Published in final edited form as:
AJR Am J Roentgenol. 2014 June ; 202(6): W586–W596. doi:10.2214/AJR.13.11147.

Educational interventions to improve screening mammography interpretation: a randomized, controlled trial

Geller BM, EdD, **Bogart A, MS**, **Carney PA, PhD**, **Sickles EA, MD**, **Smith RA, PhD**, **Monsees B, MD**, **Bassett LW, MD**, **Buist DM, PhD**, **Kerlikowske K, MD**, **Onega T, PhD**, **Yankaskas B, PhD**, **Haneuse S, PhD**, **Hill DA, PhD**, **Wallis M, MD**, and **Miglioretti DL, PhD**

Abstract

Purpose—Conduct a randomized controlled trial of educational interventions to improve performance of screening mammography interpretation.

Materials and Methods—We randomly assigned physicians who interpret mammography to one of three groups: (1) self-paced DVD; (2) live, expert-led educational session; or (3) control. The DVD and live interventions used mammography cases of varying difficulty and associated teaching points. Interpretive performance was compared using a pre-/post-test design. Sensitivity, specificity, and positive predictive value (PPV) were calculated relative to two outcomes: cancer status and consensus of three experts about recall, and each were compared using logistic regression adjusting for pre-test performance.

Results—102 radiologists completed all aspects of the trial. After adjustment for pre-intervention performance, the odds of improved sensitivity for correctly identifying a lesion relative to expert recall were 1.34 times higher for DVD participants than controls (95% confidence interval [CI]: 1.00, 1.81; P=0.050). The odds of improved PPV for correctly identifying a lesion relative to both expert recall (odds ratio [OR]=1.94, 95% CI: 1.24, 3.05; P=0.004) and cancer status (OR=1.81, 95% CI: 1.01, 3.23; P=0.045) were significantly improved for DVD participants compared to controls with no significant change in specificity. For the live-intervention group, specificity was significantly lower than the control group (OR relative to expert recall=0.80; 95% CI: 0.64, 1.00; P=0.048; OR relative to cancer=0.79; 95% CI: 0.65, 0.95; P=0.015).

Conclusion—In this randomized controlled trial, the DVD educational intervention resulted in a significant improvement in mammography interpretive screening performance on a test-set, which could translate into improved clinical interpretative performance.

Introduction

In the United States, interpretive performance of screening mammography shows significant variability (1,2). While numerous studies have sought to understand the sources of variability, such as interpretive volume, malpractice concerns, and other radiologist

Corresponding author: Berta Geller, EdD, University of Vermont, Health Promotion Research, 1 South Prospect St., Burlington, VT 05401, berta.geller@uvm.edu, Phone: 802-656-4187, Fax: 802-656-9926.

This research has not been presented nor has been accepted for a future meeting at RSNA.

characteristics (2-9), only a few have focused on practical approaches to improve interpretive performance (10-16). In two studies with multi-component interventions to enhance radiologists' interpretive performance, the components that contributed most to the observed effects could not be determined (10,14).

The Mammography Quality Standards Act (MQSA) requires physicians who interpret mammography to obtain at least 15 hours of category 1 Continuing Medical Education (CME) units in mammography every 36 months to maintain their qualifications (1). Also under MQSA, a lead interpreting physician is required to review mammography audit data and provide feedback on performance. However, the persistence of significant variability in interpretive performance has led to attempts to go beyond CME to improve interpretive skills. An IOM report suggested modifying the MQSA CME requirements to specifically focus on skills assessment using test sets with a mix of normal and abnormal findings. This assessment method would compare the performance of interpreting physicians with their peers to provide them with feedback on their interpretive skills and identify areas for improvement (17).

We conducted a randomized controlled trial of two educational interventions designed to improve performance of screening mammography interpretation. We tested the hypotheses that: (1) test sets of mammographic images developed with teaching points would increase physicians' ability to identify screening mammography findings that required recall for further work-up, (2) that mode of delivering these interventions (self-paced DVD versus live instructional program) would have different effects, and (3) that this increased ability would result in improved interpretive performance.

Materials and Methods

We conducted a three-arm, block-randomized, controlled trial to evaluate two interventions to improve screening mammography interpretative performance. A pre-/post-intervention test set design was used to evaluate the impact of a self-paced DVD and a live expert-led seminar compared to a control group. Both the live intervention and the self-paced DVD included 40 cases, 18 of which were biopsy confirmed cancers.

Study population

In 2009–2010, 300 radiologists who interpreted mammograms at a facility contributing to a Breast Cancer Surveillance Consortium (BCSC) registry between 2005 and 2006 were invited to participate. We also invited 103 non-BCSC radiologists from Oregon, Washington, North Carolina, San Francisco, and New Mexico. Institutional Review Board approval was obtained by each registry and the Statistical Coordinating Center (SCC) that performed the analysis for all study activities, including active consent to enroll radiologists and perform analytic studies. All registries and the SCC followed Health Insurance Portability and Accountability Act-compliant procedures to obtain mammography films and patient information, and received a Federal Certificate of Confidentiality and other protections for the study participants and the identities of women, physicians, and facilities related to the films we used (18).

Randomization

We randomized the 119 radiologists who completed a pre-intervention test set into one of the three groups (Figure 1). Because we only offered the live intervention once on the west coast and once on the east coast, it was not possible for some radiologists assigned to the live intervention to attend. If a radiologist provided a legitimate reason for not attending the live seminar, we re-randomized him or her to either the DVD or control group, making these groups larger than the live intervention group. Additionally, one individual who was originally randomized to the DVD group was inadvertently invited to join one of the live seminars, and accepted, so we retained this individual in the live intervention group.

Pre-intervention and post-intervention test sets

The test sets used to evaluate performance are described elsewhere (6). Briefly, a random sample of original screening mammography films from registries of the National Cancer Institute-funded BCSC (19) were digitized by the American College of Radiology (ACR) to create four pre-intervention test sets and one post-intervention test set. Three expert mammographers (BM,,EAS, LB) with extensive breast imaging experience of 32-36 years reviewed all digitized cases to determine which should be recalled for further work-up. The pre-intervention test sets consisted of four test sets, each included 109 screening mammograms, that were created as part of a larger study designed to test whether disease prevalence or case difficulty in a test set changes the association between test set performance and clinical performance (20). Our process for developing the test sets, interventions and feedback was very similar to the PERFORMS project in the United Kingdom (15,16). Before performance evaluation with the pre-intervention test set, participants took a brief survey about their training and experience. Radiologists from all three groups (DVD, Live Seminar and control) who completed one or more component (pre-intervention/post-intervention test set, an intervention) received an individualized audit feedback report after each component. These reports provided tabular and graphic results of their performance compared with the experts and with the aggregated performance of participants who interpreted the same test set or the same intervention.

