Published in final edited form as:

AJR Am J Roentgenol. 2012 July ; 199(1): W134–W141. doi:10.2214/AJR.11.7701.

Is Confidence of Mammographic Assessment a Good Predictor of Accuracy?

Berta M. Geller¹, Andy Bogart², Patricia A. Carney³, Joann G. Elmore⁴, Barbara S. Monsees⁵, and Diana L. Miglioretti^{2,6}

¹Departments of Family Medicine and Radiology, University of Vermont, 1 S Prospect St, Rm 4426, 429AR4, Burlington, VT 05401-3444.

²Biostatistics Unit, Group Health Research Institute, Seattle, WA.

³Departments of Family Medicine and Public Health and Preventive Medicine, Oregon Health & Science University, Portland, OR.

⁴Department of Internal Medicine, University of Washington, Seattle, WA.

⁵Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, MO.

⁶Department of Biostatistics, University of Washington, Seattle, WA.

Abstract

OBJECTIVE—Interpretive accuracy varies among radiologists, especially in mammography.

This study examines the relationship between radiologists' confidence in their assessments and their accuracy in interpreting mammograms.

MATERIALS AND METHODS—In this study, 119 community radiologists interpreted 109 expert-defined screening mammography examinations in test sets and rated their confidence in their assessment for each case. They also provided a global assessment of their ability to interpret mammograms. Positive predictive value (PPV) and negative predictive value (NPV) were modeled as functions of self-rated confidence on each examination using log-linear regression estimated with generalized estimating equations. Reference measures were cancer status and expert-defined need for recall. Effect modification by weekly mammography volume was examined.

RESULTS—Radiologists who self-reported higher global interpretive ability tended to interpret more mammograms per week ($p = 0.08$), were more likely to specialize ($p = 0.02$) and to have completed a fellowship in breast or women's imaging ($p = 0.05$), and had a higher PPV for cancer detection ($p = 0.01$). Examinations for which low-volume radiologists were "very confident" had a PPV of 2.93 times (95% CI, 2.01–4.27) higher than examinations they rated with neutral confidence. Trends of increasing NPVs with increasing confidence were significant for low-volume radiologists relative to noncancers ($p = 0.01$) and expert nonrecalls ($p < 0.001$). A trend of significantly increasing NPVs existed for high-volume radiologists relative to expert nonrecall ($p = 0.02$) but not relative to noncancer status ($p = 0.32$).

CONCLUSION—Confidence in mammography assessments was associated with better accuracy, especially for low-volume readers. Asking for a second opinion when confidence in an assessment is low may increase accuracy.

Keywords

accuracy; confidence; mammography; radiologist performance; screening

Although screening mammography is currently the best way to detect early breast cancer, there continues to be unexplained variability in radiologists' interpretive performance [1–3]. To better understand this variability, studies examining radiologist characteristics associated with clinical accuracy have focused on interpretive volume [4–6], years of experience [5, 7], fellowship training [2, 7], radiologists' enjoyment of interpreting mammograms [8], the balance of screening and diagnostic interpretations [4], and the influence of medical malpractice [9]. Each prior study has contributed to a slightly better understanding of the factors related to performance variability.

Radiologists have expressed different levels of perceived competence in interpreting mammograms [10, 11], but this factor has not been fully investigated. Studies show that in other specialties, physicians and medical students who are more confident in their skills and knowledge have better performance in surgical skills, venous catheterization, and musculoskeletal clinical training [12–14]. Although several studies have examined radiologists' self-reported confidence in other areas of radiology [15, 16], there is no specific research, to our knowledge, about radiologists' confidence when interpreting mammography.

On the basis of the limited published research, we hypothesized that radiologists who have higher confidence levels in their assessments of individual screening mammography examinations would be more accurate in their interpretative performance of those mammograms, with fewer missed breast cancers and fewer unnecessary recalls. In this study we assessed community radiologists' global self-reported ability to interpret mammography and their level of confidence in interpreting each case using a test set of screening mammography examinations. We evaluated whether their perceived ability or confidence was associated with positive predictive value (PPV) and negative predictive value (NPV).

Materials and Methods

Recruitment and Test Set Development

Detailed methods for identification of the study radiologists, development and administration of the test set of screening mammography examinations, and study measures are described elsewhere [17]. Briefly, the study was conducted with six mammography registries associated with the Breast Cancer Surveillance Consortium (BCSC) [18]: Carolina Mammography Registry, New Hampshire Mammography Network, New Mexico Mammography Project, Vermont Breast Cancer Surveillance System, and Group Health Cooperative in western Washington. Radiologists who interpreted mammograms at a facility contributing to a registry between January 2005 and December 2006 were eligible and invited to participate. In addition, 103 non-BCSC radiologists from Oregon, the Puget Sound region of Washington State, North Carolina, San Francisco, and New Mexico were invited to take part.

