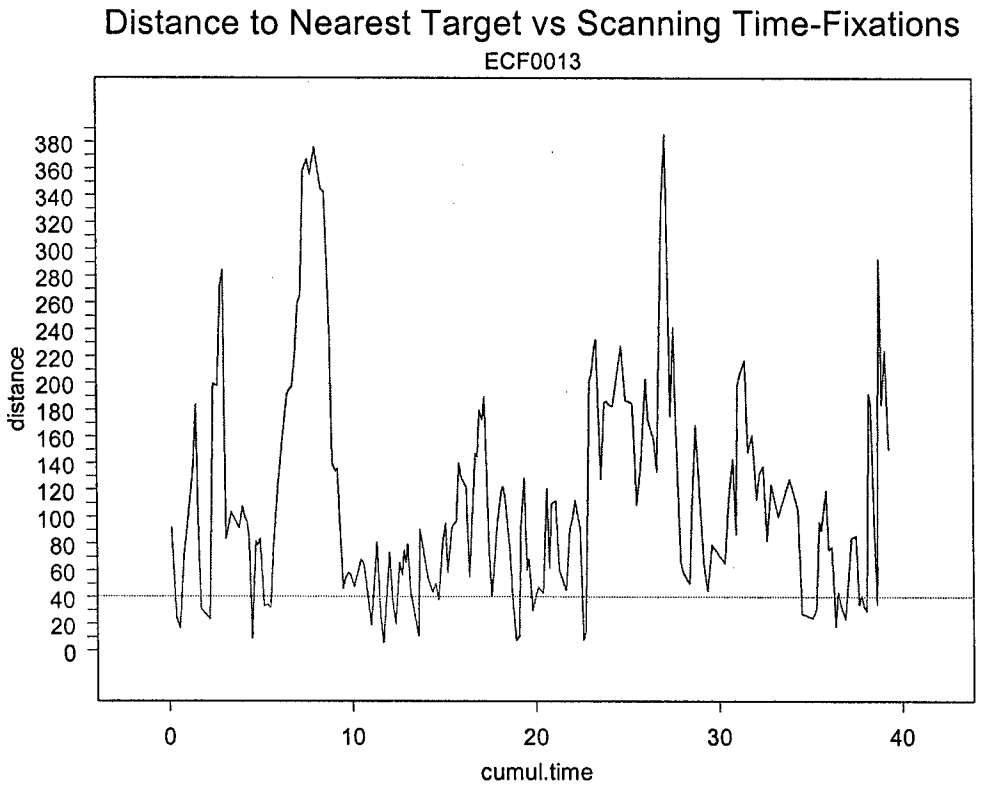
2.1. Perceptual Data

The purpose in recording eye position is to determine where the reader is directing visual attention in the displayed image (8). It is assumed that the center of attention on the image is indicated by the axis of the gaze. Properly calibrated eye position recording can relate the location of the axis of the gaze to locations in the displayed image. The resolving power of the retina is greatest in the fovea, which is a small central region of the retina on the axis of the gaze. Resolving power decreases exponentially toward the periphery. Consequently, detail is seen best on the axis of the gaze. Accurate location of the position of the axis of the gaze on the displayed image requires careful calibration and either monitoring eye-position with the head immobilized or monitoring both head and eye position.

