Published in final edited form as:
AJR Am J Roentgenol. 2012 April ; 198(4): 970–978. doi:10.2214/AJR.11.6988.

Association Between Time Spent Interpreting, Level of Confidence and Accuracy of Screening Mammography

Patricia A. Carney, PhD[1], **Andy Bogart, MS**[2], **Berta M. Geller, EdD**[3], **Sebastian Haneuse, PhD**[2], **Karla Kerlikowske, MD**[4], **Diana SM Buist, PhD**[2], **Robert Smith, PhD**[5], **Robert Rosenberg, MD**[6], **Bonnie C. Yankaskas, PhD**[7], **Tracy Onega, PhD**[8], and **Diana L. Miglioretti, PhD**[2,9]

[1)] Departments of Family Medicine and Public Health and Preventive Medicine, Oregon Health & Science University, Portland, OR

[2)] Group Health Research Institute, Seattle, WA

[3)] Office of Health Promotion Research, University of Vermont, Burlington, VT

[4)] Department of Medicine, University of California, San Francisco, San Francisco, CA

[5)] Director of Cancer Screening, American Cancer Society, Atlanta, GA

[6)] Department of Radiology, University of New Mexico, Albuquerque, NM

[7)] Department of Radiology, University of North Carolina, Chapel Hill, NC

[8)] Department of Community & Family Medicine, Dartmouth Medical School, Lebanon, NH

[9)] Department of Biostatistics, University of Washington, Seattle, WA

Abstract

Purpose—To examine the effect of time spent viewing images and level of confidence on a screening mammography test set on interpretive performance.

Materials and Methods—Radiologists from six mammography registries participated in the study and were randomized to interpret one of four test sets and complete 12 survey questions. Each test set had 109 cases of digitized four-view screening film-screen mammograms with prior comparison screening views. Viewing time for each case was defined as the cumulative time spent viewing all mammographic images before recording which visible feature, if any, was the "most significant finding". Log-linear regression fit via GEE was used to test the effect of viewing time and level of confidence in the interpretation on test set sensitivity and false-positive rate.

Results—119 radiologists completed a test set and contributed data on 11,484 interpretations. Radiologists spent more time viewing cases that had significant findings or for which they had less confidence in interpretation. Each additional minute of viewing time increased the probability of a true positive interpretation among cancer cases by 1.12 (95% CI: 1.06, 1.19, $p<0.001$), regardless of confidence in the assessment. Among radiologists who were 'very confident' in their assessment, each additional minute of viewing time increased the adjusted risk of a false positive interpretation among non-cancer cases by 1.42 (95% CI 1.21, 1.68), and this viewing-time effect diminished with decreasing confidence.

Conclusions—Longer interpretation times and higher levels of confidence in the interpretation are both associated with higher sensitivity and false positive rates in mammography screening.

Correspondence to: Patricia Carney, PhD Department of Family Medicine Oregon Health & Science University 3181 SW Sam Jackson Park Rd. Mail Code: FM Portland, OR 97239 hone: 503-494-9049 Fax: 503-494-2746 carneyp@ohsu.edu.

INTRODUCTION

Little is known about how time spent examining different types of mammographic images affects interpretive accuracy outside of comparisons between digital and screen film mammography or mammography with and without use of computer-aided detection (1-8). In one study, Saunders, *et al* (1, 2) found that incorrect detection decisions for both cancer and non-cancer cases, and incorrect decisions about needed work-up were *both* associated with longer interpretation time for screening mammography. In another study, Nodine, *et al* (3) found that a high level of confidence was associated with both shorter fixation dwell times (time spent looking at a specific area on a film and time spent initially scanning the image, and that one second of fixation dwell time (coupled with a high level of confidence) versus time spent initially scanning the image were associated with detection of true positive lesions for experienced radiologists. They also found that prolonging the search beyond the global recognition time yielded few new lesions and increased the risk of error. Kundel, *et al* (4) found that expert radiologists fixated on a cancer in 1.13 seconds and that proficient radiologists appear to use a fast holistic scanning mode rather than a search-to-find-mode. Understanding how time spent interpreting mammography affects performance could assist radiologists in avoiding viewing behaviors unlikely to improve accuracy.

Weaknesses of all the above studies include that very few experienced radiologists were included (between 1 and 6), so adjustment for radiologists characteristics known to affect performance were not addressed, and none of the studies examined the particular features being examined in either normal or abnormal images. Understanding the relationships among time spent, complexity of the images and interpretive accuracy, while adjusting for possible confounders, could aid in identifying which types of mammographic findings might benefit from a second opinion. In addition, conducting an in-depth assessment of initial radiologists' time spent and confidence in their assessment could potentially improve interpretive performance and be more effective than double reading all screening mammograms. We conducted a study with radiologists across the U.S. to examine these issues and to specifically test the hypothesis that more versus less time spent interpreting mammograms is associated with more difficult cases and lower accuracy compared to cases interpreted more quickly.

METHODS

Study Population

This study was conducted with six mammography registries (Carolina Mammography Registry, New Hampshire Mammography Network, New Mexico Mammography Project, Vermont Breast Cancer Surveillance System, and Group Health Cooperative in western Washington) associated with the National Cancer Institute funded Breast Cancer Surveillance Consortium (BCSC; http://breastscreening.cancer.gov). Data collected as part of this study were pooled at the BCSC Statistical Coordinating Center (SCC) in Seattle, WA for analysis. Each registry and the SCC received IRB approval for either active or passive consenting processes or a waiver of consent to enroll participants, link data, and perform analytic studies. All procedures are Health Insurance Portability and Accountability Act compliant and all registries and the SCC have received a Federal Certificate of Confidentiality (9) and other protection for the identities of women, physicians, and facilities that are subjects of this research. In addition, each registry and the SCC received IRB approval for all test set study activities.