Computer Requirements

Minimum mandatory computer requirements included:

- **Processor:** Pentium 4 with 1GHz or higher preferred (also newer processors such as Intel Core Duo or AMD×2)
- **RAM:** 1GB
- **Display:** Minimum 17″ (larger is better) capable of displaying 1280×1024 or higher in 32-bit color. A laptop needed to have a 15.4″ or larger screen capable of displaying 1440 ×900 or higher
- **DVD-ROM drive**
- **Hard Drive:** At least 40GB
- **Internet access** from the computer on which the test DVD was read..

We loaned laptops to radiologists who did not have computers that met these requirements.

Intervention Development

The teaching set was generated based on pre-intervention test set performance. We determined the types of mammographic findings that were commonly missed and misinterpreted and chose 40 cases, none of which appeared on the subsequent post-intervention test set, with 18 cancers visible on the mammograms, 6 non-cancers that the experts recalled for sufficiently high probability of cancer, and 16 non-cancers not recalled by the experts. Among the 18 cancers, the expert radiologist panel categorized films by their most serious findings, identifying 3 with masses, 4 with calcifications, 7 with asymmetry, and 4 with architectural distortion. Experts considered 8 of the cancers to be of intermediate difficulty and 6 as subtle. Four cases, borrowed from the Mammography Interpretive Skills Assessment (MISA) test set (21,22), were not assessed by experts for difficulty. Fifteen of the 18 cancers were interpreted clinically as true positives, and three were false negatives. Each case provided current craniocaudal (CC) and mediolateral-oblique (MLO) views and CC and MLO views from one prior comparison examination. Our expert panel wrote a teaching point for each case. All teaching points had a final review by a single expert radiologist.

Two cases in the pre-intervention test sets had more than one lesion within a breast, but cases were reviewed and experts agreed on which lesion was the most significant. Instructions provided to participants directed them to choose the most significant lesion. If there was a mass with adjacent microcalcifications, for example, the region of interest was wide and clicking within any of the area would give the radiologist credit for the finding. The test-takers could not mark multiple locations.

Self-paced DVD—A DVD was developed using the teaching set that allowed radiologists to learn the intervention content at their convenience. The DVD used the same 40 cases as the live session and 21 additional cases, including one additional MISA case, for radiologists who wanted more practice. Participants checked a box for either a positive assessment (Breast Imaging Reporting and Data System [BI-RADS] categories 0, 4, 5) or a negative assessment (BI-RADS 1, 2). For positive assessments, participants marked the location of the most important finding. After submitting an answer, the participant was immediately told if the response was correct. Next, the examination was shown again, with any important findings marked. A teaching point was displayed that explained the reason to recall or not recall the case, with explanations of any findings or the information that the case had no significant finding. Additional imaging was included with the teaching point when available (see Figure 2). Our software was designed by the ACR based on their MISA program with modifications as specified (21).

Live expert-led seminar—The live seminar was a one-day, eight-hour course taught by two experts and held in San Francisco, CA (LB, EAS) and Hanover, NH (BM, EAS). Participants reviewed teaching set cases on individual workstations in a classroom of networked computers, answering the same questions as on the DVD. When the entire group was finished with a case, data were aggregated and shared with the entire class by projecting

de-identified results on an overhead screen. Participants could compare their interpretations on their own computers to the group.

The expert instructor reviewed the correct interpretation and described the teaching points, commented on participant interpretations, and responded to questions. For example, if a large proportion recalled a benign finding (false-positive), the instructor discussed why experts would or would not recall the case.

Two months after all participants had completed the live interventions or the self-paced DVD, each participant was given audit feedback showing individual performance compared with the aggregated performance of participants in the same intervention group, relative to both expert recall and cancer status.

Among participants assigned to the Live or DVD interventions, the elapsed time between completing the intervention and receiving the follow-up test set (110 screening mammograms) ranged from 92 days to 258 days (approximately 3 to 9 months), with one exception: one subject assigned to the DVD group took an unusually long time to complete the DVD (140 days, the maximum we observed), and so once completed, had only a 28 day interval before receiving the follow-up test set in the mail. No participants received the follow-up test set prior to completing their assigned intervention.

Control group—Radiologists who were randomly assigned to the control (delayed intervention) group received audit feedback as described above on their performance on the pre-intervention and post-intervention test sets. Control subjects were sent the self-paced DVD after study completion and earned additional CME if they completed the self-paced DVD.

CME Credits—All participants received up to 24 category 1 CME credits through the University of Vermont for completing the three components: pre-intervention test set, intervention (or delayed intervention), and post-intervention test set.

Statistical Analysis

We calculated the frequency of responses to the pre-intervention survey within each randomized group. For each participant, we calculated unadjusted performance measures on the pre-intervention and post-intervention test sets, and the corresponding pre-test to post-test change. We generated kernel density smoothed distributional plots of changes in performance measures by group (23).

Performance measures of sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated relative to two outcomes: expert recall and cancer status. *Cancer cases* were defined as a diagnosis of ductal carcinoma *in situ* or invasive cancer within 12 months of the screening examination. Non-cancer cases were defined as cancer free at least 24 months following the screen. *Expert recall* included true positives and false-positive recalls that the experts deemed necessary to rule out cancer (6).

Sensitivity and PPV were calculated at the lesion level. We defined lesion-level sensitivity as the proportion of expert recalls or cancer cases that were recalled in the correct breast and with correct identification of the lesion location (i.e., the participant clicked a location on the correct breast within the lesion region defined by the experts). Specificity was defined as the proportion of non-cancer cases (or expert non-recalls) not recalled by the participant.

To account for different prevalence of expert recall and cancer among the four pre-intervention test sets, we calculated recalibrated PPV and NPV values based on each participant's observed sensitivity and specificity and a common prevalence value across all test sets, using formulas that express PPV and NPV as functions of sensitivity, specificity, and prevalence (24). Prevalence was set to values that matched the post-intervention test set: 29/110 cases for expert recall and 15/110 for analyses relative to cancer.

To compare the effect of intervention assignment on change in pre-intervention and post-intervention performance, we fit hierarchical logistic regression models. Post-test improvement was evaluated using an adjusted odds ratio (OR) comparing post-intervention to pre-intervention performance in the two intervention arms relative to the control group. Models included normally distributed random effects for radiologists and cases to accommodate correlation among examinations interpreted by the same radiologist, and among cases interpreted by different radiologists. Models were adjusted for pre-intervention test set number, and PPV and NPV models were additionally adjusted for the prevalence of expert recall or cancer in that test set.