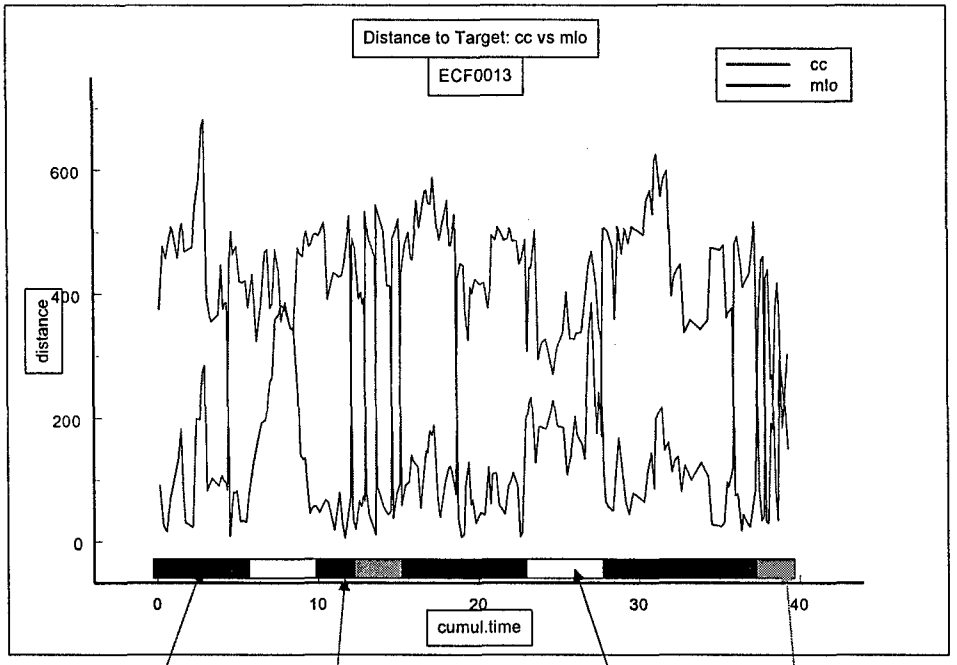
The scan path traced over the scene by the axis of the gaze consists of a series of rapid jumps (traditionally called saccades or macro-saccades) with intervening fixations when the gaze is relatively stationary. During fixation the axis of the gaze drifts and there are small corrective micro-saccades.

The raw eye-position data consist of a stream of (x,y) coordinates acquired at 50 to 60 samples per second. The data are reduced to "fixations" (x,y,t) using a nearest neighbor clustering algorithm that finds the geometric mean of (x,y) and sums the sample time. If the sample time "t" is less than 60 ms (3 raw data points), the point is considered to be part of a saccade and not part of a fixation. Blinks give spurious data that is easily recognized and are removed by a filter.

The mammogram analyzed for this report contained one lesion or "target". The distance from each fixation to the target was computed and displayed as a time series blow.



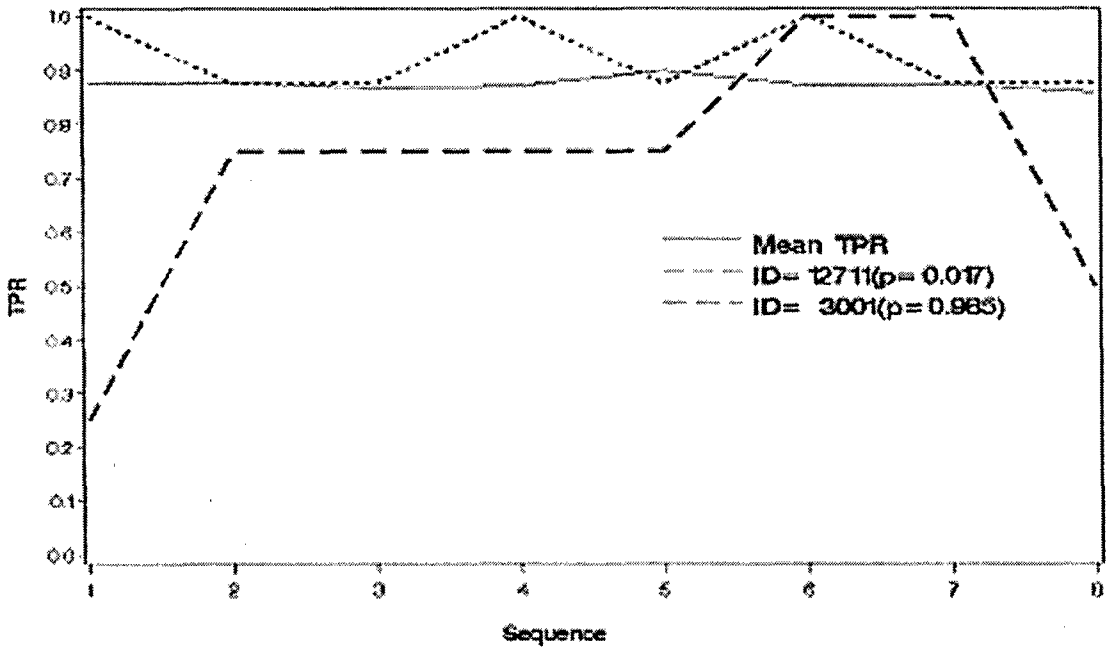
Since the mammogram consisted of two views (cc and mlo), the target appeared twice. In the graph below, we have taken the distance nearest to the two targets from each fixation.