Radiologists who interpreted mammograms at a facility contributing to any of the registries between January 2005 and December 2006 were invited to participate. We also invited 103 non-BCSC radiologists from Oregon; Puget Sound, WA; North Carolina, San Francisco, and

New Mexico. A total of 469 radiologists were invited to participate, and 148 (31.6%) consented. Among these, 119 (80.4%) completed all study activities.

Test Set Design and Development

We selected test set cases based on cancer prevalence and expert rated difficulty identifying breast cancer to create four screening mammography test sets with 109 cases in each set. This approach was used because the goal of the larger study was to assess how cancer prevalence and type of finding (subtle, intermediate and obvious) as interpreted on a test set would correlate with actual clinical practice. The results of the larger study will be reported elsewhere.

All cases were randomly selected from screening examinations performed on women aged 40 to 69 between 2000 and 2003 from the six participating BCSC mammography registries. Women who had a mastectomy and those with a prior history of breast cancer were excluded. Participating registries contributed between 42 and 84 screening mammography examinations, all of which had a mammogram within the prior 11-30 months. Of these approximately 26-48% were selected from each site for use in a test set. Examinations with stray marks or other quality issues on the films were excluded. Each screening examination selected consisted of craniocaudal (CC) and mediolateral oblique (MLO) views of each breast (4 views per woman for each of the screening and comparison examinations). For cancer cases, we selected images from exams for which cancer was diagnosed within 12 months following the mammogram. Non-cancer cases came from women who were cancer after at least two years following the mammogram. Final cases in the test sets came from 36 women known to have been diagnosed with cancer within one year of imaging, and from 94 women who remained cancer-free for two years following the imaging. Cases from these women were used more than once to configure the test sets appropriately.

The case sampling design was stratified based on clinical interpretations as true positive, true negative, false positive and false negative and were reviewed by an expert panel of radiologists (n=3), which was blinded to the original mammography interpretation and cancer status. The experts also categorized significant findings as mass, calcification, asymmetric density or architectural distortion; and as obvious, intermediate or subtle. Obvious findings were defined as those the expert panel agreed that 100% of community radiologists should identify. Intermediate finding were defined as those the expert panel agreed 25-99% of community radiologists would identify. Subtle findings were defined as those the expert panel indicated <25% of community radiologists would identify. The experts reached consensus on any interpretation for which an initial disagreement occurred.

We randomly selected 60 TP examinations and 16 FN examinations, so the experts could identify 14 obvious cancers, 15 intermediate cancers, and 9 subtle cancers for inclusion in a test sets. To include FP examinations, we selected examinations with BI-RADS assessment categories 0, 4 or 5 that were not associated with breast cancer within 24 months of mammography. The remaining examinations were TN examinations of both breasts. The composition of Test Sets 1 and 2 included: 47% obvious, 40% intermediate, 13% subtle, and the composition of Test Sets 3 and 4 included 20% obvious, 50% intermediate, 30% subtle. Test sets 1 and 2 each contained 15 cancer and 94 non-cancer cases. Test sets 3 and 4 both contained 30 cancer and 79 non-cancer cases. After cases were selected, the films were digitized by experts at the American College of Radiology using a Vidar Diagnostic Pro (10) and loaded into specially designed viewing software that allowed us to collect timing and location information associated with interpretations.

Test Set Administration

We randomized consenting radiologists to one of the four test sets. The test sets were self-administered using custom designed software distributed on a DVD. Participants were block randomized within strata defined by registry/site and whether a radiologist had reviewed at least 30 cancers in the BCSC database to ensure an equal number of radiologists with accurate measures of clinical sensitivity of mammography were reading each test set. This criterion was not used for non-BCSC participants.

Each site sent consenting radiologists the DVD along with an instruction sheet informing them of their assigned test. Radiologists used either a home or work computer or laptop provided by the study with a large size screen and high-resolution graphics (≥1280×1024, ≥3GHz, 1GB of RAM, and a video card with 128MB of memory capable of displaying full 32-bit color at the listed resolutions and a DVD reader) to show 2 images at the same time. The monitor specifications were provided to radiologists if they chose to use their home or work computers. The software developed for the study allowed radiologists to: 1) choose whether the images were displayed with right breast facing right or left and left breast facing left or right; 2) rapidly toggle (≤1 second) between the display of paired images so that visual memory is retained from one displayed pair to the next; 3) magnify a portion of the displayed image; and 4) point and click on any important abnormality to record the coordinates of findings to enable capture of whether a radiologist has identified and located the lesion of highest suspicion for cancer.

Participating radiologists were instructed to interpret test sets as they would in clinical practice. Radiologists were informed that the overall cancer rate on test sets was higher than that found in a screened population (11), but they were not informed of the specific prevalence of positive examinations or cancers in the test sets. We used this approach so that all radiologists would interpret test sets with similar knowledge of the underlying prevalence of disease instead of assuming the prevalence to be similar to their own clinical experience.

Measures

Participant Characteristics—Prior to evaluating test set digitized mammography films, the software prompted each participant to answer 12 demographic and clinical practice survey questions, including receipt of fellowship training, specialization, number of years spent interpreting mammograms, and the number of mammograms interpreted per week. In addition, we made use of the test set assignment as a radiologist-level characteristic, as it captured the case mixture and difficulty of films.

