Title

Establishing a Gold Standard for Test Sets Variation in Interpretive Agreement of Expert Mammographers

Journal

Academic Radiology, 20(6)

ISSN

1076-6332

Authors

Onega, Tracy
Anderson, Melissa L
Miglioretti, Diana L
et al.

Publication Date

2013-06-01

DOI

10.1016/j.acra.2013.01.012

Peer reviewed

Published in final edited form as:
Acad Radiol. 2013 June ; 20(6): 731–739. doi:10.1016/j.acra.2013.01.012.

Establishing a Gold-standard for Test Sets: Variation in Interpretive Agreement of Expert Mammographers

Tracy Onega[1], **Melissa L. Anderson**[2], **Diana L. Miglioretti**[2,*], **Diana S.M. Buist**[2], **Berta Geller**[3], **Andy Bogart**[2], **Robert A. Smith**[4], **Edward A. Sickles**[5], **Barbara Monsees**[6], **Lawrence Bassett**[7], **Patricia A. Carney**[8], **Karla Kerlikowske**[9], and **Bonnie C. Yankaskas**[10]

[1]Department of Community & Family Medicine, Norris Cotton Cancer Center, and The Dartmouth Institute for Health Policy & Clinical Practice, Dartmouth Medical School, Lebanon, NH

[2]Group Health Research Institute, Seattle, WA

[*]Department of Biostatistics, University of Washington, Seattle, WA

[3]Departments of Family Medicine and Radiology, University of Vermont, Burlington, VT

[4]American Cancer Society

[5]Department of Radiology, University of California, San Francisco, CA

[6]Department of Radiology, Washington University, St. Louis, MO

[7]Department of Radiology, University of California, Los Angeles, CA

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

[9]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA and General Internal Medicine Section, Department of Veterans Affairs, University of California, San Francisco, CA

[10]Department of Radiology, University of North Carolina at Chapel Hill, NC

Abstract

Rationale and Objectives—Test sets for assessing and improving radiologic image interpretation have been used for decades and typically evaluate performance relative to gold-standard interpretations by experts. To assess test sets for screening mammography, a gold-standard for whether a woman should be recalled for additional work-up is needed, given that interval cancers may be occult on mammography and some findings ultimately determined to be benign require additional imaging to determine if biopsy is warranted. Using experts to set a gold-standard assumes little variation occurs in their interpretations, but this has not been explicitly studied in mammography.

Materials and Methods—Using digitized films from 314 screening mammography exams (n=143 cancer cases) performed in the Breast Cancer Surveillance Consortium, we evaluated interpretive agreement among three expert radiologists who independently assessed whether each

CORRESPONDENCE: Tracy Onega, PhD, Dartmouth Medical School, HB 7927, Rubin 8 – DHMC, One Medical Center Dr. Lebanon, NH 03756, Phone: 603-653-3671, Fax: 603-653-9020, Tracy.L.Onega@dartmouth.edu.

examination should be recalled, and the lesion location, finding type (mass, calcification, asymmetric density, or architectural distortion), and interpretive difficulty in the recalled images.

Results—Agreement among the three expert pairs for recall/no recall was higher for cancer cases (mean 74.3 ± 6.5) than for non-cancers (mean 62.6 ± 7.1). Complete agreement on recall, lesion location, finding type and difficulty ranged from 36.4%–42.0% for cancer cases and from 43.9%–65.6% for non-cancer cases. Two of three experts agreed on recall and lesion location for 95.1% of cancer cases and 91.8% of non-cancer cases, but all three experts agreed on only 55.2% of cancer cases and 42.1% of non-cancer cases.

Conclusion—Variability in expert interpretive is notable. A minimum of three independent experts combined with a consensus should be used for establishing any gold-standard interpretation for test sets, especially for non-cancer cases.

Keywords

mammography; test set; expert opinion; gold standard; variation

INTRODUCTION

In radiology, test sets have been used for decades to assess and improve interpretive performance (1,2). Typically, the gold-standard for interpretation is either based on observed patient outcomes, or is based on expert review where a panel of experts comes to consensus on the interpretation. In the latter, the consensus decision becomes the gold-standard and provides the basis for measuring individual performance. Little is known about the extent to which expert radiologists vary in interpretive assessments. Importantly, agreement among mammography experts has not been examined in the context of test set development in screening mammography, though test sets are frequently used for educational purposes.

The high prevalence of screening mammography use in the population and wide variability in radiologists' interpretive performance of mammography (3,4) makes the issue of testing radiologists for interpretation ability clinically important. Test sets are also useful for evaluating interventions aimed at improving interpretation, because it is difficult to assess changes in screening mammography performance in clinical practice due to low within-practice breast cancer prevalence and the long lag-time for obtaining true cancer status. For screening mammography test sets, breast cancer status may be considered the ultimate gold-standard, but it also unrealistically applies a diagnostic standard of performance to a screening test. In contrast, a gold standard for whether or not the exam should be recalled for additional work-up and the location and type of any significant findings would be more clinically relevant, because it measures performance based on the fact that some interval cancers are occult on prior mammography and some findings ultimately determined to be benign required additional imaging to determine if biopsy is warranted. The nature of screening makes having clear objective criteria for a recall decision to evaluate a suspicious finding or identification of significant findings difficult, but using biopsy results within one year of screening as the gold standard unrealistically judges all false negatives and false positives as avoidable errors.