Tests of statistical significance were two-sided with an alpha level of 0.05. All analyses were conducted using R version 2.15.0 (25,26) and SAS 9.2 software (SAS Institute Inc., Cary, NC, USA).

Results

Of 119 subjects who participated in the intervention and control groups, 102 (85.7%) completed the post-intervention test set (Figure 1). Among those who completed the all the components approximately 38% had read mammograms for 10 years or fewer, and 86% described their ability to perceive and determine the importance of mammographic findings as average or above average (Table 1). More than 80% of participants read digital screening mammograms in clinical practice. Twenty-three radiologists assigned to the DVD intervention also completed at least one extra case while 12 completed all 21 extra cases (data not shown).

Lesion-level sensitivity relative to expert recall increased by an average of more than 10% in the live- and DVD-intervention groups and 7% in the control group (Table 2). Average specificity decreased relative to both expert recall and cancer in the live-intervention group, while average specificity increased in both the DVD and control groups. Average increases in recalibrated PPV were higher in the DVD and control groups than in the live-intervention group for both outcomes. Recalibrated NPV increased on average for the live-intervention and control groups, but showed a small average decrease in the DVD group relative to both expert recall and cancer.

Improvements in sensitivity were evident for all three groups relative to expert recall (Figure 3 panel a), but not relative to cancer status (panel c). No obvious differences appeared in the distribution of the changes in specificity relative to expert recall or cancer status (panels b and d).

Sensitivity relative to expert recall improved significantly after the DVD intervention according to the adjusted models (Table 3). The odds of recalling an expert-recalled case at the lesion level increased 1.34 times more in the DVD group (95% confidence interval [CI]: 1.00, 1.81) compared to the corresponding increase in the control group (P=0.05), after adjustment for performance on the pre-intervention test set. Improvement in sensitivity in the live-intervention group did not differ from the control group relative to expert recall or cancer status. Specificity in the live-intervention group was significantly lower than in the control group (OR relative to expert recall = 0.80; 95% CI: 0.64, 1.00; P=0.048; OR relative to cancer status 0.79; 95% CI: 0.65, 0.95; P=0.015), after adjusting for baseline performance. The odds of expert recall given recall by participants (PPV) in the DVD group nearly doubled (OR=1.94; 95% CI: 1.24, 3.05) compared to the control group l(P=0.004), and improved by a factor of 1.81 (95% CI: 1.01, 3.23) relative to cancer status (P=0.045). No significant differences were observed in the changes in recalibrated NPV across the groups.

Finally, we determined the proportion of participants who improved by any amount for sensitivity, specificity, recalibrated PPV, and recalibrated NPV, along with adjusted improvement ORs (Table 4). After adjustment for pre-intervention test set assignment, there were no significant differences in the odds of improving between the live- or DVD-intervention groups and the control group. Results were, however, consistent with the live-intervention participants improving more frequently in sensitivity and recalibrated NPV relative to both outcomes, and with the control group improving most in re-calibrated PPV for both outcomes.

Discussion

We tested the effects of two case-based educational interventions designed to improve interpretive screening mammography performance. After adjustment for pre-intervention test set performance, the improvement in sensitivity relative to expert recall, and PPV relative to both expert recall and cancer were significantly higher for the DVD-intervention participants than for control-group participants, with no significant change in specificity. Thus, our study showed a positive effect on interpretive performance.

The interpretive performance of mammography varies considerably among radiologists. Although differences in patient populations could be partially responsible, most variability probably is related to interpreting physician skills. Roughly 62% of all U.S. radiologists interpret mammography as part of their workload, but only 10.5% consider themselves breast imaging specialists (27). In fact, breast imaging specialists interpret only 30% of mammograms in the U.S. (27). This underscores the importance of methods that improve the skills of general radiologists who interpret screening mammograms.

Our report is only the second published study to use a randomized controlled design for interventions to improve mammography interpretive performance (28). Although only 102 radiologists completed our study, it is the largest to date designed to measure improved mammography performance. Carney et al.(28) tested a 1.5-hour Internet-based CME designed to decrease recall rate. Participation was very low and no differences were found between the control and intervention group as measured by pre- and post-clinical recall rate (28).

Our findings differ from several other studies that were designed to improve mammography interpretation but used study designs that were less rigorous than a randomized controlled trial. Linver and colleagues (12) found that attendance at an extensive three- to four-day instructor-led educational program in breast imaging resulted in significant improvement in sensitivity but no change in PPV. Their intensive intervention used a pre-intervention and post-intervention clinical audit to measure change but included only 12 radiologists from a single practice with no control group. In another study conducted by Berg et al. (11), a one-day CME course increased cancer detection among 21 radiologists but did not decrease the false positive rate. Berg et al. used only a single test set, which contained all cases with abnormal findings. Testing was administered immediately before and after the teaching intervention, and again at 2 to 3 months for a small subset of participants. The Linver study was published in 1992, when mammography CME courses were new, so unlike our study and Berg et al., participants had little or no exposure to previous mammography CME (12). Advantages of our study were the use of different test sets for evaluating pre-intervention and post-intervention performance, with interpretation 3–6 months before and after the intervention. Unlike many previous studies, we included normal mammograms to more closely resemble clinical practice.

Three studies used audit feedback to improve mammography interpretation (10,14-16,29) and found that feedback accompanied with extensive clinical review improves mammography interpretation. However, none of the audit feedback interventions were rigorously tested and in general, this method is very resource intensive.

The strengths of our study include that we tested two interventions, one that used a convenient DVD that allowed participants to proceed at their own pace, and another that used a live, expert-led intervention in which radiologists had the opportunity for more personalized interaction. We found that the DVD, which is a relatively inexpensive, easy-to-disseminate method for delivering CME, was effective in improving mammography interpretation. Another strength of our study is that our participants were a large, diverse group of community radiologists from across the country. In addition, the cases used in our test sets and intervention were randomly selected from clinical practice to represent the diversity of real-world practice.

We focused our results on the lesion level, because detection of the lesion itself is most clinically relevant for improvements in mammography performance. When a woman is recalled for an abnormal finding on mammography, the additional imaging (e.g., diagnostic magnification views and/or ultrasound) focuses first on the region of interest (although the extent of disease is addressed during the diagnostic evaluation). Therefore, it is important to

detect and identify the correct lesion location. The size of the areas we accepted for correct recall was generous, but specific. Improving or changing the exact location of the radiologist's click would not change the results. If the radiologist chose a finding within the broad area, they were given credit if appropriately recalled.