Outcome Measures—Our analyses focus on mammographic test set performance. As participants viewed individual cases in the test set, they were prompted to identify the most significant visible breast abnormality, and to decide whether or not the patient should be recalled for additional work-up. The decision to recall constituted a positive test result for our analyses. Recall decisions on mammography exams were modeled conditionally based on the patients' true cancer status, and other relevant covariates described in the data analysis section, to estimate effect of the time spent viewing films on sensitivity and false-positive probability.

Time Measurement and Image Viewing Process—The test-set software randomly presented the images in a similar manner to digital mammogram interpretation using a single monitor. Each case was presented in a sequence including MLO and CC views of both breasts simultaneously, followed by MLO and CC views of each breast paired with the analogous image from the previous exam to assess whether changes from the prior mammogram were apparent. Figure 1 illustrates an example case with image presentation

shown from the Test Set software. Images could be magnified as needed. The software recorded the length of time, measured in seconds, the user spent viewing each individual film, which was defined as the cumulative time spent viewing all mammographic images and identifying which, if any, visible feature was the most significant finding by a mouse click.

Radiologists were encouraged by a pop-up message within the program to examine all available images, including both current and comparison views, before indicating their decisions about any of the individual images to ensure viewing consistency during the study. Because the assessment software did not have a pause feature, any time spent away from the computer during the completion of the test set was added to the cumulative time associated with the most recently started exam. To minimize the impact on our analyses, we assessed viewing time in a controlled setting with seven radiologists viewing 320 cases where interrupted viewing time was not possible and found that >98% of interpreters completed viewing all study images for each case within five minutes, which our expert radiologists concurred with. Examinations for which viewing time exceeded 5 minutes (n=1,443) were excluded from our analyses since they likely represented a mix of uninterrupted and interrupted viewing durations.

Other Exam-Level Characteristics—Breast density was categorized as almost entirely fat, <25%; scattered fibroglandular densities, 25–50%; heterogeneously dense, 51–75%; or extremely dense, >75% (12). Users reported confidence in their assessment on each exam as either not at all confident, not very confident, neutral, confident, or very confident. For our analyses, we combined the responses for ‘Not at all confident’ and ‘Not very confident’ to form a ‘Not Confident’ category.

We recorded the expert-assessed lesion type for each case as one of either mass; calcification, asymmetry, or architectural distortion. This variable was classified as missing when the expert consensus indicated there was no significant finding.

Data Analysis

All but one radiologist, who reviewed only 104 cases, reviewed all 109, resulting in 12,966 interpretations. Of these, 1,443 (11%) observations were excluded because their duration exceeded five minutes, and an additional 39 (0.3%) were excluded due to errors in time recording caused by computer problems. In all, 11,484 (89%) interpretations were suitable for analyses.

We calculated frequency distributions for responses to the 12 demographic questions by test set assignment. To address the primary scientific question of the effect of viewing time on radiologist’s test set performance, we modeled the relative risk of a positive assessment (recall) using log-linear regression. To account for the correlation within both radiologists and exams, we implemented an extension of GEE developed specifically for analysis of non-nested multilevel data (13, 14).

We regressed a binary indicator of recall on viewing time and examination-level, patient-level, and radiologist-level covariates using a log-link function to estimate relative risks. Our models made a Poisson variance assumption, which yields valid variance estimates for relative risk in analyses of common binary outcomes (15), and applied the robust Huber-White sandwich variance estimator. We separately modeled sensitivity and false-positive rates. For both cancer and non-cancer exams, we modeled the probability of recall as a function of viewing time, with and without adjustment for: radiologist’s confidence on the exam; the radiologist’s assessment of breast density; the expert-identified lesion type; and the radiologist’s fellowship category, specialization, years interpreting mammograms,

number of mammograms read per week, and random test set assignment. We expressed viewing time in the model using a single linear term, and expressed each categorical covariate with an appropriate group of indicator variables.

We hypothesized that the effect of viewing time on the probability of recall may differ according to the radiologist's confidence interpreting the exam and assessed whether confidence modified the effect of viewing time on the risk of recall by including interaction terms between viewing time and each indicator associated with confidence. In analyses where interactions were statistically significant, we calculated confidence-level specific estimates of the relative risk of recall associated with a one-minute increase in viewing time. Where the interaction was not significant, we omitted it from the regression models, and estimated a time effect across levels of confidence. All analyses were conducted using the R statistical software, version 2.10.0 (16, 17).

RESULTS

Seventy-six of 119 participating radiologists (64%) reported interpreting mammograms for more than 10 years, and 86 (72%) reported reading at least 50 mammograms per week (Table 1). Fifteen radiologists (13%) reported completion or plans to complete a fellowship in breast or women's imaging.

Figure 2 presents the mean and inter-quartile range (IQR) for viewing time for the 2,291 exams of images belonging to 36 women known to have been diagnosed with cancer within one year of imaging, and for 9,193 exams of images from 94 women who remained cancer-free for two years following the imaging. Results are shown by cancer status and the confidence level each radiologist selected following the exam, with horizontal reference lines indicating the 25^{th}, 50^{th}, and 75^{th} percentiles of the viewing time distribution among cancer and non-cancer exams. Median times associated with exams resulting in a positive assessment, shown with a solid vertical bar, were higher than negative exams, shown with a dotted bar, for cancers and non-cancers and across all levels of confidence. Among both cancer and non-cancer exams, median viewing times were shorter for exams on which radiologists endorsed greater confidence, and longer for exams on which readers were less confident.