A total of 469 radiologists were invited to participate, and 148 (31.6%) consented. Among these, 119 (80.4%) completed this study. Each site received institutional review board (IRB) approval to obtain active consent from radiologists to interpret test sets and to link their

information and test set performance to the mammograms they interpreted in clinical practice (for BCSC radiologists). In addition, each registry and the BCSC Statistical Coordinating Center received IRB approval for either active or passive consenting processes or a waiver of consent to enroll women undergoing mammography at a BCSC facility, link data, and perform analytic studies. All procedures were HIPAA-compliant and all registries and the Statistical Coordinating Center received a Federal Certificate of Confidentiality and other protection for the identities of women, physicians, and facilities who are the subjects of this research.

The test set cases ($n = 130$) were randomly selected from film-screen screening mammography examinations of 40- to 69-year-old women between 2000 and 2003 from the six participating BCSC registries for a larger study that will be reported elsewhere. All cases had a comparison screening mammographic examination within 2 years before the examination. Women who had breast augmentation or a history of breast cancer were excluded.

Each case consisted of craniocaudal and mediolateral oblique views of each breast (4 views per woman for each of the screening and comparison examinations). For cancer test set cases ($n = 36$), we selected images from examinations for which invasive breast cancer or ductal carcinoma in situ (DCIS) breast cancer was diagnosed within 12 months after mammography. Noncancer cases ($n = 94$) were women who had at least 2 years of cancer-free follow-up after screening mammography. A panel of three expert radiologists reviewed and agreed on all the cases to be recalled. The final four mammography test sets included 109 cases, 91 of which were in all four. Test set cases varied by cancer prevalence and expert-rated difficulty to identify breast cancer. We used this approach because the goal of the larger study was to assess how cancer prevalence and difficulty of findings (subtle, intermediate, or obvious) interpreted on a test set would correlate with clinical practice.

Identifying information was removed from the mammography films, and the films were professionally digitized by American College of Radiology (ACR) staff and uploaded into custom-designed software created in conjunction with the ACR for viewing and collecting radiologist's interpretations. Details of the process, software, and test set development and composition are described elsewhere [19]. Participating radiologists viewed cases presented in a random order. Each case was presented in a sequence including mediolateral oblique and craniocaudal views of both breasts simultaneously, followed by mediolateral oblique and craniocaudal views of each breast paired with the analogous image from a previous examination to assess whether changes from the prior mammographic examination were apparent.

Consenting radiologists were randomized to one of the four test sets that they interpreted using self-administered, custom-designed software on a digital video disc (DVD). Radiologists were informed that the overall cancer rate of the test sets was higher than that found in a screened population, but they were not informed of the specific prevalence of positive examinations or cancers in the test sets to avoid influencing their interpretations.

Radiologists used either a home or work computer or a laptop provided by the study with a large screen and high-resolution graphics to show two images simultaneously (screen resolution, 1280×1024 ; processor speed, 3 GHz; at least 1 GB of random access memory [RAM], a 128-MB video card capable of displaying 32-bit color, and a DVD reader).

Before interpreting cases, radiologists answered demographic and clinical practice survey questions including self-reported global ability to perceive and identify important mammographic findings. We linked our study data to sex information for 93 of the 119

participating radiologists who had participated in a prior study [20]. For the remaining 26, each study site obtained IRB approval as necessary to collect and provide missing sex data, resulting in complete capture of sex.

For each test set case, radiologists indicated whether they would recall the patient (i.e., BI-RADS assessment category of 0, 4, or 5, which was considered “positive”) or not (BI-RADS assessment category of 1 or 2, which was considered “negative”) [21]. BI-RADS category 3 was not an assessment choice. In addition, they rated their level of confidence in their recall—no recall assessment as not at all confident, not very confident, neutral, confident, or very confident. Because of the low number of responses in the “not at all confident” and “not very confident” categories, these were combined into a “not confident” category for analysis.

Reference Measures

We measured interpretive performance using PPV and NPV of a recall. We used two separate reference measures: first, the development of cancer within 12 months of imaging; and, second, either the development of a cancer or a recall of a noncancer case recommended by our panel of three expert radiologists, which we term “expert recall” in the remainder of the article.

Statistical Analysis

We calculated frequency distributions for radiologists’ demographic characteristics and clinical experience both overall and according to their self-reported global ability to perceive and identify important mammographic findings. We used the Fisher exact test to assess the significance of differences observed among participants who rated their interpretive ability as average, above average, or expert.