Although our sample was large and diverse, we had based our power calculations on 160 radiologists participating (Supplemental table 1: power analyses); therefore, we lacked statistical power for detecting potentially important differences in performance. It is generally accepted that power calculations should not be redone after the study is completed (29). Instead, our focus was on reporting effect sizes with measures of precision (e.g., 95% confidence intervals). Despite having less statistical power than originally planned, we were able to show statistically significant differences, because the effect sizes from the interventions were higher than we anticipated in our initial sample size calculations. Had we had a larger sample size, we may have detected statistically significant improvements in the live intervention group as well. The re-randomization process may have led to a bias in the live intervention group, because busier radiologists may have been more likely to have legitimate reasons for being unable to participate. However, the DVD vs. control group comparison should not be biased, because radiologists who were unable to participate in the live intervention were randomly reassigned to one of these two groups. Our analyses used an intervention exposure approach instead of an intent-to-treat (ITT) approach. As a sensitivity analysis we conducted the ITT analysis and found the results to be in a similar direction (Supplemental Table 2).

More radiologists in the DVD and control groups did not complete all components of the study compared with the live intervention. It is very challenging to recruit and retain busy physicians to fully complete studies, particularly a study like this that required up to 24 hours over a two year time period. Radiologists who made a commitment to come to our live intervention had to get release time for the specific day and in some instances advanced airplane reservations. We can only surmise that the draw of an intervention taught by recognized experts in the field plus having freed up time in their schedule made it more likely that the radiologists would attend the live intervention. Those in the control and the DVD groups did not necessarily get release time to complete the study and therefore had to fit it into a busy schedule making it more difficult to complete.

We used digitized analogue films, as does the Mammography Case Review educational sets of the American College of Radiology (ACR) (22) and PERFORMS(15), rather than digital images. Thus, the film quality was inferior to digital images and original film mammograms. The available data when the study began showed that only about 25% of BCSC radiologists were interpreting digital mammograms. This information influenced our decision to use digitized films, although by the time the interventions were implemented more than 80% of participants were reading digital mammograms in screening practice. For future interventions, we recommend digitally acquired images since they are readily accessible and most closely approximate current imaging quality and conditions.

Improvement in performance was seen even in the control group. There are at least three potential explanations. First, all groups, including the control group, received an outcome

audit feedback of the results of the pre-intervention test set (described on page 10). Second, the control group gained experience with reading the digitized films and operating the software from the pre-intervention test set. Third, there were four pre-intervention tests of varying difficulty and only one post-intervention test set of average difficulty such that some of the controls may have improved if they took the difficult pre-intervention test set. For these reasons, we adjusted for test set and case-difficulty in the statistical analysis and we compared changes in the intervention groups relative to the control group, which would remove these factors as driving our results.

In conclusion, continuing to develop and test methods to effectively improve interpretive performance is important because over 130 million women are screened with mammography in the U.S. every 1–2 years (3) and breast cancer remains a leading cause of cancer death. Results from our randomized controlled trial suggest that interpretive performance can be improved by educational interventions based on actual clinical cases with findings that are frequently misinterpreted. These findings and the improvements in interpretive performance observed by previous investigators suggest that direct interventions focused on improving the accuracy of mammography interpretation should be the highest priority for meeting CME requirements under MQSA. We plan to extend our results by measuring whether clinical recall rates changed after the interventions. We will need to wait several years to calculate clinical sensitivity and specificity using true cancer status because of a lag time in data from cancer registries. CME is increasingly being offered through the Internet, providing easy access to educational opportunities and expanding the dissemination possibilities for CME interventions such as those described in this study (30).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

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 (R01CA107623; K05CA104699; Breast Cancer Surveillance Consortium: U01CA63740, U01CA86076, U01CA86082, U01CA70013, U01CA69976, U01CA63731, U01CA70040, HHSN261201100031C). The collection of cancer and vital status 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. 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.

References

1. Mammography quality standards act of 1992. Mammography facilities requirement for accrediting bodies, and quality standards and certifying requirements: interim rules (21 CFR 900). Vol. 58. Government Printing Office; Washington, DC: Dec 21. 1993 p. 57558-72.Federal REgister1992 1992. Report No.: 102-539
2. Elmore J, Jackson S, Abraham L, Miglioretti D, Carney P, Geller B, et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology. 2009; 253(3):641–51. [PubMed: 19864507]