Table 2 illustrates the median viewing times and IQRs for all examinations in groups defined by expert-identified finding type and participant-rated BI-RADS (12) breast density. Median viewing times for exams containing any expert findings were longer than for those with none, with the exception of those containing a mass. Fatty and extremely dense breasts have similar properties in terms of time spent viewing. Both have shorter viewing times than scattered and heterogeneously dense for cases with no findings and most cases with findings. Among exams with expert findings, those containing masses corresponded to the shortest median viewing times, and calcifications to the longest, except in cases where breasts were almost entirely fatty for which asymmetries had the longest median viewing time.

Among cases with cancer, each additional minute of viewing time increased the adjusted probability of a true positive assessment by a factor of 1.12 (95% CI: 1.06, 1.19), $p<0.001$ (Table 3). This effect did not significantly differ by radiologist confidence in assessment before ($p=0.88$) or after ($p=0.73$) adjustment for expert-identified lesion type, reader-rated breast density, fellowship category, specialization, years interpreting mammograms, mammograms read per week, and random test set assignment. Confidence in the assessment was significantly associated with a true positive assessment both before ($p=0.003$) and after ($p<0.001$) covariate adjustment. Radiologists who reported being very confident in their

assessment were 1.32 (95% CI: 1.16, 1.50) times more likely to have correctly recalled the patient than those reporting neutral confidence.

The relative risk of false positive assessments, estimated from examinations of women who remained free of breast cancer for one year after imaging, illustrated that the relationship between viewing time and false positive probability differed significantly by radiologist confidence (Table 3), both before ($p=0.016$) and after ($p=0.039$) adjustment for the factors mentioned above. The unadjusted association between confidence and the risk of a false positive was statistically significant ($p<0.001$). False positive exams were about half (RR=0.55, 95% CI: 0.36, 0.86) as likely among very confident examiners as among those reporting neutral confidence. Those who reported being not very confident or not at all confident were 1.41 (95%CI: 1.17, 1.69) times more likely to recall the patient.

The effect of viewing time on the probability of a false positive assessment was significant across all confidence levels, both before and after covariate adjustment. For those who were 'Very confident' in their assessment, each additional minute of viewing time increased the adjusted risk of a false positive by a factor of 1.42 (95% CI: 1.21, 1.68). Relative risk estimates diminished monotonically according to falling confidence. For those reporting 'Confident' assessments, the risk increased by a factor of 1.40 (95% CI: 1.29, 1.52), for 'Neutral' assessments by a factor of 1.38 (95% CI: 1.29, 1.48), and for those who were either 'Not very confident' or 'Not at all confident' in their assessment, by a factor of 1.20 (95% CI: 1.10, 1.31).

DISCUSSION

This study is the largest conducted to-date with 119 radiologists and 11,484 interpretations analyzed to examine the relationships among viewing time, type of finding, confidence and accuracy when interpreting mammography. We found that radiologists spent more time viewing mammographic findings that they ultimately recalled rather than those that those they did not recall and that higher confidence was usually associated with shorter viewing times. Among cancer cases, increasing viewing time increased the probability of a true positive interpretation. We also found that among non-cancer exams for which radiologists felt 'very confident' in their assessment, each additional minute of viewing time increased the adjusted risk of a false positive interpretation by 42% and that this effect diminished according to decreasing confidence.

These findings illustrate the complex relationship between view time and confidence. While increased viewing time resulted in a small increase in sensitivity, it decreased specificity to a much larger extent. Thus, radiologists may not benefit from spending more time on an interpretation they are not confident in, but may benefit from asking a colleague for a second opinion, which may assist less experienced radiologists in gaining knowledge about and confidence in their interpretations. Addressing confidence in an educational setting may be challenging as most continuing medical education is designed to address knowledge deficits that may or may not exist, and changes in knowledge often do not translate to improvements in skill (18). Interestingly, our findings do not appear to be related to fellowship training, years of experience, or specialization in breast imaging. One might hypothesize that such educational experiences should shorten the time needed to interpret a mammogram, but little exists in the literature on this important topic. Double reading is used extensively outside the U.S. (19, 20), which appears to reduce recall rates without affecting cancer detection. Double reading is not done routinely in the U.S., principally because radiologists are not reimbursed for it; though, double reading on cases that take a long time to interpret might improve specificity.

Our study differs from those of Kundel, Nodine *et* al and others (2, 3, 21-23), who have studied eye position and fixation dwell times using a computer eye-head tracking system. While we also used an interactive computer system that included time assessments, we did not specifically measure eye movements. Like these investigators, we found that longer time spent on the interpretation yielded lower specificity though the 2002 study only included nine radiologists and six of them were trainees (2). Another difference is that these investigators did not include specific measures of confidence as we did in our study.

In another study, Castella *et al* (24) studied the influence of signal variability on human and model observers for detection tasks using simulated masses superimposed on both real patient mammographic backgrounds and synthesized mammographic backgrounds with clustered lumpy backgrounds. They found that human observers' performance did not vary when benign masses were superimposed on real images or on the synthesized background. Uncertainty and variability in signal shape did not significantly affect human performance though variability in signal size did. Our findings differ in that we found level of confidence, a concept reflective of uncertaintly, in the interpretation influenced interpretive accuracy.

Our study shows that interpretation using a test set methodology involves variation in time spent, level of confidence and accuracy. More time spent and lower confidence appears to result in much higher recall rates with many more false positive exams and only a small increase in sensitivity. Interventions that recognize this issue could reduce false positive cases without altering sensitivity. Options such as selective requests for a second opinion should be tested. These might involve academic detailing (university-based educational outreach involving face-to-face education by trained health care professional) (25-27), which has shown improved performance in physicians' use of pharmacologic agents.