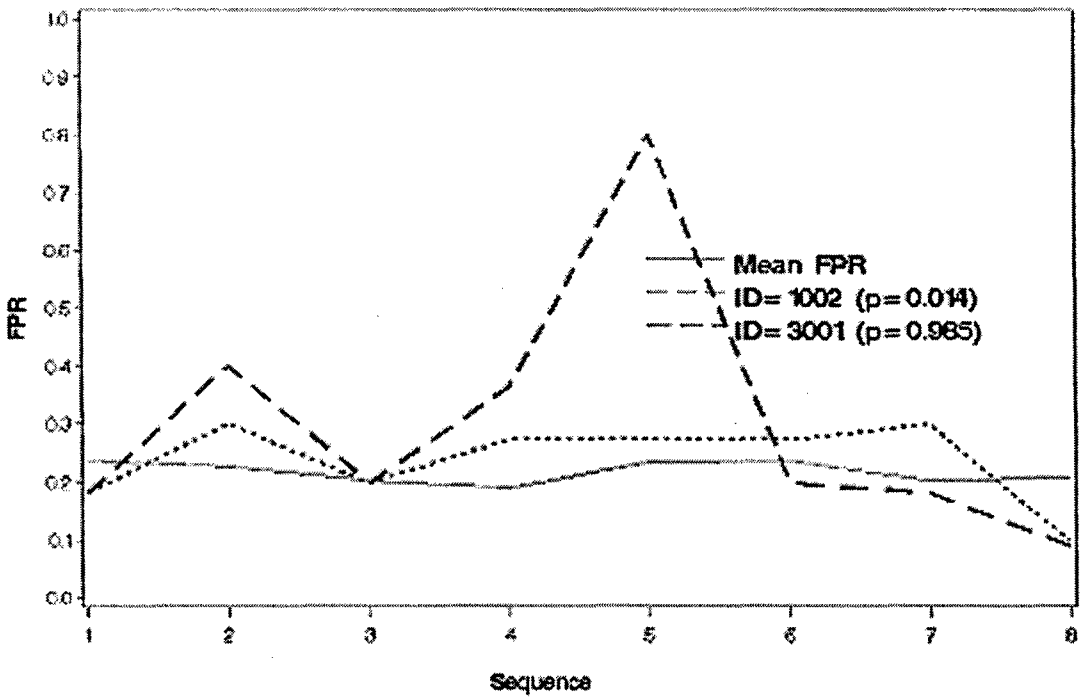


Initial analysis of visual scanning data for each of the two views separately across time revealed clearly defined "epochs" of visual "sampling" between two views of the standard mammogram. A final "epoch" was observed to be characterized by rapid sampling across the two views in proximity to the target:

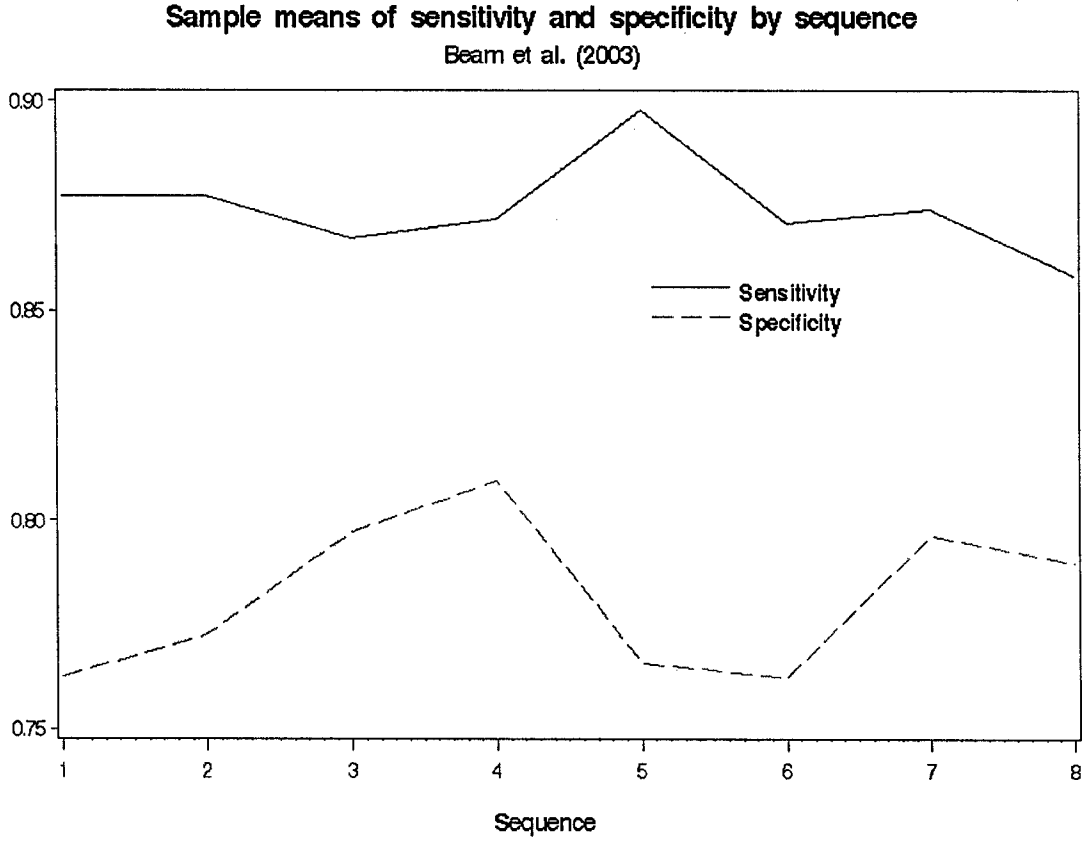
2.2. Interpretation Data

We obtained data from 110 radiologists reading 148 screening mammograms in a reading experiment. The mammograms were presented on 8 mammoviewers. We computed the true positive fraction (tp) (the proportion of breast cancer cases given a recommendation for recall) and the false positive fraction (fp) (the proportion of women without breast cancer who were given a recommendation for recall) for the interpretations given at each mammoviewer. Our data show variability in temporal patterns among the radiologists. For example, the time series graph of one radiologist showed a declining trend in both tp and fp, suggesting an increase in the threshold used by the radiologist to recommend callback-i.e., the radiologist appears to become more stringent with their callbacks. Another radiologist was very constant in having nearly perfect true positive proportion (tp=1.0) across the 8 mammoviewers. Interestingly, this reader's fp was low, indicating the reader had high skill that was consistent throughout the reading experiment.