3. Buist D, Anderson M, Haneuse S, Sickles E, Smith R, Carney P, et al. Influence of Annual Interpretive Volume on Screening Mammography Performance in the United States. Radiology. 2011; 259(1):72–84. [PubMed: 21343539]
4. Barlow WE, Chi C, Carney PA, Taplin SH, D'Orsi C, Cutter G, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst. Dec 15; 2004 96(24):1840–50. [PubMed: 15601640]
5. Smith-Bindman R, Chu P, Miglioretti DL, Quale C, Rosenberg RD, Cutter G, et al. Physician predictors of mammographic accuracy. J Natl Cancer Inst. Mar 2; 2005 97(5):358–67. [PubMed: 15741572]
6. Carney P, Geller B, Bogart A, Kerlikowske K, Rosenberg R, Buist D, et al. Association between time spent, confidence and accuracy of screening mammography. AJR Am J Roentgenol. 2012; 198(4):970–8. [PubMed: 22451568]
7. Geller BM, Bowles EJ, Sohng HY, Brenner RJ, Miglioretti DL, Carney PA, et al. Radiologists' performance and their enjoyment of interpreting screening mammograms. AJR Am J Roentgenol. Feb; 2009 192(2):361–9. [PubMed: 19155395]
8. Elmore J, Taplin S, Barlow W, Cutter G, D'Orsi C, Hendrick R, et al. Does litigation influence medical practice? The influence of community radiologists' medical malpractice perceptions and experience on screening mammography. Radiology. 2005; 236(1):37–46. [PubMed: 15987961]
9. Miglioretti DL, Gard CC, Carney PA, Onega TL, Buist DS, Sickles EA, et al. When radiologists perform best: the learning curve in screening mammogram interpretation. Radiology. Dec; 2009 253(3):632–40. [PubMed: 19789234]
10. Adcock KA. Initiative to Improve Mammogram Interpretation. Permanente Journal. 2004; 8:12–8.
11. Berg WA, D'Orsi CJ, Jackson VP, Bassett LW, Beam CA, Lewis RS, et al. Does training in the Breast Imaging Reporting and Data System (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography? Radiology. Sep; 2002 224(3):871–80. [PubMed: 12202727]
12. Linver M, Paster S, Rosenberg R, Key C, Stidley C, King W. Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,633 cases. Radiology. 1992; 184:39–43. [PubMed: 1609100]
13. Urban N, Longton G, Crowe A, Drucker M, Lehman C, Peacock S, et al. Computer-assisted mammography feedback program (CAMFP) an electronic tool for continuing medical education. Acad Radiol. 2007; 14(9):1036–42. [PubMed: 17707310]
14. Perry N. Interpretive Skills in the National Health Service Breast Screening Programme: performance indicators and remedial measures. Seminars in Breast Disease. 2003; 6(3):08–113.
15. Gale A. PERFORMS: A self-assessment scheme for radiologists in breast screening. Seminars in Breast Disease: Improving and monitoring mammographic interpretative skills. 2003; 6:148–52.
16. Scott H, Gale A. Breast screening: PERFORMS identifies key mammographic training needs. Br J Radiology. 2006; 79:S127–S33.
17. Nass, S.; Ball, J. Improving Breast Imaging Quality Standards. Institute of Medicine; Washington DC: 2005.
18. Carney PA, Geller BM, Moffett H, Ganger M, Sewell M, Barlow WE, et al. Current Medicolegal and Confidentiality Issues in Large, Multicenter Research Programs. Am J Epidemiol. Aug 15; 2000 152(4):371–8. 2000. [PubMed: 10968382]
19. Ballard-Barbash R, Taplin SH, Yankaskas BC, Ernster VL, Rosenberg RD, Carney PA, et al. Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database. AJR Am J Roentgenol. Oct; 1997 169(4):1001–8. [PubMed: 9308451]
20. Onega T, Anderson M, Miglioretti D, Buist D, Geller B, Bogart A, et al. Establishing a gold standard for test sets: variation in interpretive agreement of expert mammographers. Academic Radiology. 2013; 20(6):731–9. [PubMed: 23664400]
21. Radiology, ACo. Mammography Case Review. 2012. Available from: http://www.acr.org/mcr
22. Sickles E. The American College of Radiology's Mammography Interpretive Skills Assessment (MISA) Examination. Seminars in Breast Disease. 2003; 6(3):133–9.
23. Sheather S, MC J. A reliable data-based bandwidth selection method for kernel density estimation. J Roy Statist Soc B. 1991; 53(3):683–90.

24. van Belle, G.; Fisher, L.; Heagerty, P.; Lumley, T. Biostatistics: A Methodology for the Health Sciences. Wiley-Blackwell; Hoboken, New Jersey: 2004.
25. R RDCT. A language and environment for statistical computing Vienna. R Foundation for Statistical Computing; Austria: 2011.
26. R Development Core Team. R: A language and environment for statistical computing. 2012. Available from: http://www.R-project.org/
27. Lewis R, Sunshine J, Bhargavan M. A portrait of breast imaging specialists and of the interpretation of mammography in the United States. AJR. 2006; 187:W456–68. [PubMed: 17056875]
28. Carney P, Abraham A, Cook A, Feig S, Sickles E, Miglioretti D, et al. Impact of an educational intervention designed to reduce unnecessary recall during screening mammography. Acad Radiology. In press.
29. van der Horst F, Hendriks J, Rijken H, Holland R. Breast cancer screening in the Netherlands: Audit and training of radiologists. Seminars in Breast Disease. 2003; 6(3):114–2.
30. Casebeer L, Brown J, Roepke N, Grimes C, Henson B, Palmore R, et al. Evidenced-based choices of physicians: a comparative analysis of physicians participating in Internet CME and non-participants. BMC Medical Education. 2010; 10(42)http://www.biomedcentral.com/1472-6920/10/42

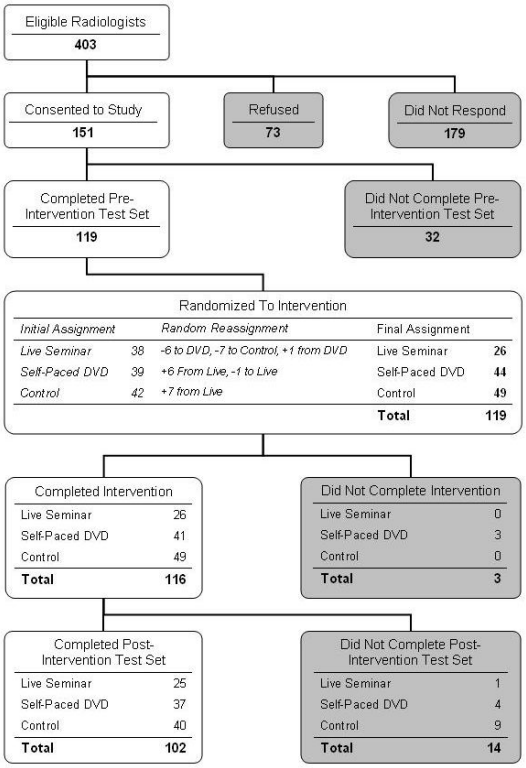
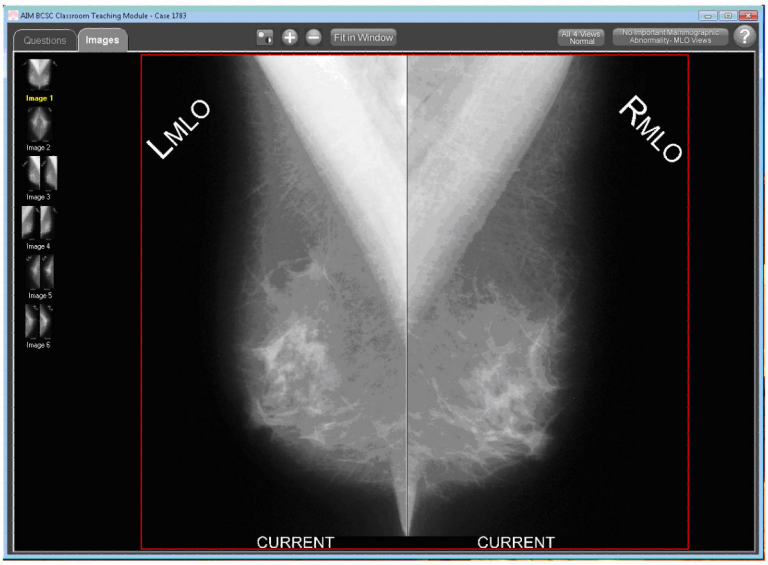
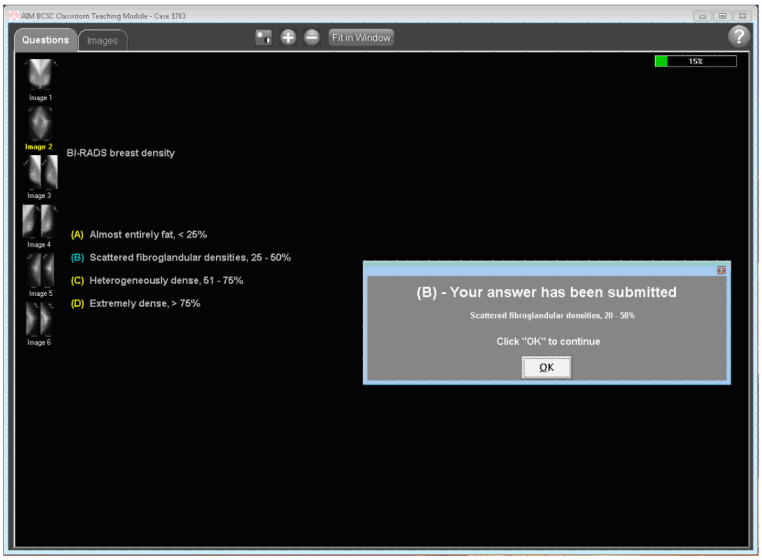