The strengths of our study include the large number of participating radiologists from around the U.S and the large number of images our analysis was based upon. Another strength is that we designed the test set software so that it was similar to interpreting digital mammography, which is now being used in greater than 70% of mammography facilities across the U.S., so it is similar to clinical practice (28). Limitations include that physicians could not be kept blind to the fact that their interpretations of the test set were being studied, so their interpretations could have been influenced by the Hawthorne effect (knowledge that they were participating in a study could have affected their interpretive behavior) (29). In addition, though we used the American College of Radiology Guidelines for computer monitor display capabilities for diagnostic radiology (30), image quality in this field is advancing rapidly and monitors did improve over the study period. Similarly, radiologists use of two monitors for interpretive viewing increased over this time period, which may have resulted in variability of study findings. We also used a control setting with seven radiologists to identify a cut point for maximum viewing time and found that >98% viewed the cases within this time period. Only two percent took longer than five minutes in our controlled setting. Though this time might be longer when radiologists interpreted in their home or work settings, we do not think it will have affected our findings to any significant degree.

In conclusion, longer interpretation times and higher levels of confidence appear to be independently associated with a small increase in sensitivity to detect cancers in screening mammography; however, longer interpretation times also appear to be associated with a much greater risk of false positives, and this association increases in magnitude with higher levels of confidence.

Acknowledgments

This work was supported by the American Cancer Society, made possible by a generous donation from the Longaberger Company's Horizon of Hope® Campaign (SIRSG-07-271, SIRSG-07-272, SIRSG-07-273, SIRSG-07-274-01, SIRSG-07-275, SIRGS-06-281, SIRSG-09-270-01, SIRSG-09-271-01, SIRSG-06-290-04), the Breast Cancer Stamp Fund, and the National Cancer Institute Breast Cancer Surveillance Consortium (U01CA63740, U01CA86076, U01CA86082, U01CA70013, U01CA69976, U01CA63731, U01CA70040). The collection of cancer data used in this study was supported in part by several state public health departments and cancer registries throughout the U.S. For a full description of these sources, please see: http://www.breastscreening.cancer.gov/work/acknowledgement.html <http://www.breastscreening.cancer.gov/work/acknowledgement.html> . The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health. We thank the participating women, mammography facilities and radiologists for the data they have provided for this study. A list of the BCSC investigators and procedures for requesting BCSC data for research purposes are provided at: http://breastscreening.cancer.gov/ <http://breastscreening.cancer.gov/> ."

REFERENCES

1. Saunders RS Jr, Baker JA, Delong DM, Johnson JP, Samei E. Does image quality matter? Impact of resolution and noise on mammographic task performance. Medical Physics. 2007; 4(10):3971–3981. [PubMed: 17985642]
2. Saunders RS, Samei E. Improving mammographic decision accuracy by incorporating observer ratings with interpretation time. British Journal of Radiology. Dec; 2006 79(Spec No 2):S117–22. [PubMed: 17209116]
3. Nodine CF, Mello-Thoms C, Kundel HL, Weinstein SP. Time course of perception and decision making during mammographic interpretation. American Journal of Roentgenology. 2002; 179(4): 917–923. [PubMed: 12239037]
4. Kundel HL, Nodine CF, Krupinski EA, Mello-Thoms C. Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms. Academic Radiology. 2008; 15(7):881–886. [PubMed: 18572124]
5. Tchou PM, Haygood TM, Atkinson EN, Stephens TW, Davis PL, Arribas EM, Geiser WR, Whitman GJ. Interpretation time of computer aided-detection at screening mammography. Radiology. 2010; 257(1):40–46. [PubMed: 20679448]
6. Haygood TM, Wang J, Atkinson EN, Lane D, Stephens TW, Patel P, Whitman GJ. Timed effeciency of interpretation of digital and film-screen screening mammograms. AJR. 2009; 192(1): 216–220. [PubMed: 19098202]
7. Berns EA, Hendrick RE, Solari M, Barke L, Reddy D, Wolfman J, Segal L, DeLeon L, Benjamin S, Willis L. Digital and screen-film mammography: Comparison of image acquisition and interpretation times. AJR. 2006; 187(1):38–41. [PubMed: 16794152]
8. Ishlyama M, Tsunoda-Shimizu H, Kikuchi M, Saida Y, Hiramatsu S. Comparison of reading time between screen-film mammography and soft-copied full-field digital mammography. Breast Cancer. 2009; 16(1):58–61. [PubMed: 18836795]
9. Carney PA, Geller BM, Moffett H, Ganger M, Sewell M, Barlow WE, Taplin SH, Sisk C, Ernster VL, Wilke HA, Yankaskas B, Poplack SP, Urban N, West MM, Rosenberg RD, Michael S, Mercurio TD, Ballard-Barbash R. Current Medico-legal and Confidentiality Issues in Large Multi-center Research Programs. American Journal of Epidemiology. 2000; 152(4):371–378. [PubMed: 10968382]
10. American College of Radiology use of Diagnostic Profilm Digitization. http://www.vidar.com/film/diagnosticpro-advantage.html
11. Poplack SP, Tosteson AN, Grove M, Wells WA, Carney PA. The Practice of Mammography in 53,803 Women from the New Hampshire Mammography Network. Radiology. 2000; 217:832–840. [PubMed: 11110951]
12. American College of Radiology, Breast Imaging Reporting and Data System (BI-RADS), Copyright. 2004.
13. Miglioretti DL, Heagerty PJ. Marginal modeling of multi-level binary data with time-varying covariates. Biostatistics. 2004; 5(3):381–983. [PubMed: 15208201]