Using experts to set the gold-standard for recall presumes there is little variation in their interpretive threshold. However, little empirical evidence exists upon which to base this assumption for screening mammography, since most studies where gold-standards have been established were not limited to experts. Experts may often use different nomenclature to describe essentially the same mammographic finding. However, nomenclature variability used by expert radiologists interpreting the same set of mammograms has not been studied. Many studies have examined radiologist agreement of mammographic interpretation, but

these did not specifically focus on expert agreement, which we would expect to be higher compared with non-expert radiologists, or gold standard development for test sets (5–10). An older study in chest radiography assessed methods for determining "truth" and found that no single method based on radiologists' interpretation was adequate (11). Agreement among experts should be examined to inform the development of standards for the number of experts to include when setting a gold standard interpretation to assess test set data.

Our objective was to develop a gold standard interpretation for mammography using an expert panel for a project that included developing tests sets to evaluate the interpretive accuracy of community radiologists. We specifically assessed interpretive agreement among three expert radiologists who independently reviewed the same digitized film screening mammograms to set the gold-standard. We examined agreement for recall, lesion location, finding type, and case difficulty. We compared the degree to which requiring agreement for two of three, versus three of three, experts affected setting the gold-standard. Given that test sets are used to measure interpretive accuracy based on a gold standard interpretation, an empirical look at levels of agreement among experts in developing the gold standard is informative to accuracy assessment.

MATERIALS AND METHODS

Protection of Study Subjects

The screening mammography films, digitized for this study, came from five mammography registries located in San Francisco, California; western Washington; North Carolina; Vermont; and New Hampshire, which are associated with the National Cancer Institute-funded Breast Cancer Surveillance Consortium (BCSC) (12,13). Each registry and the Statistical Coordinating Center (SCC), where the analysis was performed, received institutional review board (IRB) approval for either active or passive consenting processes or a waiver of consent to enroll participants, link data, and perform analytic studies. The expert radiologists consented to participate in the study. All registries follow procedures that are Health Insurance Portability and Accountability Act (HIPAA) compliant, and all registries and the SCC have received a Federal Certificate of Confidentiality (14) and other protection for the identities of women, physicians, and facilities related to the images used in this research.

Mammograms Reviewed by Experts

The test set development is described in detail elsewhere (15). Briefly, the test sets were developed for the Assessing and Improving Mammography (AIM) study in which we randomly selected films from women aged 40–69 years who received screening between 2000 and 2003. All exams were required to have comparison films from the prior 11–30 months. Films were excluded if the woman had a history of breast cancer or breast augmentation. Cancer cases included diagnoses of ductal carcinoma *in situ* (DCIS) or invasive breast cancer within 12 months from the date of the screening exam and all non-cancer cases remained cancer free for 24 months following the screening exam. A total of 314 cases (each having eight images: four current and four prior) were digitized by specialists at the American College of Radiology and placed on a DVD with software designed to record the review decisions of the expert radiologists.

Expert Review Process

Three expert radiologists participated in this study, all are senior radiologists in academic medical centers who teach and specialize in breast imaging. Each radiologist independently reviewed the DVD of 314 cases on either their personal or work computer, using computer specifications that were the same as those for radiologists participating in the test set study: a

large screen size, high-resolution graphics (1280×1024, 3 GHz, 1 GB of RAM, and a video card with 128 MB of memory capable of displaying full 32-bit color at the listed resolutions), and a DVD reader. The software let the radiologists view the craniocaudal and mediolateral oblique films in their preferred placement of left and right breasts and allowed image enlargement and comparison images. Experts were blinded to the original clinical mammography interpretations and cancer status of all mammograms.

The radiologists evaluated each mammogram for recall (yes or no), as defined by the American College of Radiology BI-RADS® atlas: recall = assessment codes 0, 4 and 5; and no recall = codes 1 and 2 (16). When they chose recall, they also gave the laterality, location, and finding type (mass, calcification, architectural distortion, or asymmetry) of the most important finding and the level of difficulty of identifying the finding (obvious, intermediate, subtle). The experts specified location of the most important finding in two ways: 1) They clicked on the lesion in the computer image of the film, the coordinates of which were stored by the computer software; and 2) they identified the quadrant where the lesion was located (upper outer, upper inner, lower outer, lower inner, or unable to determine).

After the independent expert reviews were complete, their interpretations were evaluated for agreement. Complete agreement among the initial reviews was defined as cases for which at least two of the three experts agreed on each of the indicators (recall, location, finding type, and difficulty). The experts came together for an in-person consensus meeting to discuss cases (n=68) with no initial agreement. The experts discussed these cases until they achieved consensus. The “gold-standard” interpretation for the test set was established using the majority opinion for cases with agreement of the initial review and the consensus interpretation for cases evaluated at the in-person consensus meeting. We used this approach to establish the gold standard because two readers have been shown to be superior to one in clinical practice, with either consensus by a third reader to resolve discordant interpretations or having positive interpretations overrule negative interpretations in discordant pairs. (16,17) The evidence suggests little additional value of having more than 3 readers.(18)

Analysis

Data from the experts’ independent and consensus reviews were sent to the SCC for analysis, which included description of agreement among the three experts on recall, and on location, finding type, and difficulty for the recalled cases. For cases in which a significant finding was noted and recall was recommended, lesion location was assessed for agreement by comparing click locations on the mammogram to determine whether the same lesion was indicated. Two reviewers independently evaluated agreement of the click locations as: 1) clearly indicating the same lesion (Figure 1); 2) clearly indicating different lesions (for example, clicks with different laterality) (Figure 2); or 3) agreement uncertain, further review needed (Figure 3). All films where concordance was uncertain or the experts clicked different film views (one expert clicked the CC and the other clicked the MLO) (Figure 3), were reviewed by an independent expert radiologist affiliated with the study team but who did not participate in the development of the test set. Review for lesion location agreement was necessary for 34 (22.8%) of the 149 films for which at least two experts indicated recall.