We examined the effect of self-reported global ability and examination-level confidence on test set PPV and NPV using log-linear regression. To model PPV relative to 12-month cancer status, we fit log-linear generalized estimating equation (GEE) models on the subset of test set cases recalled by participants, with 12-month cancer status as a binary outcome, assuming a robust Poisson variance. This approach yields valid variance estimates for relative risk of common binary outcomes [22]. To model PPV relative to recall as determined by our expert panel, we repeated this analysis on the same subset of test set examinations using expert recall as the binary outcome. We modeled the NPV with respect to both the cancer-status and expert-recall reference measures by fitting similar models on the subset of examinations that were not recalled by participants. Models for the effect of self-reported global ability were adjusted for the prevalence of the appropriate reference measure (cancer or expert recall), which differed according to test set assignment. Models for the effect of examination-level confidence were adjusted for radiologist sex, test set assignment, practice specialization, fellowship training, years of mammography interpretation, and use of digital mammography equipment in practice. We estimated the relative PPV and NPV for each examination-level confidence using neutral confidence as the referent category. Trend tests were conducted by refitting the models treating confidence as an ordinal categorical variable with values of from 1 to 4.

We hypothesized that the effect of examination-level confidence on interpretive accuracy would differ between participants who interpret mammograms frequently versus those who do not. Therefore, we allowed an interaction between examination-level confidence and weekly interpretive volume, which we dichotomized into low volume (0–99 mammograms read per week) and high volume (≥ 100 mammograms read per week).

To account for potential correlation both within examinations (the same case interpreted by different radiologists) and within radiologists (different cases interpreted by the same radiologist), all regression models were estimated using an extension of GEEs that accommodates this nonnested clustering of radiologists and cases [23, 24]. The method relies on an independent working correlation structure and makes use of the robust Huber-White sandwich estimator of regression parameter standard errors [25]. Tests of statistical significance were two-sided with an alpha level of 0.05. All analyses were conducted using statistics software [26] (R, version 2.10, R Project for Statistical Computing; or SAS, version 9.2, SAS Institute).

Results

A total of 119 participating radiologists completed a mammography test set. Interpretation data were missing for seven cases (one radiologist reviewed images from only 104 patients instead of 109 and two reviews were excluded because of missing examination-level confidence ratings), resulting in a total of 12,964 reviews available for analysis.

All participants except five rated themselves as having at least average global ability to perceive and determine the importance of mammographic findings. Two indicated that they were below average, and three indicated that they were not sure. Table 1 presents information on the radiologists' clinical experience both overall and by their self-rated global mammographic interpretive ability. Despite random assignment to test sets, self-rated global ability differed significantly by test set: A higher proportion of radiologists rated their ability as average for test sets 1 and 2, above average for test sets 3 and 4, and expert for tests 1 and 4 ($p = 0.02$). Participants expressing higher global interpretive ability tended to read more mammograms per week ($p = 0.08$), were more likely to specialize in their radiology work ($p = 0.02$), and were more likely to have completed a fellowship in breast or women's imaging ($p = 0.05$). Other pretest set survey characteristics did not differ significantly by self-reported interpretive ability.

In Table 2 we present estimates of PPV and NPV by global ability ratings relative to both cancer status and expert recall as reference measures. The prevalence of each reference measure differed by test set, so raw estimates of PPV and NPV are not comparable across test set groups because both PPV and NPV are functions of disease prevalence. Test sets 1 and 2 had a relatively lower prevalence of cancer (15/109) and of expert recalls (36/109), whereas test sets 3 and 4 each had higher cancer prevalence (30/109) and more expert recalls (51/109). To allow comparisons across test sets, we present estimates that are adjusted to a common prevalence of cancer (15/109) and expert recall (36/109).

The mean PPV relative to cancer differed significantly across groups defined by the three global ability rating levels ($p = 0.01$): Radiologists reporting above average ability had 1.06 times (95% CI, 1.02–1.09) higher PPV than the average ability group and those reporting expert ability had 1.08 times (95% CI, 1.00–1.16) higher PPV than those reporting average ability. Those reporting expert ability did not differ significantly from the above-average group, having 1.02 times (95% CI, 0.96–1.08) higher PPV, with a CI straddling the null value of 1. These groups did not differ significantly in their mean PPVs relative to expert recall or in their mean NPVs relative to either reference measure.