14. Miglioretti DL, Heagerty PJ. Marginal modeling of non-nested multilevel data using standard software. American Journal of Epidemiology. 2007; 165(4):453–463. [PubMed: 17121864]
15. Louise-Anne McNutt. Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes. Am J Epidemiol. 2003; 157:940–943. [PubMed: 12746247]
16. R Development Core Team. R Foundation for Statistical Computing. Vienna; Austria: 2009. R: A language and environment for statistical computing. ISBN 3-900051-07-0, URL http://www.R-project.org.
17. Højsgaard S, Halekoh U, Yan J. The R Package geepack for Generalized Estimating Equations Journal of Statistical Software. 2005; 15(2):1–11.
18. Straus SE, Tetroe J, Graham I. Defining knowledge translation. CMAJ. 2009; 181:3–4.
19. Smith-Bindman R, Chu PW, Miglioretti DL, Sickles EA, Blanks R, Ballard-Barbash R, Bobo JK, Lee NC, Wallis M, Patnick J, Kerlikowske K. Comparison of screening mammography in the United States and the United Kingdom. JAMA. 2003; 290(16):2129–2137. [PubMed: 14570948]
20. Hofvind S, Geller BM, Vacek PM, Thresen S, Skaane P. Using the European Guidelines to Evaluate the Norwegian Breast Cancer Screening Program. Eur J Epidemiol. 2007; 22(7):447–55. [PubMed: 17594526]
21. Kundel HL, Nodine CF, Conant EF, Weinstein SP. Holistic component of image perception in mammogram interpretation: gaze-tracking study. Radiology. 2007; 242(2):396–402. [PubMed: 17255410]
22. Mello-Thoms C, Dunn S, Nodine CF, Kundel HL, Weinstein SP. The perception of breast cancer: what differentiates missed from reported cancers in mammography? Academic Radiology. 2002; 9(9):1004–1012. [PubMed: 12238541]
23. Mello-Thoms C, Dunn SM, Nodine CF, Kundel HL. An analysis of perceptual errors in reading mammograms using quasi-local spatial frequency spectra. Journal of Digital Imaging. 2001; 14(3): 117–123. [PubMed: 11720333]
24. Castella C, Eckstein MP, Abbey CK, Kinkel K, Verdun FR, Saunders RS, Samei E, Bochud FO. Mass detection on mammograms: influence of signal shape uncertainty on human and model observers. Journal of the Optical Society of America, A, Optics, Image Science, & Vision. Feb; 2009 26(2):425–36.
25. Naughton C, Feely J, Bennett K. A RCT evaluating the effectiveness and cost-effectiveness of academic detailing versus postal prescribing feedback in changing GP antibiotic prescribing. Journal of Evaluation in Clinical Practice. 2009; 15(5):807–812. [PubMed: 19811593]
26. Baum S, Harder S. Appropriate dosing in patients with impaired renal function on medical wards before and after an educational intervention. International Journal of Clinical Pharmacology & Therapeutics. 2010; 48(1):29–35. [PubMed: 20040337]
27. Frich JC, Hoye S, Lindbaek M, Straand J. General practitioners and tutors' experiences with peer group academic detailing: a qualitative study. BMC Family Practice. 2010; 11:12. [PubMed: 20152015]
28. American College of Radiology. Oct. 2010 Personal Communication
29. The Hawthorne Effect And The Overestimation of Treatment Effectiveness. Psychology Today. Nov 27. 2010 http://www.psychologytoday.com/blog/overcoming-pain/201011/the-hawthorne-effect-and-the-overestimation-treatment-effectiveness
30. American College of Radiology Guidelines for Computer Monitor Display Capabilities for Diagnostic Radiology. http://www.acr.org/secondarymainmenucategories/quality_safety/guidelines/med_p hys/electronic_practice.aspx