We used descriptive statistics to summarize the age, current hormone therapy use, menopausal status, and breast density of the women whose mammograms this study included. For cases where cancer was detected in the follow-up year, we described cancer characteristics including tumor type (invasive or in situ), size, nodal status, grade, and estrogen receptor (ER) and progesterone receptor (PR) status.

To describe variability of assessments between experts, we defined five successive levels of agreement: 1) *woman-level agreement* required experts to agree only on whether or not the woman should be recalled; 2) *breast-level agreement* additionally required selection of the same laterality; 3) *lesion-level agreement* further required that experts clicked on the same lesion; 4) *finding-level agreement* required that experts identified the same lesion and agreed on finding type; and 5) *complete agreement* required finding-level agreement, with additional agreement on the difficulty rating (Table 1).

To characterize pair-wise variability in expert mammography assessment, we reviewed agreement for each possible unique pair of radiologists, summarizing the proportion of mammograms where the experts agree for each of the different levels of agreement, stratified by cancer status.

We compared two methods for defining the gold-standard: 1) require all three experts to agree; or 2) use majority opinion, requiring only that any two experts agree. To describe the agreement among all three experts, we report two sets of results based on these two strategies for defining the gold-standard. For each of the five levels of agreement defined above, we report the proportion of mammograms for which all three experts agree, and the proportion for which any two agree (Table 1). Differences in findings nomenclature were summarized by the experts' finding types where there was lesion-level agreement.

The set of cases for review included 46% (N=143) with cancer and 54% (N=171) without cancer. All analyses were done using Stata/SE version 10.1 (Stata Corp. LP, College Station, Texas).

RESULTS

Of the 314 exams, 33% of films were from women ages 40–49, 41% from women ages 50–59, and 25% from women ages 60–69 years; 46% were on hormones at the time of the mammogram, and 77% were postmenopausal (Table 2). Of the 143 cancers diagnosed within a year of the screening mammogram, 81% (N=116) were invasive and 19% (N=27) DCIS. Of the 116 invasive cancers, 34% (N=37) were ≤1 cm in size, and 23% (N=72) were >2 cm; size was missing for 6% (N=7). Lymph node status was positive in 29% (N=32), grade 1 or 2 in 67% (N=66), and ER/PR status was positive/positive in 71.4% (N=60) and negative/negative in 16% (N=13) (Table 3).

Recall rates varied among the expert radiologists: 34.4% for expert 1, 51.3% for expert 2, and 65.7% for expert 3. Among the 314 exams reviewed, 24.5% (N=77) were recalled by none of the three experts, 28.0% (N=88) by one expert, 20.1% (N=63) by two experts, and 27.4% (N=86) by all three experts.

Table 4 summarizes pair-wise agreement in the expert reviews, showing the variability in agreement observed among pairs of experts. Agreement not to recall without cancer among the three unique pairings of the radiologists was 40.9%, 49.1%, and 63.7%. Pair-wise agreement to recall with cancer was 53.9%, 57.3%, and 72.7%. Overall agreement (agree to recall, or agree not to recall) at the woman level was 57.9%, 59.1%, and 70.8% for films without cancer, and 69.9%, 71.3%, and 81.8% for films with cancer. For all pairs, overall breast-level and lesion-level agreement was slightly lower than woman-level agreement. Overall complete agreement, which required lesion-level agreement plus agreement on finding type and difficulty rating, was 36.4%, 37.1%, and 42.0% for cancers and 43.9%, 53.8%, 65.6% for non-cancers. The higher level of complete agreement among non-cancer cases was due to lower recall rate among these mammograms.

Table 5 outlines agreement for the two gold-standard setting scenarios we tested (all three experts agreed, or any two of the three experts agreed). All three experts agreed to recall 75 of 143 (52.4%) cancer cases. All three also agreed on no-recall for 13 (9.1%) of the cancer cases, resulting in an overall woman-level agreement of 88/143 (61.5%) of cancers. Unlike pair-wise agreement, agreement tends to be greater for cancer cases than non-cancer cases when all three experts are required to agree.

Agreement among three experts was considerably higher if requiring only two of the three radiologists, rather than all three, to agree. Among cancer cases, agreement was 79.0% for two of three independent expert readings and 52.4% for three of three; for non-cancer cases, agreement was 21.1% for two of three experts and 6.4% for all three.