TPR
Sequence
Mean TPR
ID= 12711(p= 0.017)
ID= 3001(p= 0.985)


FPR
Sequence
Mean FPR
ID= 1002 (p= 0.014)
ID= 3001 (p= 0.985)

The following graph depicts the behavior of the mean of the group of radiologists vs. sequence. Interestingly, there is a distinct spike in both mean sensitivity and specificity. In addition, a "trough" in mean specificity is observed at the time of the spike in mean sensitivity. This could come about through changing threshold selection. Relaxing the threshold to gain sensitivity also increases the rate of false positives, and thus decreases specificity. Yet, it is important to point out that some of the other segments of the two curves do not exhibit this "tradeoff" relationship. For example, from the 3rd to the 4th sequence, both mean sensitivity and mean specificity increase. This parallelism is also seen from the 6th to 7th sequence. Parallelism in apparent decreases in mean sensitivity and mean specificity was observed in the last two sequences (7th to 8th).

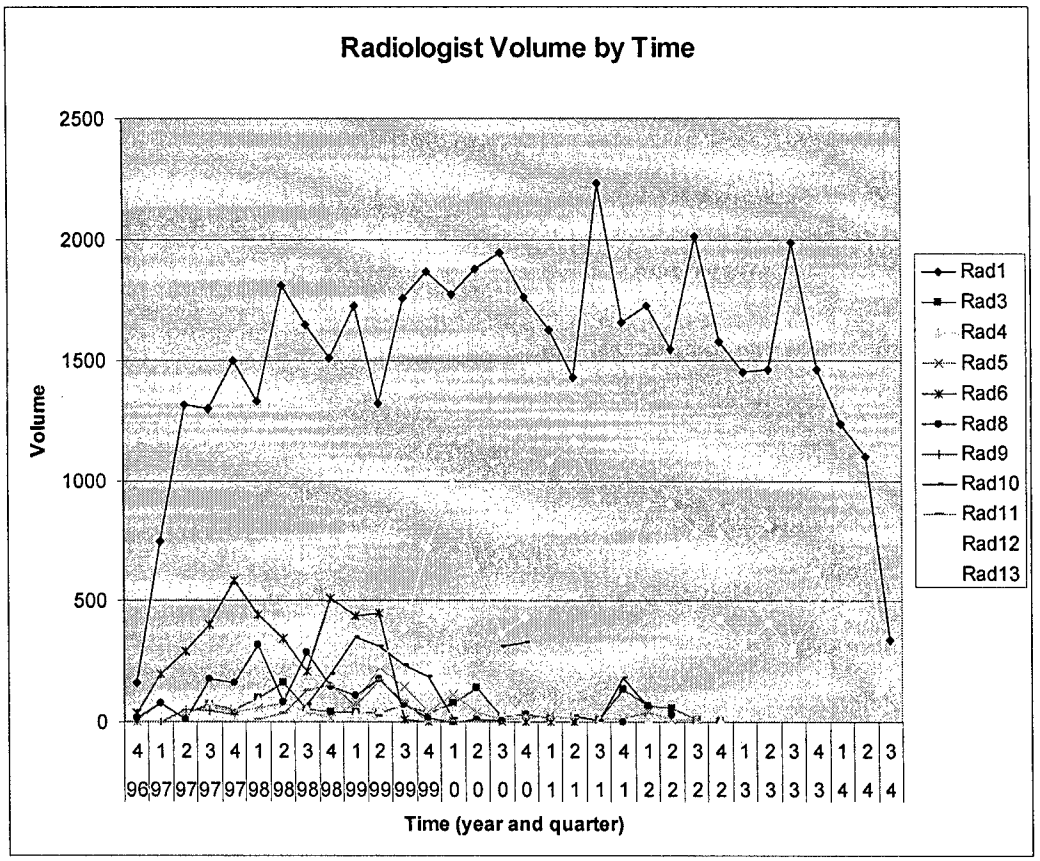


2.3. Audit data

Presently, radiologists who interpret mammograms are required by federal law to track the outcome of the cases recalled at screening for further work-up. We acquired data from a large screening program. The data was collected over 7 years for 13 radiologists. The data analyzed consist of the volume and the recall for each radiologist by quarter of the year. Visual examination of the trends of two measures of the performance of a large screening program over a five year period suggests that whereas the proportion of women recalled at screening might have been stable, the "yield" of the screening program (i.e., the proportion of those called back from screening who were determined to have breast cancer-often referred to as the "Positive Predictive Value" or "PPV") was much more variable.