The frequency distributions of examination-specific confidence ratings for the two reference measures (cancer status and expert recall) are shown separately for positively and negatively interpreted test cases in Table 3. Confidence in positive assessments tended to be higher for cases with cancer within 12 months and lower for cases that remained cancer-free for 24 months ($p < 0.001$). Analysis by expert recall showed a similar pattern: Confidence in

positive assessments for cases also recalled by the panel of experts tended to be higher than cases not recalled by the expert panel ($p < 0.001$). Among negative assessments, the distribution of confidence ratings was similar for both cancer and noncancer cases ($p = 0.15$). Examination-level confidence ratings on negative assessments differed by expert recall status, with higher confidence ratings on examinations of images that the experts also chose not to recall ($p = 0.01$).

Adjusted estimates of relative PPVs and NPVs are shown separately by reference measures and radiologist volume in Figure 1. For both low- and high-volume radiologists, the adjusted PPVs for cancer increased with radiologists' perceived confidence in their interpretation. Among low-volume participants, the PPVs of reviews for which they were very confident was 2.93 times (95% CI, 2.01–4.27) higher than for reviews they rated with neutral confidence. Similarly, positive assessments rated by high-volume readers as very confident were 2.86 times (95% CI, 1.91–4.27) more likely to be a cancer case than positive assessments they rated with neutral confidence. Trends of increasing relative PPVs with increasing confidence were significant for low- and high-volume readers for both referent outcome measures (each $p < 0.001$). The pattern of increasing PPV with increasing confidence did not differ significantly by volume for cancer status ($p = 0.80$) or for recall by the expert panel ($p = 0.47$).

The relationship between confidence and NPV differed significantly by interpretive volume—both in identifying noncancer cases ($p = 0.01$) and in identifying cases not recalled by experts ($p = 0.004$). Estimates of NPVs among high-volume readers did not vary significantly by examination-level confidence ratings for either reference measure relative to neutral confidence readers. Among low-volume readers, however, negative assessments in which the reader was very confident were 1.07 times (95% CI, 1.03–1.12) more likely to be a noncancer case and 1.25 times (95% CI, 1.13–1.37) more likely to be an expert nonrecall than negative assessments in which confidence was neutral. Negative assessments in which low-volume readers were not confident were less likely to be an expert nonrecall than negative assessments that were confidence neutral (relative ratio, 0.86; 95% CI, 0.75–0.98) but did not differ significantly in adjusted NPV in likelihood of being a noncancer. Trends of increasing NPVs with increasing confidence were significant for low-volume readers for noncancers ($p = 0.01$) and expert nonrecall cases ($p < 0.001$). A significant increasing trend in NPV existed for high-volume radiologists relative to expert nonrecall ($p = 0.02$) but not relative to noncancer status ($p = 0.32$).

Discussion

In this study we found that radiologists who rated their global interpretive ability as above average or as expert had a higher PPV for cancer than radiologists who rated their ability as average or below average. This finding suggests that radiologists' perceptions of their performance match their ability to recall patients for workup on the basis of screening mammograms. We also found that radiologists' confidence in their assessments to recall a patient on the basis of screening mammography appeared to be higher for both cancer and noncancer cases that a panel of experts determined should be recalled and to be lower for cases for which the expert panel determined that recall was unwarranted. Interestingly, the relationship between confidence and accuracy of a negative assessment depended on whether the radiologists interpreted 100 or more mammograms per week or fewer than 100 mammograms per week.

Measuring radiologists' self-confidence in interpreting individual screening mammograms may have several practical applications. Because we found that self-reported ratings of confidence are associated with both PPV and NPV, we believe that when radiologists have

low confidence in their assessment they may benefit from asking a colleague to review the case to assist in the recall decision. In addition to potentially improving assessment accuracy, this strategy might provide an opportunity for radiologists to learn from each other by stimulating discussion about whether the findings are suspicious enough to recall. It is likely that many radiologists currently do this informally in clinical practice when another radiologist is available. Digital images may increase the possibility of seeking a second opinion because they are easier to share with radiologists outside one's practice.

The mammographic interpretive volume of radiologists appeared to modify the effect of confidence on NPV. When radiologists interpreted a mammography examination as negative, confidence levels did not affect NPV for radiologists with a high weekly volume of mammograms. However, the NPV increased with self-reported confidence for low-volume radiologists. These results suggest that confidence in correctly assessing a true-negative examination is particularly important for the low-volume radiologists and that before deciding against recall they might particularly benefit from reviewing with a colleague the examinations for which they have low confidence in their assessment.

In our current study we found that the PPV for cancer differed by the global measure of self-rated ability to perceive and determine the importance of mammographic findings. Among high-performing individuals, such as competitive athletes, performance is often linked to self-efficacy [27–29], which is a generalized belief in one's ability to succeed in a particular situation, similar to our global measure. Our study shows that both global self-efficacy and confidence related to a radiologist's ability to correctly interpret an individual patient's mammographic examination are associated with increased PPV.