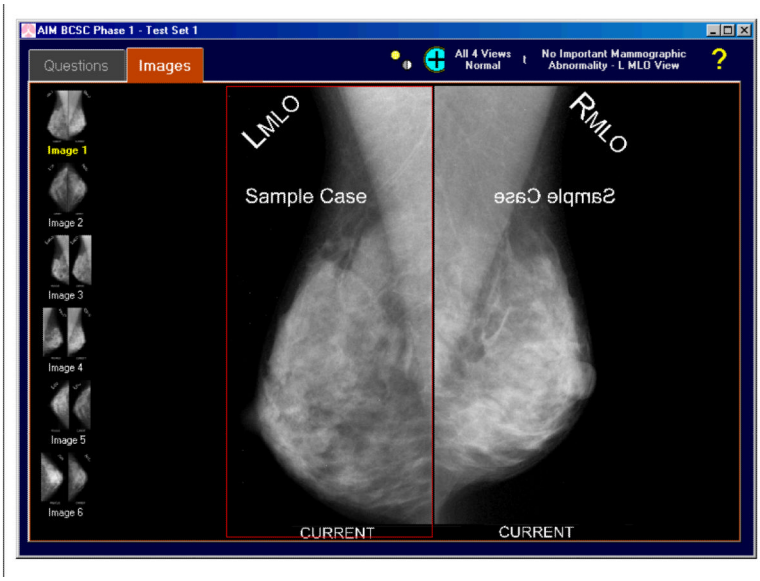

Figure 1.
Example Case of Image Presentation in the Test Set Software

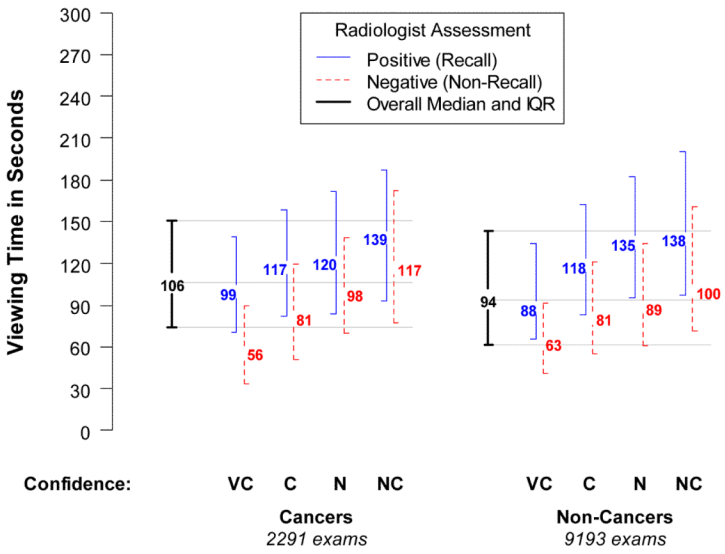

Figure 2.

Median and interquartile range of viewing times by true cancer status and radiologist confidence.

Confidence categories: VC= very confident, C= confident, N = neutral, NC = not confident.

Table 1

Radiologist Characteristics at Baseline by Randomized Test Set Assignment

	Test Set*			
	1 (n=30)	**2 (n=34)**	**3 (n=28)**	**4 (n=27)**
Number of mammography examinations read/week	n (%)	n (%)	n (%)	n (%)
Up to 10	2 (7%)	2 (6%)	2 (7%)	2 (7%)
11-49	8 (27%)	10 (29%)	4 (14%)	3 (11%)
50-99	7 (23%)	10 (29%)	10 (36%)	9 (33%)
100-199	7 (23%)	8 (24%)	9 (32%)	7 (26%)
200 or more	6 (20%)	4 (12%)	3 (11%)	6 (22%)
Category 1 CME hours in mammography received in the past three years (not including this program)				
None	1 (3%)	0 (0%)	3 (11%)	0 (0%)
1-10	2 (7%)	3 (9%)	2 (7%)	0 (0%)
11-15	5 (17%)	6 (18%)	6 (21%)	4 (15%)
16-30	10 (33%)	14 (41%)	8 (29%)	13 (48%)
31 or more	12 (40%)	11 (32%)	9 (32%)	10 (37%)
Specialization in radiology				
Generalist (no specialization)	10 (33%)	6 (18%)	5 (18%)	5 (19%)
Primarily generalist (some specialization)	9 (30%)	15 (44%)	12 (43%)	12 (44%)
Primarily Specialist (some general work)	9 (30%)	11 (32%)	9 (32%)	7 (26%)
Breast Specialist (no general work)	2 (7%)	2 (6%)	2 (7%)	3 (11%)
Current main practice type				
Community practice radiology group	16 (53%)	19 (56%)	15 (54%)	16 (59%)
Academic radiology group	5 (17%)	4 (12%)	4 (14%)	4 (15%)
Radiologist in a multispecialty group	7 (23%)	5 (15%)	6 (21%)	3 (11%)
Solo radiology practice	0 (0%)	1 (3%)	2 (7%)	3 (11%)
Locum tenens	2 (7%)	4 (12%)	1 (4%)	0 (0%)
Other	0 (0%)	1 (3%)	0 (0%)	1 (4%)
Completed post-residency fellowship				
Yes, in Breast or Women's Imaging	3 (10%)	2 (6%)	4 (14%)	6 (22%)
Yes, other	14 (47%)	16 (47%)	13 (46%)	8 (30%)
No	13 (43%)	16 (47%)	11 (39%)	13 (48%)
Number of days per week working in breast imaging in past year				
1 day or less	9 (30%)	12 (35%)	8 (29%)	5 (19%)
2 days	5 (17%)	5 (15%)	6 (21%)	8 (30%)
3 days	7 (23%)	7 (21%)	9 (32%)	5 (19%)
4 days	3 (10%)	6 (18%)	2 (7%)	4 (15%)
5 days	6 (20%)	4 (12%)	3 (11%)	5 (19%)
Years interpreting mammograms				

	Test Set*			
	1 (n=30)	**2 (n=34)**	**3 (n=28)**	**4 (n=27)**
1-5 years	9 (30%)	7 (21%)	6 (21%)	3 (11%)
6-10 years	2 (7%)	6 (18%)	5 (18%)	5 (19%)
11-20 years	13 (43%)	14 (41%)	13 (46%)	11 (41%)
21-30 years	3 (10%)	4 (12%)	4 (14%)	6 (22%)
31 years or more	3 (10%)	3 (9%)	0 (0%)	2 (7%)
Estimated number of mammograms per year interpreted over the last 5 years				
Don't know	10 (33%)	6 (18%)	5 (18%)	8 (30%)
0 to 1000	2 (7%)	8 (24%)	2 (7%)	3 (11%)
1001 to 2000	7 (23%)	10 (29%)	5 (18%)	4 (15%)
2001 to 3000	5 (17%)	2 (6%)	11 (39%)	4 (15%)
More than 3000	6 (20%)	8 (24%)	5 (18%)	8 (30%)
Self-rated ability to perceive & determine importance of mammographic findings				
Not sure	2 (7%)	0 (0%)	0 (0%)	1 (4%)
Below Average	0 (0%)	0 (0%)	1 (4%)	1 (4%)
Average	15 (50%)	20 (59%)	12 (43%)	6 (22%)
Above Average	9 (30%)	12 (35%)	15 (54%)	14 (52%)
Expert	4 (13%)	2 (6%)	0 (0%)	15 (19%)
Currently interprets digital screening exams	3 (10%)	7 (21%)	6 (21%)	7 (26%)
Where did you review this DVD application				
Home computer	11 (37%)	14 (41%)	13 (46%)	12 (44%)
Office computer	3 (10%)	5 (15%)	1 (4%)	4 (15%)
Radiology reading room workstation	2 (7%)	3 (9%)	0 (0%)	1 (4%)
Study Laptop	12 (40%)	12 (35%)	13 (46%)	9 (33%)
Other	2 (7%)	0 (0%)	1 (4%)	1 (4%)