When the standard for agreement was based on at least two of three expert readings, lesion-level agreement occurred in 91.8% of non-cancer cases and 95.1% of cancer cases; and complete agreement in 86.0% of non-cancer and 73.4% of cancer cases (Table 5). This contrasted with agreement based on three of three expert reviews, in which lesion-level agreement was 42.1% for non-cancers and 55.2% for cancers, and complete agreement was 38.6% for non-cancers and 21.0% for cancer cases.

We examined the proportion of cases for which the three of three expert consensus decision agreed with the decision when only two of three agreed prior to going to consensus review. For cancer cases, the consensus (three-of-three) decision agreed with the two of three radiologists' decision in 93% of the cancer cases and 53% of the non-cancer cases at the breast level. When cancer is present (based on objective pathology), there is less variability among experts as to whether or not to recall, and thus fewer experts may suffice for setting the recall standard for cancer cases.

Table 6 shows the variability of finding types with lesion-level agreement. The distribution of finding types for cases where experts agreed on type is shown separately for lesions that were recalled by two experts and those that were recalled by three experts. Similarly, when experts disagreed on finding type, the combinations of finding types are shown by the number of experts who recalled the lesion. In 54 cases, two experts recalled the same lesion and the other expert did not; and in 74 cases, all three experts recalled the same lesion. When all three experts agreed on the lesion, they agreed on finding type for 51/74 (69%) of the identified lesions; when two of three radiologists agreed on the lesion, they agreed on the type for 38/54 (70%). For lesions recalled by all three experts who disagreed on the finding type, the disagreement was mass versus asymmetry 70% (16/23) of the time. When two radiologists recalled the lesion and disagreed on finding, the disagreement was most frequently related to characterizing the findings as a mass versus asymmetry 38% (6/16) and calcification versus architectural distortion 25% (4/16).

DISCUSSION

We assessed variability in interpretive agreement among three expert breast imagers to define a gold-standard for test sets to be used in a study of screening mammography performance. Our results showed the expected increased variability in agreement among the experts when stricter levels of agreement were required to set the gold-standard. Our results also showed that requiring two of three, rather than complete consensus when setting a gold-standard for test sets results in a notable decrease in the number of cases needing review and discussion. Complete agreement, which required lesion-level agreement and agreement on finding type and difficulty rating, was notably higher for non-cancer cases than cancer cases. Mass versus asymmetry was the leading difference in describing finding type when interpretation varied. Our results demonstrate that interpretive variation exists, even among

experts, and test set development benefits from the development of *a priori* rules to guide setting a gold-standard, such as nomenclature for benign findings. Our results suggest that at least three experts are needed for developing a gold-standard mammography interpretation, especially for non-cancer cases; however, the optimal number of experts and the proportion of agreement needed in gold-standard development are still questions requiring further research.

Variation among experts to develop a gold-standard has not been examined previously, but is important to quantify for several reasons: 1) Gold-standard interpretations have a degree of uncertainty, which is not accounted for in measuring accuracy of test-takers; 2) cancer cases can be considered to have objective 'truth', when the lesion is visible on the images, but non-cancer cases do not; and 3) optimal approaches to developing gold-standard interpretations should be identified to help standardize test set development so results are comparable across studies and establish achievable and practical quality standards for clinical practice. Careful review of the literature shows little attention to variation among experts in developing gold-standards, particularly for mammography. Several studies have examined expert interpretation variability in clinical practice. A British study examined agreement among three experienced radiologists for interpreting emergency radiographs (19). They found 51% to 74% concordance of expert pairs and concluded that this inter-observer variation should be accounted for in setting quality standards and designing assessment tools (19). Similarly, a study of six thyroid pathology experts found considerable inter-observer variability, with unanimous agreement in 13% of cases and majority agreement in 40% (20). This scant literature suggests that expert interpretive variation is higher than might be expected and should be recognized as a potential source of variability in performance studies (21,22).

Although pathology-confirmed breast cancer status is the ultimate gold-standard for mammography, having perfect mammography performance relative to true cancer status is impossible given limitations in the technology, and also undesirable in performance studies. Radiologists should not be penalized on a test set for not recalling a cancer case that is unable to be detected by mammography. Likewise, experts may agree that some findings on screening views should be recalled for further evaluation to determine if they should be biopsied – even if it is later determined that these findings are benign. That is, a proportion of non-cancer cases are "appropriate recalls", yet determining this requires a subjective assessment best developed by expert panels. Thus, for screening mammography test sets, we believe it is important to have a gold-standard interpretation for whether the exams should be recalled for additional work-up, based on expert opinion, given that in typical clinical practice some exams that ultimately are determined to not be cancer, still require workup. Our results suggest that at least three experts should be used to develop this type of gold-standard.

Our study, and that of Elsheikh et al., (20) demonstrate that requiring agreement of more experts decreases the proportion of cases in which agreement is found. A tradeoff exists: including more experts and therefore potentially better approximating "truth" yet reducing agreement, versus maximizing agreement, but including more "noise" by having fewer expert opinions. This tradeoff can be approached with various strategies. For example, test sets for the ACR's Mammography Case Review are developed by onsite consensus among six to nine experts (23).