Many factors may be related to radiologists' confidence in their assessments, and confidence may affect radiologists' mammographic interpretation behavior in surprising ways. A recent study using the same dataset as this study examined the relationship between viewing time and accuracy and reported that confidence was negatively related to increased viewing time [17]. For those who were very confident in their assessment of an individual examination, each additional minute of viewing time increased the adjusted risk of a false-positive examination by a factor of 1.42 (95% CI, 1.21–1.68). As confidence levels dropped, so did the risk of false-positive examinations with each additional minute of viewing time.

There are several limitations to this study. First, using a test set is an artificial setting and the relationship between confidence and accuracy may be different in the clinical setting. Digitized films were interpreted on computers that might have made the images more difficult to interpret than on workstations, lowering the radiologists' confidence in their assessments. However, the advantage of using a test set is that radiologists interpreted mostly the same cases, which helped control for interradiologist variability found in clinical practice.

In conclusion, our study examined the association of radiologists' interpretive accuracy with their self-reported level of confidence at the examination level and to a global measure of their ability. We found that PPV increased with increasing levels of confidence. We also found that radiologists who interpreted fewer than 100 mammograms a week were significantly more accurate in the negative assessments (i.e., higher NPV) for which they were more confident. Asking for a second reading when confidence in an assessment of negative mammographic findings is low could potentially increase cancer detection. Asking for a second reading when confidence in an assessment of recall is low could potentially reduce unnecessary recalls. However, radiologists may not always have a colleague available to review a case and getting a second opinion is not currently reimbursed by Medicare or by most private insurance companies. Although this study was performed with

digitized film-screen mammograms, it is likely that the same results would be observed in the interpretation of full-field digital mammograms. Conversion to digital mammography would facilitate intrapractice and teleradiology consultative review [30], enabling double reading of cases in which the primary interpreting radiologist has a low level of confidence. Reimbursement of such a practice may improve the overall accuracy of mammographic interpretation even if the interpreting radiologist asks for review of only the troublesome cases.

Acknowledgments

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/>.

We also thank Jose Cayere and Amy Buzby and the American College of Radiology for technical assistance in developing and supporting implementation of the test sets. Their work was invaluable to the success of this project.

This work was supported by the American Cancer Society and made possible by a generous donation from the Longaberger Company's Horizon of Hope Campaign (SIRGS-07-271-01, SIRGS-07-272-01, SIRGS-07-274-01, SIRGS-07-275-01, SIRGS-06-281-01, ACS A1-07-362). It was also supported by the National Cancer Institute Breast Cancer Surveillance Consortium (BCSC) (U01CA63740, U01CA86076, U01CA86082, U01CA63736, 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 United States. For a full description of these sources, please see: <http://breastscreening.cancer.gov/work/acknowledgement.html>.

References

1. Duijm L, Louwman M, Groenewoud J, et al. Interobserver variability in mammography screening and effect of type and number of readers on screening outcome. *Br J Cancer*. 2009; 100:901–907. [PubMed: 19259088]
2. Saunders RS, Samei E. Improving mammographic decision accuracy by incorporating observer ratings with interpretation time. *Br J Radiol*. 2006; 79(spec no 2):S117–S122. [PubMed: 17209116]
3. Beam C, Layde P, Sullivan D. Variability in the interpretation of screening mammograms by US radiologists: findings from a national sample. *Arch Intern Med*. 1996; 156:209–213. [PubMed: 8546556]
4. Buist D, Anderson M, Haneuse S, et al. Influence of annual interpretive volume on screening mammography performance in the United States. *Radiology*. 2011; 259:72–84. [PubMed: 21343539]
5. Tchou PM, Haygood TM, Atkinson EN, et al. Interpretation time of computer aided-detection at screening mammography. *Radiology*. 2010; 257:40–46. [PubMed: 20679448]
6. Haygood TM, Wang J, Atkinson EN, et al. Timed efficiency of interpretation of digital and film-screen screening mammograms. *AJR*. 2009; 192:216–220. [PubMed: 19098202]
7. Berns EA, Hendrick RE, Solari M, et al. Digital and screen-film mammography: comparison of image acquisition and interpretation times. *AJR*. 2006; 187:38–41. [PubMed: 16794152]
8. Ishiyama 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; 16:58–61. [PubMed: 18836795]
9. Carney PA, Geller BM, Moffett H, et al. Current medico-legal and confidentiality issues in large multi-center research programs. *Am J Epidemiol*. 2000; 152:371–378. [PubMed: 10968382]
10. Feig S. Adverse effects of screening mammography. *Radiol Clin North Am*. 2004; 42:807–819. [PubMed: 15337417]
11. Luo P, Eikman E, Kealy W, et al. Analysis of a mammography teaching program based on an affordance design model. *Acad Radiol*. 2006; 13:1542–1552. [PubMed: 17138123]
12. Antonoff M, Shelstad R, Schmitz C, et al. A novel critical skills curriculum for surgical interns incorporating simulation training improves readiness for acute inpatient care. *J Surg Educ*. 2009; 66:248–254. [PubMed: 20005496]