*Participants were block randomized within strata defined by registry/site and whether or not a radiologist had reviewed at least 30 cancers in the BCSC database to balance the number of radiologists with accurate measures of clinical sensitivity of mammography across test sets. Non-BCSC radiologists were considered to be in the group of radiologists who did not review at least 30 cancers.

Table 2

Median Viewing Time and Inter-quartile Range in Seconds Overall and by Radiologist-assessed Breast Density and Expert-assessed Lesion Type

	No Abnormalities*	Any Lesion Type	Mass	Calcification	Asymmetry	Architectural Distortion
All Cases	91 (60, 140) *n = 6,985*	103 (71, 151) *n = 4,499*	90 (64, 131) *n = 900*	111 (77, 164) *n = 1,170*	108 (72, 151) *n = 1,429*	100 (69, 151) *n = 1,000*
By breast density						
Almost Entirely Fat (<25%)	80 (53, 123) *n = 1,615 (23%)*	98 (63, 142) *n = 326 (7%)*	84 (58, 112) *n = 78 (9%)*	104 (80, 128) *n = 49 (4%)*	122 (70, 163) *n = 96 (7%)*	89 (58, 148) *n = 103 (10%)*
Scattered FG Densities (25-50%)	97 (65, 148) *n = 2,569 (37%)*	105 (71, 152) *n = 1,856 (41%)*	94 (64, 134) *n = 503 (56%)*	111 (79, 162) *n = 485 (41%)*	108 (74, 150) *n = 517 (36%)*	101 (71, 161) *n = 351 (35%)*
Heterogeneously Dense (51-75%)	96 (63, 148) *n = 2,127 (30%)*	104 (72, 150) *n = 1,638 (36%)*	93 (66, 133) *n = 266 (30%)*	113 (77, 164) *n = 461 (39%)*	107 (72, 151) *n = 465 (33%)*	102 (73, 146) *n = 446 (45%)*
Extremely Dense (>75%)	86 (58, 134) *n = 673 (10%)*	102 (69, 154) *n = 679 (15%)*	82 (63, 103) *n = 53 (6%)*	110 (74, 177) *n = 175 (15%)*	104 (71, 147) *n = 351 (25%)*	98 (61, 149) *n = 100 (10%)*

FG = fibroglandular Percents shown are column percents

*One of the 6,985 lesion-free exams was not assessed for density

Table 3

Relative risk of recall among cancer films and non-cancer films

	Unadjusted RR	p	Adjusted RR†	p
Among Cancer Films:				
Viewing Time, 1 minute increase	1.10 (1.04 ,1.17)	0.002	1.12 (1.06 ,1.19)	<0.001
Radiologist confidence on current exam		0.003		<0.001
Very Confident	1.30 (1.11 ,1.53)		1.32 (1.16 ,1.50)	
Confident	1.15 (1.04 ,1.27)		1.13 (1.04 ,1.23)	
Neutral		1 (referent)	1 (referent)	
Not Confident‡	0.93 (0.75 ,1.17)		0.94 (0.77 ,1.16)	
Among Non-Cancer Films:				
Viewing Time, 1 minute increase	1.42 (1.32 ,1.53)	<0.001	–	
Radiologist confidence on current exam		<0.001		
Very Confident	0.55 (0.36 ,0.86)		–	
Confident	0.68 (0.56 ,0.82)		–	
Neutral		1 (referent)	–	
Not Confident	1.41 (1.17 ,1.69)		–	
Confidence-Specific Viewtime Effects				
Viewing Time, 1 minute increase		0.016*		0.039*
on Very Confident exams	1.50 (1.30 ,1.74)		1.42 (1.21 ,1.68)	
on Confident exams	1.41 (1.29 ,1.54)		1.40 (1.29 ,1.52)	
on Neutral exams	1.38 (1.30 ,1.48)		1.38 (1.29 ,1.48)	
on Not Confident exams	1.19 (1.08 ,1.31)		1.20 (1.09 ,1.31)	

† Models are adjusted for expert-identified lesion type, reader-rated breast density, fellowship category, specialization, years interpreting mammograms, mammograms read per week, and random test set assignment. No significant interaction was found between viewing time and confidence among cancer films (p=0.73). Adjusted model for non-cancer films contains an interaction (p=0.039) between viewing time and exam-level confidence, and so confidence-specific viewtime estimates are presented here.

‡ Radiologists reporting ‘Not very confident’ or ‘Not at all confident’ are combined in the ‘Not confident’ group

* P value from test for interaction between confidence and viewing time