Our study had several limitations. First, the expert radiologists were reviewing digitized images of film studies and were viewing them on personal computers. The resulting digitized images were not of diagnostic quality, and this may have increased variability among the experts. However, each was viewing the same set of digitized images; so image

quality would affect the expert radiologists equally and parallel the conditions study radiologists viewed when taking the test set. Finally, although poor quality images were excluded in our study, our study images were not selected based on whether the findings were readily visible, as is the case for CME cases designed for review on a personal computer, such as in the ACR Mammography Case Review product (23). Instead, our goal was to select a clinically representative sample of films with and without cancer, as was consistent with the overarching goals of the test set study design.

Key questions arise from our study, such as whether we should expect non-expert radiologists interpreting images in a test set to be measured against a gold-standard set by experts given even experts often do not agree. Lack of perfect concordance in judgment has been documented in most medical specialties, and thus it is reasonable to benchmark the level of agreement between experts, even if imperfect. We feel that gold-standard interpretation for mammography test sets should not be based on cancer status alone, since some cancers may not be visible on mammography and some findings that turn out to be benign did require work-up and recalling the patient was the clinically correct decision. Consensus meetings of at least three experts seem to be a good method to approximate the "truth," given interpretive variability even among experts. However, the test takers are evaluated based on their individual interpretation compared to a consensus interpretation.

Regardless of the approach used to develop a gold-standard, interpretation in clinical settings for which objective truth is more difficult raises an important issue regarding how best to evaluate non-expert physicians against a standard that may vary substantively even among experts. The results of this study demonstrate the need to rigorously address this methodological issue in mammography and other clinical arenas.

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, HHSN261201100031C). 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://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. The authors 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. We thank Rebecca Hughes for her scientific editing of the manuscript. We also 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/." Challenges in Establishing Gold Standards for Test Sets: Variation in Interpretive Agreement of Expert Mammographers

References

1. Elmore JG, Miglioretti DL, Carney PA. Does practice make perfect when interpreting mammography? Part II. JNCI. 2003; 95(4):250–252. [PubMed: 12591973]
2. Nodine CF, Kundel HL, Mello-Thomas C, Weinstein SP, Orel SG, Sullivan DC, Conant EF. How experience and training influence mammography expertise. Academic Radiology. 1999; 6(10):575–585. [PubMed: 10516859] Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. NEJM. 1994; 331:1493–9. [PubMed: 7969300]
3. Elmore JG, Jackson SL, Abraham L, Miglioretti DL, Carney PA, geller BM, et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology. 2009; 253:641–651. [PubMed: 19864507]

4. Baker JA, Kornguth PJ, Floyd CE Jr. Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description. Am J Roentgenol. 1996; 166:773–8. [PubMed: 8610547]
5. Berg WA, Campassi C, Langenberg P, Sexton MJ. Breast Imaging Reporting and Data System: Inter- and intraobserver variability in feature analysis and final assessment. Am J Roentgenol. 2000; 174:1769–77. [PubMed: 10845521]
6. Kerlikowske K, Grady D, Barclay J, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. J Nat Cancer Inst. 1998; 90:1801–9. [PubMed: 9839520]
7. Kundel HL, Polansky M. Measurement of observer agreement. Radiology. 2003; 228:303–8. [PubMed: 12819342]
8. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med. 1994; 331:1493–9. [PubMed: 7969300]
9. Ciatto S, Houssami N, Apruzzese A, Bassetti E, Brancato B, Carozzi F, et al. Reader variability in reporting breast imaging according to BI-RADS® assessment categories (the Florence experience). The Breast. 2006; 15:44–51. [PubMed: 16076556]
10. Revesz G, Kundel HL, Bonitatibus M. The effect of verification on the assessment of imaging techniques. Investigative Radiology. 1990; 25:461–464. [PubMed: 2185196]
11. Ballard-Barbash R, Taplin SH, Yankaskas BC, et al. Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database. AJR Am J Roentgenol. 1997 Oct; 169(4):1001–1008. [PubMed: 9308451]
12. National Cancer Institute. [Accessed April 14, 2009] Breast Cancer Surveillance Consortium Homepage. http://breastscreening.cancer.gov/
13. 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]
14. Carney PA, Bogart TA, Geller BM, Haneuse S, Kerlikowske K, Buist DSM, Smith R, Rosenberg RD, Yankaskas BC, Onega T, Miglioretti DL. Association Between Time Spent Interpreting, Level of Confidence, and Accuracy of Screening Mammography. AJR. (in press).
15. American College of Radiology (ACR). ACR Breast Imaging and Reporting and Data System, Breast Imaging Atlas. 4. Reston, VA: American College of Radiology; 2003. ACR BI-RADS - Mammography.
16. Hukkinen K, Kivisaari L, Vehmas T. Impact of the number of readers on mammography interpretation. Acta Radiol. 2006; 47:655–659. [PubMed: 16950700]
17. Shaw CM, Flanagan FL, Fenlon HM, McNicholas MM. Consensus review of discordant findings maximizes cancer detection rate in double-reader screening mammography: Irish National Breast Screening Program experience. Radiology. 2009; 250:354–362. [PubMed: 19188311]
18. Duijm LE, Louwman MW, Groenewoud JH, van de Poll-Franse LV, Fracheboud J, Coebergh JW. Inter-observer variability in mammography screening and effect of type and number of readers on screening outcome. British journal of cancer. 2009; 100:901–907. [PubMed: 19259088]
19. Robinson PJA, Wilson D, Coral A, Murphy A, Verow P. Variation between experienced observers in the interpretation of accident and emergency radiographs. The British Journal of Radiology. 1999; 72:323–330. [PubMed: 10474490]
20. Elsheikh TM, Asa SL, Chan JKC, DeLellis RA, Heffess CS, LiVolsi VA, Wenig BM. Interobserver and intraobserver variation among experts in the diagnosis of thyroid follicular lesions with borderline nuclear features of papillary carcinoma. Am J Clin Pathol. 2008; 130:736–744. [PubMed: 18854266]
21. Brealey S, Scally AJ. Bias in plain film reading performance studies. The British Journal of Radiology. 2001; 74:307–316. [PubMed: 11387147]
22. Brealey SD, Scally AJ, Hahn S, Godfrey C. Evidence of reference standard related bias in studies of plain radiograph reading performance: a meta-regression. The British Journal of Radiology. 2007; 80:406–413. [PubMed: 17151064]