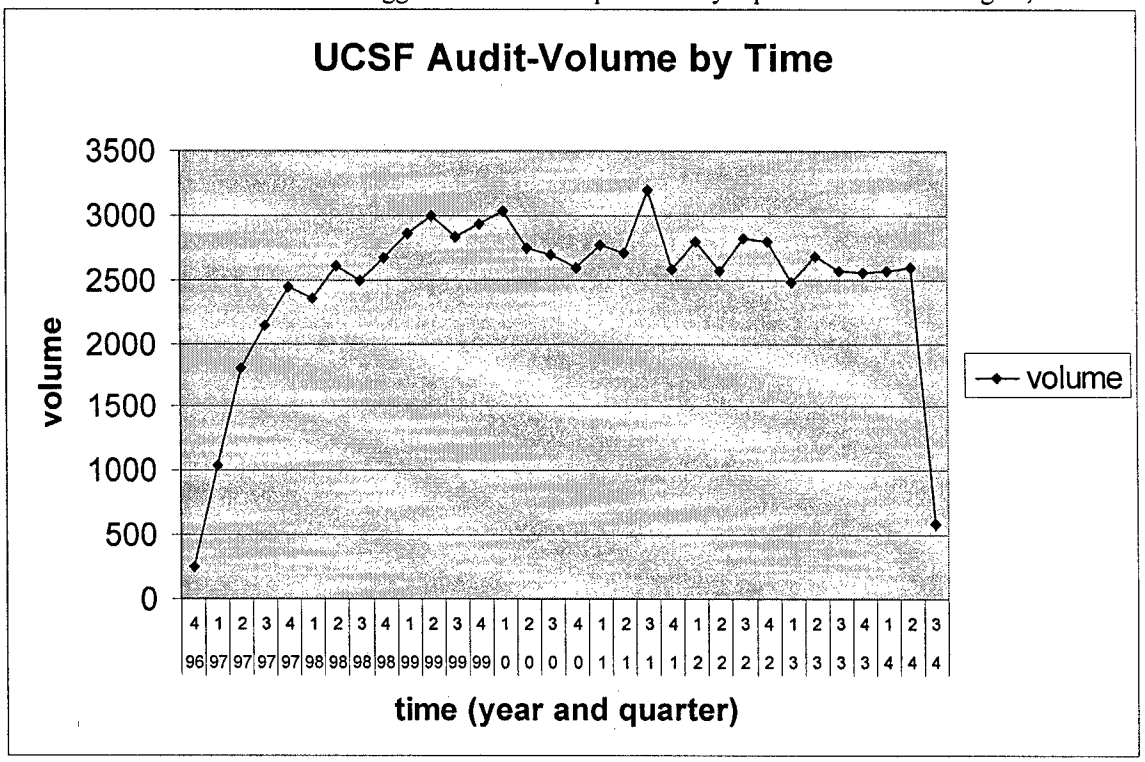
2.3.1. Radiologist Volume Across Time

This graph depicts radiologist volume across time. It shows that the majority of the volume is attributable to two radiologists. If we ignore the values at the ends of the series (they might be incomplete observations), we note a declining trend during the latter time periods, which was preceded by an initial increasing trend. It is also important to observe that the other radiologists terminated reading before this downturn. However, one new radiologist began near the end of the time period.