13. Millington S, Wong R, Kassen B, et al. Improving internal medicine residents' performance, knowledge, and confidence in central venous catheterization using simulators. *J Hosp Med*. 2009; 4:410–416. [PubMed: 19753575]
14. Burke J, Fayaz S, Graham K, et al. Peer-assisted learning in the acquisition of clinical skills: a supplementary approach to musculoskeletal system training. *Med Teach*. 2007; 29:577–582. [PubMed: 17978969]
15. Nietert P, Ravenel J, Taylor K, et al. Influence of nodule detection software on radiologists' confidence in identifying pulmonary nodules with computed tomography. *J Thorac Imaging*. 2011; 26:48–53. [PubMed: 20498624]
16. Peña BM, Taylor G. Radiologists' confidence in interpretation of sonography and CT in suspected pediatric appendicitis. *AJR*. 2000; 175:71–74. [PubMed: 10882249]
17. Carney PA, Bogart TA, Geller BM, et al. Association between time spent interpreting, level of confidence, and accuracy of screening mammography. *AJR*. 2012; 198:970–978. [PubMed: 22451568]
18. National Cancer Institute. [Accessed March 9, 2012] Breast Cancer Surveillance Consortium Website. Breast Cancer Surveillance Consortium. breastscreening.cancer.gov. Updated January 27, 2012.
19. Højsgaard S, Halekoh U, Yan J. The R Package geepack for generalized estimating equations. *J Stat Software*. 2005; 15:1–11.
20. Straus SE, Tetroe J, Graham I. Defining knowledge translation. *CMAJ*. 2009; 181:3–4.
21. D'Orsi, C.; Bassett, L.; Berg, WA., et al. ACR BI-RADS-Breast Imaging Atlas. 4th ed.. Reston, VA: American College of Radiology; 2003.
22. McNutt L, Wu C, Xue X, et al. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol*. 2003; 157:940–943. [PubMed: 12746247]
23. Kundel HL, Nodine CF, Conant EF, Weinstein SP. Holistic component of image perception in mammogram interpretation: gaze-tracking study. *Radiology*. 2007; 242:396–402. [PubMed: 17255410]
24. Mello-Thoms C, Dunn S, Nodine CF, Kundel HL, Weinstein SP. The perception of breast cancer: what differentiates missed from reported cancers in mammography? *Acad Radiol*. 2002; 9:1004–1012. [PubMed: 12238541]
25. Diggle, P.; Heagerty, P.; Liang, K., et al. The analysis of longitudinal data. 2nd ed.. New York, NY: Oxford University Press; 2002.
26. R Development Core Team R. A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2011.
27. Taylor J. Predicting athletic performance with self-confidence and somatic and cognitive anxiety as a function of motor and physiological requirements in six sports. *J Pers*. 1987; 3:1–15.
28. Mahoney M, Avenier M. Psychology of the elite athlete: an exploratory study. *Cognit Ther Res*. 1977; 1:135–141.
29. Bandura A. Self-efficacy towards a unifying theory of behavioral change. *Psychol Rev*. 1977; 84:191–215. [PubMed: 847061]
30. Joffe S, Burak J, Rackson M, et al. The effect of international teleradiology attending radiologist coverage on radiology residents' perceptions of night call. *J Am Coll Radiol*. 2006; 3:872–878. [PubMed: 17412187]

Fig. 1.

Relative positive predictive values (PPVs) and negative predictive values (NPVs) by examination-level confidence and weekly mammographic volume for two reference measures (i.e., cancer status and expert recall). Low volume indicates 0–99 mammograms read per week and high volume, 100 or more mammograms read per week.