23. Sickles EA. The American College of Radiology's mammography interpretive skills assessment (MISA) examination. Semin Breast Dis. 6:133–139.

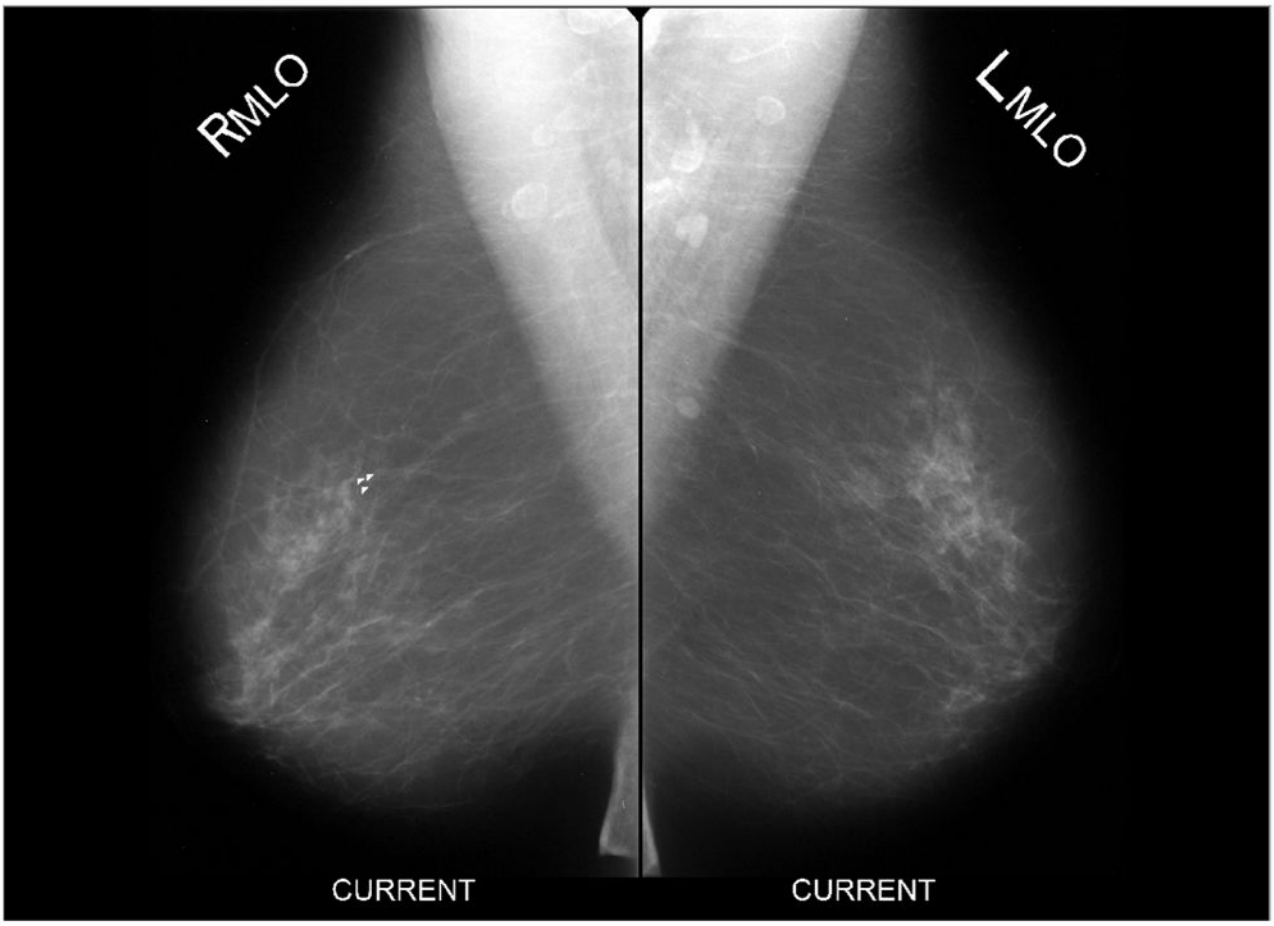

Figure 1.
Expert agreement: three experts recalled, same lesion, as indicated by the three click (▲) locations.