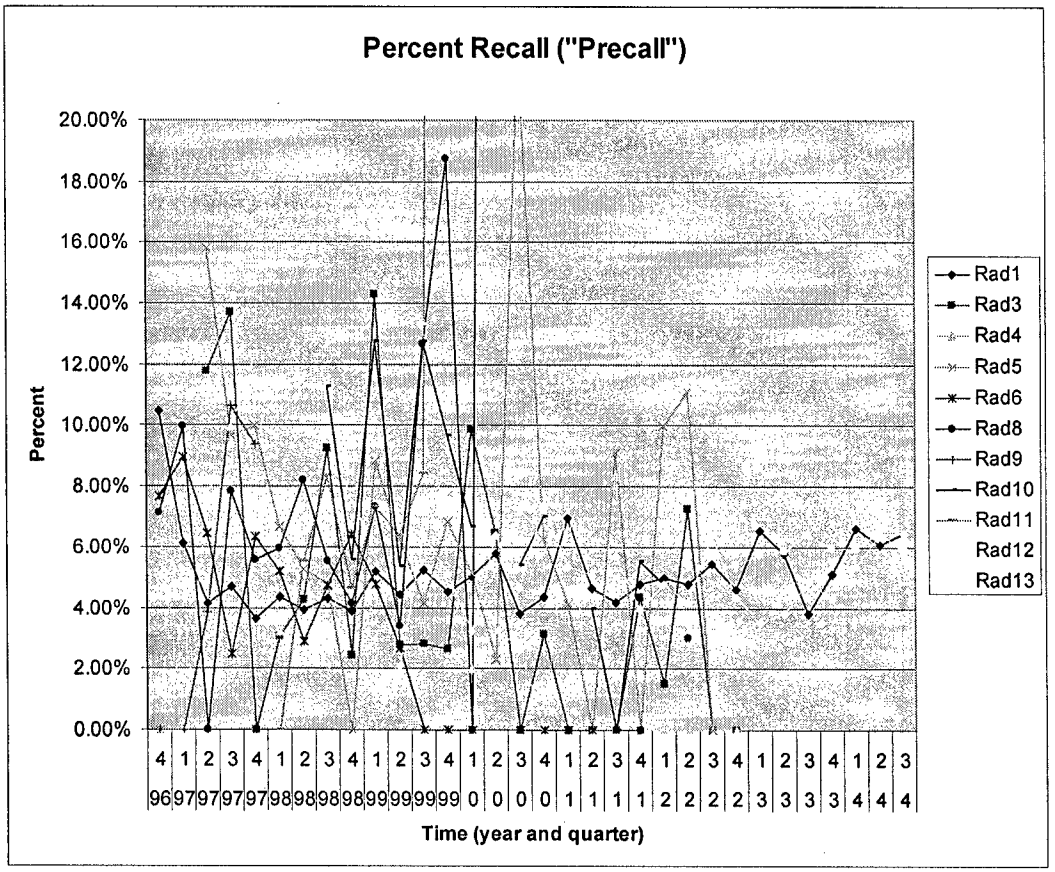
2.3.2. Practice Volume Across Time

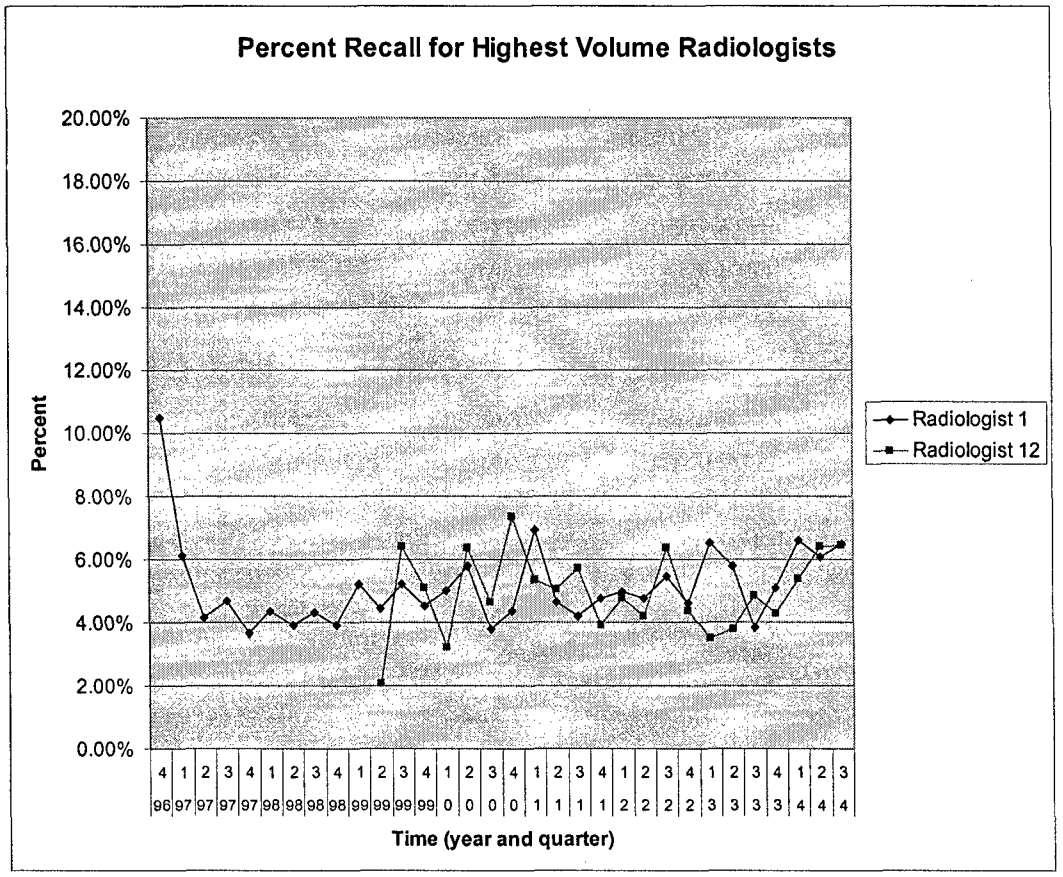
It can be readily observed from the plot of the entire practice across time (i.e., the sum of the volumes of the individual radiologists) that there is indeed a downward trend suggested which was preceded by a period of increase. Again, it is important to exclude from consideration the two extremes, starting and ending, since they are probably incomplete.



In addition, a stabilization period is suggested at the very end of the time period and a spike in volume apparently occurred in the 3rd quarter of 2001. This is the quarter in which the events of "9/11" occurred. It might be worthwhile investigating whether or not this finding of a spike in volume soon after September 11, 2001 is observed in other mammography practices as well as in other healthcare services.

2.3.3. Radiologist Recall Across Time





3. Discussion

Our study found time-related patterns to the interpretation of mammograms. This research is important to breast cancer research and to the breast cancer advocate since it opens new opportunities for improving the early detection of breast cancer by delineating basic trends that heretofore have not been known.

We are now working on statistical modeling of this data and expect more insights will be gained.

4. References

1. Kundel HL, Nodine CF, Carmody DP. Visual scanning pattern recognition and decision making in pulmonary nodule detection. Investigative Radiology 1978; 13:175-181.
2. Krupinski EA, Kundel HL, Judy PF, Nodine CF. Key issues for image perception research. Radiology 1998; 209:611-612.
3. Nodine CF, Kundel HL, Lauver SC, Toto LC. Nature of expertise in searching mammograms for breast masses. Academic Radiology 1996; 3:1000-1006.
4. Beam CA, Conant EF, Sickles EA, Factors affecting radiologist variability in screening mammography. Academic Radiology, 2002; 9: 531-540.
5. Sickles EA. Auditing your practice. In Kopans DB, Mendelson EB, eds. A categorical course in breast imaging. Oak Brook, IL, Radiological Society of North America, 1995:81-91.