TABLE 1
Participants' Characteristics by Self-Rated Ability to Perceive and Determine Importance of Mammographic Findings

Participants' Characteristics and Survey Responses	All Participants ^a (n = 119)	Self-Rated Ability to Perceive and Determine the Importance of Mammographic Findings			p ^b
		Average (n = 53)	Above Average (n = 50)	Expert (n = 11)	
Female, no. (%)	56 (47)	21 (40)	29 (58)	5 (45)	0.17
Test set, no. (%)					0.02
1	30 (25)	15 (28)	9 (18)	4 (36)	
2	34 (29)	20 (38)	12 (24)	2 (18)	
3	28 (24)	12 (23)	15 (30)	0 (0)	
4	27 (23)	6 (11)	14 (28)	5 (45)	
Number of mammography examinations I read a week, no. (%)					0.08
Up to 10	8 (7)	3 (6)	3 (6)	0 (0)	
11–49	25 (21)	14 (26)	8 (16)	1 (9)	
50–99	36 (30)	20 (38)	13 (26)	3 (27)	
100–199	31 (26)	13 (25)	16 (32)	2 (18)	
200 or more	19 (16)	3 (6)	10 (20)	5 (45)	
How many category 1 hours in mammography have you received in the past 3 years (not including this program), no. (%)					0.28
none	4 (3)	3 (6)	0 (0)	0 (0)	
1–10	7 (6)	3 (6)	3 (6)	1 (9)	
11–15	21 (18)	12 (23)	7 (14)	0 (0)	
16–30	45 (38)	21 (40)	18 (36)	5 (45)	
31 or more	42 (35)	14 (26)	22 (44)	5 (45)	
How specialized are you in your clinical work in radiology?, no. (%)					0.02
Generalist (with no specialization)	26 (22)	17 (32)	7 (14)	0 (0)	
Primarily generalist (with some specialization)	48 (40)	22 (42)	20 (40)	4 (36)	
Primarily specialist (with some general work)	36 (30)	13 (25)	18 (36)	4 (36)	
Breast specialist (no general work)	9 (8)	1 (2)	5 (10)	3 (27)	
Which one of the following BEST describes your current main practice?, no. (%)					0.22
Community practice radiology group	66 (55)	33 (62)	27 (54)	3 (27)	
Academic radiology group	17 (14)	6 (11)	7 (14)	4 (36)	

Participants' Characteristics and Survey Responses	All Participants ^a (n = 119)	Self-Rated Ability to Perceive and Determine the Importance of Mammographic Findings			<i>p</i> ^b
		Average (n = 53)	Above Average (n = 50)	Expert (n = 11)	
Radiologist in a multispecialty group	21 (18)	8 (15)	11 (22)	2 (18)	0.05
Solo radiology practice	6 (5)	2 (4)	2 (4)	1 (9)	
Locum tenens	7 (6)	4 (8)	1 (2)	1 (9)	
Other	2 (2)	0 (0)	2 (4)	0 (0)	
Completed a fellowship after residency training, no. (%)					
Yes, in breast or women's imaging	15 (13)	5 (9)	4 (8)	5 (45)	0.26
Yes, other	51 (43)	21 (40)	24 (48)	3 (27)	
No	53 (45)	27 (51)	22 (44)	3 (27)	
How many years have you been interpreting mammograms?, no. (%)					
1–5 years	25 (21)	17 (32)	5 (10)	2 (18)	
6–10 years	18 (15)	7 (13)	9 (18)	1 (9)	0.18
11–20 years	51 (43)	19 (36)	23 (46)	6 (55)	
21–30 years	17 (14)	6 (11)	10 (20)	1 (9)	
31 years or more	8 (7)	4 (8)	3 (6)	1 (9)	
Please estimate the average number of mammograms per year you interpreted over the last 5 years, no. (%)					
Don't know	29 (24)	12 (23)	12 (24)	1 (9)	0.82
0–1000	15 (13)	11 (21)	3 (6)	1 (9)	
1001–2000	26 (22)	13 (25)	10 (20)	2 (18)	
2001–3000	22 (18)	10 (19)	10 (20)	2 (18)	
More than 3000	27 (23)	7 (13)	15 (30)	5 (45)	
Do you currently interpret digital screening examinations?, no. (%)					0.01
Yes	96 (81)	45 (85)	40 (80)	9 (82)	
PPV and NPV, mean (95% CI)					
Relative to cancer, prevalence set to 15/109 = 13.8%					
PPV	26.4 (17.5–39.8)	25.7 (17.0–39.0)	27.2 (18.2–40.7)	27.7 (18.3–41.9)	
NPV	95.1 (91.6–98.7)	94.9 (91.2–98.7)	94.9 (91.9–99.0)	95.0 (91.6–98.6)	0.55
Relative to expert recall, prevalence set to 36/109 = 33%					
PPV	56.5 (46.6–68.5)	55.9 (45.8–68.1)	57.2 (47.3–69.2)	58.8 (48.7–71.1)	0.29

Participants' Characteristics and Survey Responses	All Participants ^a (<i>n</i> = 119)	Self-Rated Ability to Perceive and Determine the Importance of Mammographic Findings			<i>p</i> ^b
		Average (<i>n</i> = 53)	Above Average (<i>n</i> = 50)	Expert (<i>n</i> = 11)	
NPV	83.4 (76.9–90.4)	83.5 (77.0–90.6)	83.4 (76.7–90.6)	83.0 (76.4–90.1)	0.92

Note—PPV = positive predictive value, NPV = negative predictive value.