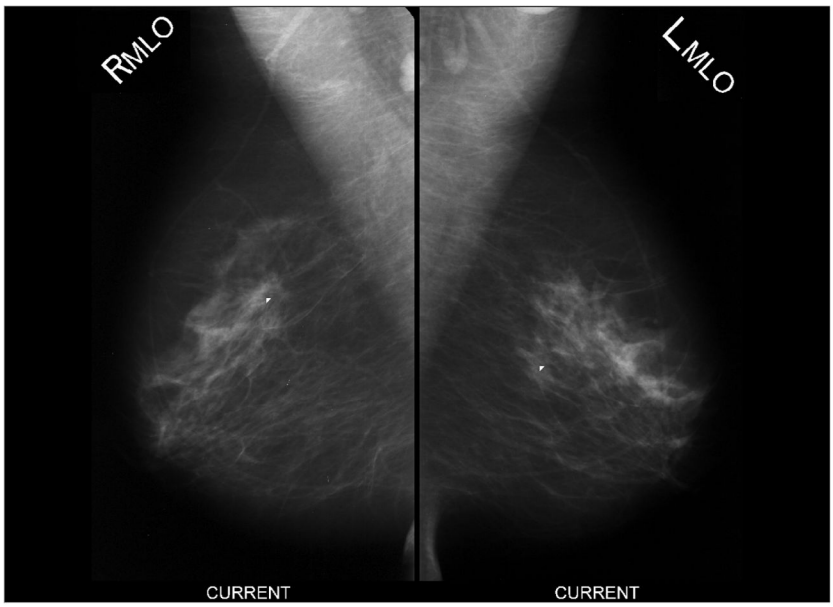

Figure 2.
No expert agreement: two experts recalled, different lesions as shown by click (▲) locations in different breasts.

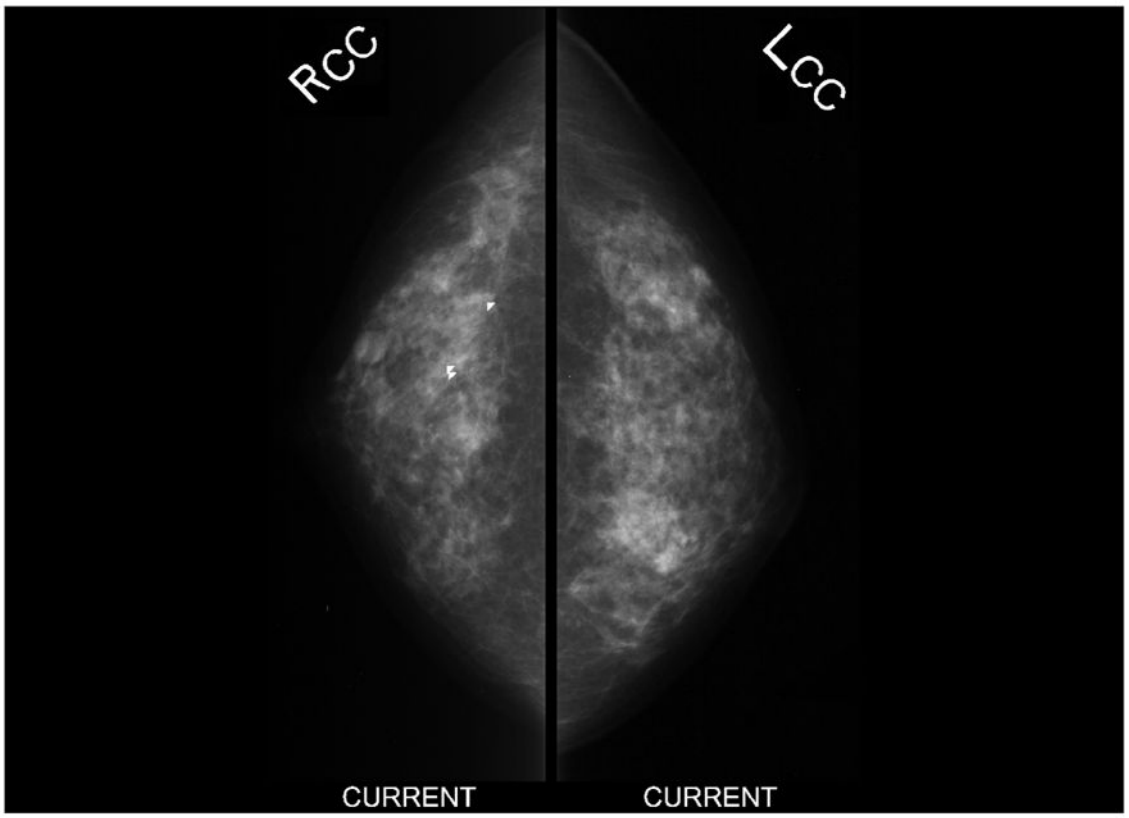

Figure 3.
Expert agreement in need of review: 3 experts recall, two agree, but unclear whether 3rd expert has indicated the same lesion, as indicated by the three click (▲) locations.