Figure 1. Flow chart of study participants





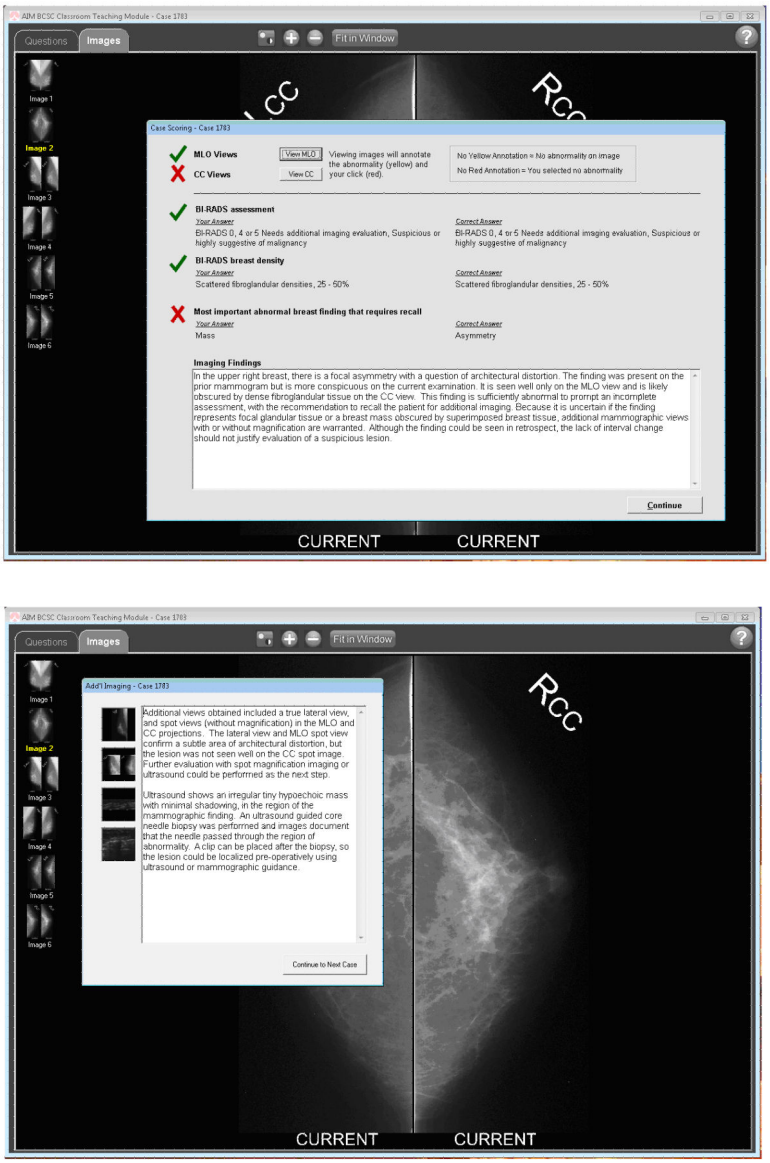


Figure 2. Example test set case with teaching points

(a) Access to multiple views of cases; (b) question about breast density; (c) Case score and teaching point; (d) Additional imaging and teaching point.