^aThree participants indicated “Not sure” and two indicated “Below average” in response to the item “Rate how you would describe your ability to perceive and determine the importance of mammographic findings.” These five individuals are only included in the total of 119 participants.

^b*p* values for categorical measures are from Fisher exact test of tables excluding the five participants mentioned in footnote a; *p* values for PPV and NPV comparisons are from 2 degrees-of-freedom Wald tests from generalized estimating equation models of performance regressed on self-rated confidence adjusted for outcome prevalence. These models excluded the five individuals mentioned in footnote a and accommodated nonnested correlation within patient and reader as described in the text.

TABLE 2

Positive Predictive Value (PPV) and Negative Predictive Value (NPV) Overall and by Self-Rated Ability Group for Two Reference Measures

Reference Measure	All Participants ^a (n = 119)	Self-Rated Ability to Perceive and Determine the Importance of Mammographic Findings			<i>p</i> ^b
		Average (n = 53)	Above Average (n = 50)	Expert (n = 11)	
Cancer, prevalence set to 15/109 = 13.8%					
PPV	26.4 (17.5–39.8)	25.7 (17.0–39.0)	27.2 (18.2–40.7)	27.7 (18.3–41.9)	0.01
NPV	95.1 (91.6–98.7)	94.9 (91.2–98.7)	94.9 (91.9–99.0)	95.0 (91.6–98.6)	0.55
Expert recall, prevalence set to 36/109 = 33%					
PPV	56.5 (46.6–68.5)	55.9 (45.8–68.1)	57.2 (47.3–69.2)	58.8 (48.7–71.1)	0.29
NPV	83.4 (76.9–90.4)	83.5 (77.0–90.6)	83.4 (76.7–90.6)	83.0 (76.4–90.1)	0.92

^aThree participants indicated “Not sure” and two indicated “Below average” in response to the item “Rate how you would describe your ability to perceive and determine the importance of mammographic findings.” These five individuals are included in the total of 119 participants.

^b*p* values for PPV and NPV comparisons are from 2 degrees-of-freedom Wald tests from generalized estimating equation models of performance regressed on self-rated perceived ability, adjusted for outcome prevalence, where the null hypothesis is that all three ability groups have identical PPV or NPV after adjustment for differences in the prevalence of cases in the test sets viewed. These models excluded the five individuals mentioned in footnote a and accommodated nonnested correlation within patient and reader as described in the text.

TABLE 3

Frequencies of Examination-Level Confidence Ratings for Mammographic Test Set Examinations by Radiologist Assessment for Two Reference Outcome Measures

Examination-Level Confidence Ratings	Cancer Status as Reference Measure		Expert Recall as Reference Measure	
	Cancer	Noncancer	Expert Recalls	Expert Nonrecalls
Positive assessments, no. of examinations	1995	3580	3555	2020
Examination-level confidence, no. (%)				
Not confident ^a	86 (4)	476 (13)	223 (6)	339 (17)
Neutral	416 (21)	1390 (39)	894 (25)	912 (45)
Confident	967 (48)	1477 (41)	1739 (49)	705 (35)
Very confident	526 (26)	237 (7)	699 (20)	64 (3)
<i>p</i> ^b	< 0.001		< 0.001	
Negative assessments, no. of examinations	614	6775	1552	5837
Examination-level confidence, no. (%)				
Not confident ^a	49 (8)	333 (5)	134 (9)	248 (4)
Neutral	201 (33)	1908 (28)	509 (33)	1600 (27)
Confident	293 (48)	3742 (55)	767 (49)	3268 (56)
Very confident	71 (12)	792 (12)	142 (9)	721 (12)
<i>p</i> ^b	0.15		0.01	

Note—Examinations were conducted on images from 36 cancer cases and 94 noncancer cases. A panel of three expert radiologists assessed all cases: They recalled all cancer cases and 21 noncancer cases.

^a“Not at all confident” and “Not very confident” are combined in the “Not confident” category.

^b*p* values are from 3 degrees-of-freedom Wald tests from generalized estimating equation models of the relevant binary reference measure, regressed on examination-level confidence rating adjusted for outcome prevalence. The models accommodated nonnested correlation within patient and reader as described in the text.