Table 1

Measurement of expert agreement and gold standard development

Five Successive Levels of Agreement	Agreement Criteria(on)
Woman-Level	Recall
Breast-Level	Recall and laterality
Lesion-Level	Recall, laterality, location
Finding-Level	Recall, laterality, location, finding type
Complete Agreement	Recall, laterality, location, finding type, difficulty
Two Methods for Establishing Gold Standard Interpretation	
- Require all three experts to agree	
- Majority opinion, requiring only two of three experts to agree	

Table 2

Characteristics of women whose mammograms were reviewed by the panel of experts

	n	%
Total	**314**	**100**
Age		
40–44	47	15
45–49	57	18.2
50–54	64	20.4
55–59	66	21
60–64	50	15.9
65–69	30	9.6
Current HT use		
No	192	63.8
Yes	109	36.2
(Missing)‡	13	(4.1)
Postmenopausal		
No	101	32.6
Yes	209	67.4
(Missing) ‡	4	(1.3)
Breast density†		
BI-RADS 1	11	4.4
BI-RADS 2	93	34.3
BI-RADS 3	140	50.1
BI-RADS 4	28	10.2
(Missing) ‡	42	(13.4)
Cancer w/in a year of screen		
No	171	54.5
Yes	143	45.5

*Cancer diagnosed within 12 months of screening mammogram

†Breast density from clinical assessment

‡Missing values not included in column percentages

Table 3

Cancer characteristics of cancer cases reviewed by the panel of experts

	n	%
Number of cancers	**143**	**100**
Cancer histologic type		
Ductal carcinoma in situ (DCIS)	27	18.9
All Invasive	116	81.1
Cancer size† (mm)		
≤ 5	13	11.9
6–10	24	22.0
11–15	25	22.9
16–20	22	20.2
> 20	25	22.9
Unknown‡	7	(6.0)
Axillary lymph node status†		
Negative	79	71.2
Positive	32	28.8
Unknown‡	5	(4.3)
Grade†		
1: Well differentiated	20	20.2
2: Moderately differentiated	46	46.5
3: Poorly differentiated	32	32.3
4: Undifferentiated	1	1.0
Unknown‡	17	(14.7)
ER/PR status†		
ER+/PR+	60	71.4
ER+/PR−	11	13.1
ER−/PR+	0	0.0
ER−/PR−	13	15.5
Unknown‡	32	(27.6)

†Invasive cancers only

‡Unknown values not included in column percentages

Table 4

Pair-wise agreement of expert reviews

	No Cancer (N=171)			Cancer (N=143)		
	Radiologist Pair			Radiologist Pair		
	(1 and 2)	(1 and 3)	(2 and 3)	(1 and 2)	(1 and 3)	(2 and 3)
	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)
Recall (Woman-level)						
Agree to recall	29 (17.0)	12 (7.0)	17 (9.9)	104 (72.7)	77 (53.9)	82 (57.3)
Agree on no-recall	70 (40.9)	109 (63.7)	84 (49.1)	13 (9.1)	25 (17.5)	18 (12.6)
Disagree on recall	72 (42.1)	50 (29.2)	70 (40.9)	26 (18.2)	41 (28.7)	43 (30.1)
Overall agreement						
Woman-level	99 (57.9)	121 (70.8)	101 (59.1)	117 (81.8)	102 (71.3)	100 (69.9)
Breast-level	96 (56.1)	121 (70.8)	99 (57.9)	110 (76.9)	95 (66.4)	96 (67.1)
Lesion-level	86 (50.3)	118 (69.0)	97 (56.7)	105 (73.4)	94 (65.7)	95 (66.4)
Finding-level	80 (46.8)	116 (67.8)	95 (55.6)	81 (56.6)	81 (56.6)	80 (55.9)
Complete	75 (43.9)	112 (65.5)	92 (53.8)	53 (37.1)	60 (42.0)	52 (36.4)

Table 5

Agreement among all 3 experts.

	All 3 agree		Majority opinion: Require any 2 of 3 agree	
	No Cancer N=171	Cancer N=143	No Cancer N=171	Cancer N=143
Recall	**N (%)**	**N (%)**	**N (%)**	**N (%)**
Agree on no-recall	64 (37.4)	13 (9.1)	135 (78.9)	30 (21.0)
Disagree on recall	96 (56.1)	55 (38.5)	N/A	N/A
Agree to recall	11 (6.4)	75 (52.4)	36 (21.1)	113 (79.0)
Overall agreement				
Woman-level	75 (43.9)	88 (61.5)	171 (100.0)	143 (100.0)
Breast-level	75 (43.9)	81 (56.6)	166 (97.1)	139 (97.2)
Lesion-level	72 (42.1)	79 (55.2)	157 (91.8)	136 (95.1)
Finding-level	70 (40.9)	58 (40.6)	151 (88.3)	126 (88.1)
Complete	66 (38.6)	30 (21.0)	147 (86.0)	105 (73.4)

Table 6

Agreement in finding type when experts recall the same lesion

	Number of experts who recalled the lesion			
	3 experts		2 of 3 experts	
Findings[†]	**N**	**%**[*]	**N**	**%**[*]
All agree	**51**	**69**	**38**	**70**
all C	21	41	13	34
all M	19	37	10	26
all AD	7	14	4	11
all AS	4	8	11	29
Disagree	**23**	**31**	**16**	**30**
C, M	1	4	2	13
C, AD	0	0	4	25
C, AS	3	13	0	0
M, AD	1	4	2	13
M, AS	16	70	6	38
AS, AD	2	9	2	13
Total	**74**	**100**	**54**	**100**

[†]C=Calcification; M=Mass; AD=Architectural distortion; AS=Asymmetry.

[*]Percentages shown combination of finding types given by the experts, separately by whether or not there was agreement.