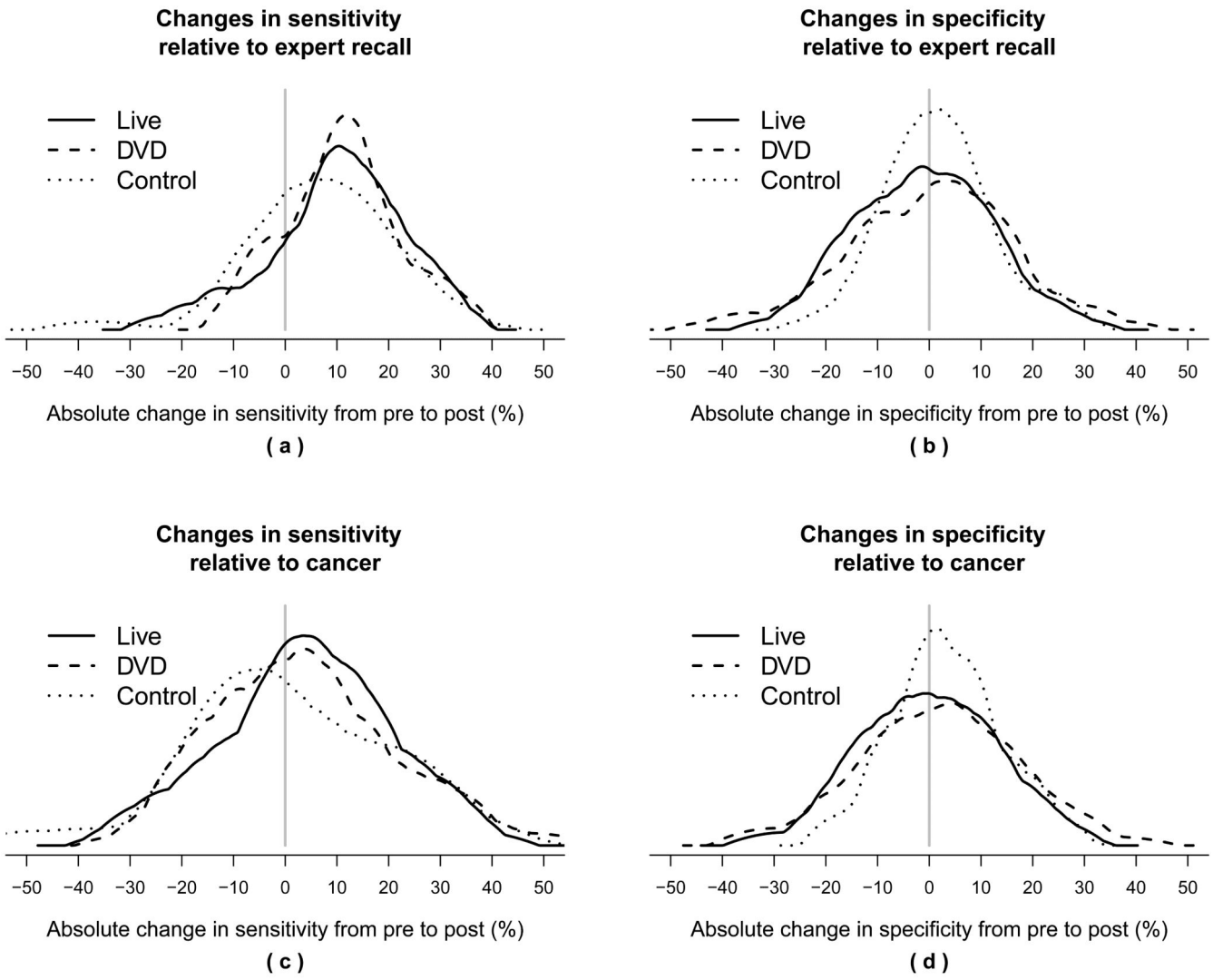


Figure 3. Graphs of test set performance measures pre- and post-intervention by intervention group and relative to expert recall and cancer status

Increases in sensitivity or specificity are indicated by a shift to the right. Solid line, live intervention; dashed line, DVD intervention; dotted line, control.

Table 1

Radiologist characteristics overall and by random intervention assignment

Physician and Practice Characteristics	All Participants *N*=101	Live Seminar *N*=25	Self-Paced DVD *N*=36	Control Group *N*=40
Test Set, n (%)				
1 (15 cancers, lower difficulty)	25 (24.8%)	5 (20.0%)	10 (27.8%)	10 (25%)
2 (15 cancers, higher difficulty)	30 (29.7%)	7 (28.0%)	10 (27.8%)	13 (32.5%)
3 (30 cancers, lower difficulty)	25 (24.8%)	7 (28.0%)	10 (27.8%)	8 (20%)
4 (30 cancers, higher difficulty)	21 (20.8%)	6 (24.0%)	6 (16.7%)	9 (22.5%)
Years interpreting mammograms, n (%)				
1-5 years	22 (21.8%)	6 (24.0%)	9 (25.0%)	7 (17.5%)
6-10 years	17 (16.8%)	6 (24.0%)	7 (19.4%)	4 (10.0%)
11-20 years	44 (43.6%)	8 (32.0%)	18 (50.0%)	18 (45.0%)
21-30 years	14 (13.9%)	4 (16.0%)	0 (0.0%)	10 (25.0%)
31 years or more	4 (4.0%)	1 (4.0%)	2 (5.6%)	1 (2.5%)
Self-rated ability to perceive & determine importance of mammographic findings, n (%)				
Not sure	2 (2.0%)	0 (0.0%)	1 (2.8%)	1 (2.5%)
Below Average	2 (2.0%)	1 (4.0%)	1 (2.8%)	0 (0.0%)
Average	47 (46.5%)	11 (44.0%)	17 (47.2%)	19 (47.5%)
Above Average	40 (39.6%)	11 (44.0%)	14 (38.9%)	15 (37.5%)
Expert	10 (9.9%)	2 (8.0%)	3 (8.3%)	5 (12.5%)
Number of mammography examinations read per week, n (%)				
Up to 10	6 (5.9%)	2 (8.0%)	2 (5.6%)	2 (5.0%)
11-49	20 (19.8%)	9 (36.0%)	5 (13.9%)	6 (15.0%)
50-99	34 (33.7%)	6 (24.0%)	14 (38.9%)	14 (35.0%)
100-199	26 (25.7%)	5 (20.0%)	11 (30.6%)	10 (25.0%)
200 or more	15 (14.9%)	3 (12.0%)	4 (11.1%)	8 (20.0%)
Category 1 CME hours in mammography received in the past three years (not including this program), n (%)				
none	4 (4.0%)	1 (4.0%)	3 (8.3%)	0 (0.0%)
1-10	7 (6.9%)	2 (8.0%)	2 (5.6%)	3 (7.5%)
11-15	18 (17.8%)	4 (16.0%)	6 (16.7%)	8 (20.0%)
16-30	34 (33.7%)	10 (40.0%)	11 (30.6%)	13 (32.5%)
31 or more	38 (37.6%)	8 (32.0%)	14 (38.9%)	16 (40%)
Specialization in radiology, n (%)				
Generalist (with no specialization)	21 (20.8%)	7 (28.0%)	5 (13.9%)	9 (22.5%)
Primarily generalist (with some specialization)	45 (44.6%)	11 (44.0%)	17 (47.2%)	17 (42.5%)
Primarily Specialist (with some general work)	27 (26.7%)	6 (24.0%)	10 (27.8%)	11 (27.5%)
Breast Specialist (no general work)	8 (7.9%)	1 (4.0%)	4 (11.1%)	3 (7.5%)
Main practice type at time of survey, n (%)				
Community practice radiology group	56 (55.4%)	17 (68.0%)	19 (52.8%)	20 (50.0%)
Academic radiology group	15 (14.9%)	1 (4.0%)	7 (19.4%)	7 (17.5%)

Physician and Practice Characteristics	All Participants *N*=101	Live Seminar *N*=25	Self-Paced DVD *N*=36	Control Group *N*=40
Radiologist in a multispecialty group	19 (18.8%)	3 (12.0%)	9 (25%)	7 (17.5%)
Solo radiology practice	5 (5.0%)	1 (4.0%)	1 (2.8%)	3 (7.5%)
Locum tenens	5 (5.0%)	2 (8.0%)	0 (0.0%)	3 (7.5%)
Other	1 (1.0%)	1 (4.0%)	0 (0.0%)	0 (0.0%)
Interpreted digital screening exams at time of survey, n (%)	81 (80.2%)	22 (88%)	29 (80.6%)	30 (75%)

Table 2

Summary of performance measures pre and post intervention, by intervention group relative to two reference outcomes

	Relative to Expert Recall			Relative to Cancer		
	Live n=25	**Self-Paced DVD n=36**	**Control n=40**	**Live Intervention n=25**	**Self-Paced DVD n=36**	**Control n=40**
Sensitivity, mean (sd)						
pre-intervention	51.3 (11.2)	52.3 (10.9)	54.4 (11.0)	62.0 (15.0)	66.1 (13.2)	65.5 (13.9)
post-intervention	61.2 (9.9)	63.1 (9.8)	61.4 (11.2)	66.1 (14.6)	69.8 (14.9)	65.8 (15.9)
change	10.0 (12.8)	10.9 (10.7)	7.0 (13.5)	4.1 (15.5)	3.7 (16.7)	0.3 (19.2)
Specificity, mean (sd)						
pre-intervention	77.9 (12.1)	74.8 (13.9)	73.1 (11.0)	69.3 (12.0)	65.8 (13.8)	64.1 (10.2)
post-intervention	76.8 (9.5)	75.0 (14.8)	74.9 (12.6)	69.2 (9.0)	67.8 (14.0)	67.5 (12.1)
change	−1.2 (12.3)	0.2 (15.5)	1.8 (10.3)	−0.1 (12.5)	1.9 (14.8)	3.4 (10)
Re-calibrated PPV[a], mean (sd)						
pre-intervention	48.5 (11.2)	46.2 (12.1)	44.4 (12.5)	25.6 (6.2)	25.3 (7.7)	23.4 (6.4)
post-intervention	50.7 (11.2)	51.5 (13.3)	49.4 (10.7)	26.2 (6.0)	27.5 (8.3)	25.4 (6.7)
change	2.2 (12.8)	5.3 (14.3)	4.9 (9.2)	0.6 (7.2)	2.2 (8.2)	2.1 (6.7)
Re-calibrated NPV[a], mean (sd)						
pre-intervention	88.3 (4.9)	90.1 (4.3)	89.0 (5.1)	94.4 (2.5)	95.3 (2.1)	94.8 (2.5)
post-intervention	89.5 (6.0)	90.0 (5.0)	89.7 (5.4)	95.0 (3.0)	95.3 (2.5)	95.1 (2.7)
change	1.2 (5.6)	−0.2 (5.1)	0.7 (6.0)	0.6 (2.8)	−0.1 (2.6)	0.4 (3.0)

Pre- and post-intervention performance measures have been rounded to the nearest tenth. Differences were computed at the subject level, with means and standard deviations calculated prior to rounding, resulting in apparent offsets of ±0.1 relative to the pre- and post-intervention values in some change values shown.

[a]Recalibrated PPV and NPV are calculated using the observed subject-specific estimates of sensitivity and specificity, at the appropriate level of analysis, and assume a cancer prevalence of (15/110) and an expert recall prevalence of (29/110), both of which correspond to the prevalence values present in the post-intervention test set.

Table 3

Comparison of the effects of intervention on various performance measures relative to two reference outcomes

	Relative to Expert Recall		Relative to Cancer	
	Adjusted OR[a]	p	Adjusted OR[a]	p
Sensitivity				
Live vs. Control	1.24 (0.90, 1.72)	0.190	1.22 (0.78, 1.90)	0.384
DVD vs. Control	1.34 (1.00, 1.81)	0.050	1.28 (0.85, 1.92)	0.237
Specificity				
Live vs. Control	0.80 (0.64, 1.00)	0.048	0.79 (0.65, 0.95)	0.015
DVD vs. Control	0.90 (0.74, 1.10)	0.299	0.92 (0.77, 1.09)	0.343
PPV				
Live vs. Control	1.13 (0.69, 1.86)	0.631	1.11 (0.59, 2.09)	0.743
DVD vs. Control	1.94 (1.24, 3.05)	0.004	1.81 (1.01, 3.23)	0.045
NPV				
Live vs. Control	1.08 (0.84, 1.39)	0.547	1.06 (0.74, 1.51)	0.752
DVD vs. Control	0.96 (0.77, 1.21)	0.760	0.94 (0.67, 1.30)	0.694

[a]Statistics shown here are performance measure odds ratios associated with intervention, adjusted for performance on the pre-intervention test sets. In the case of sensitivity, for example, the odds of participant recall given expert recall increased from pre-intervention to post-intervention in the Live group 1.24 (95% CI: 0.90, 1.72) times more than the corresponding increase in the Control group; this comparison was not statistically significant however, as the confidence interval included 1 (p=0.190). All models were adjusted for pre-intervention test set, and PPV and NPV models were additionally adjusted for the prevalence of cancer or expert recall.

Table 4

Number of participants showing improvement from pre- to post-intervention by intervention group, and adjusted improvement odds ratios

	Number improving from pre-intervention to post-intervention			Adjusted Improvement Odds Ratios (95% CI)[a]	
	Live n=25	Self-Paced DVD n=36	Control n=40	Live vs. Control	DVD vs. Control
Relative to Expert Recall					
Sensitivity	21 (84%)	29 (81%)	27 (68%)	2.6 (0.7, 9.5)	2.0 (0.7, 9.5)
Specificity	13 (52%)	19 (51%)	23 (58%)	0.8 (0.3, 2.3)	0.8 (0.3, 2.3)
Re-calibrated PPV[b]	16 (64%)	23 (64%)	28 (70%)	0.7 (0.3, 2.2)	0.7 (0.3, 2.2)
Re-calibrated NPV[b]	14 (56%)	16 (43%)	18 (45%)	1.6 (0.6, 4.5)	1.0 (0.4, 4.5)
Relative to Cancer					
Sensitivity	15 (60%)	18 (50%)	16 (40%)	2.3 (0.8, 6.8)	1.7 (0.6, 6.8)
Specificity	14 (56%)	20 (54%)	25 (62%)	0.8 (0.3, 2.2)	0.7 (0.3, 2.2)
Re-calibrated PPV[b]	16 (64%)	23 (64%)	26 (65%)	0.9 (0.3, 2.6)	1.0 (0.4, 2.6)
Re-calibrated NPV[b]	14 (56%)	16 (43%)	18 (45%)	1.6 (0.6, 4.5)	1.0 (0.4, 4.5)

[a] Odds ratio estimates are from logistic models of a binary indicator of improvement from pre- to post-intervention, regressed on intervention group, and are adjusted for pre-intervention test set assignment.

[b] Recalibrated PPV and NPV are calculated using the observed subject-specific estimates of sensitivity and specificity, and assume a cancer prevalence of (15/110) and an expert recall prevalence of (29/110), both of which correspond to the prevalence values present in the post-intervention test set.