HHS Public Access

Author manuscript

AJR Am J Roentgenol. Author manuscript; available in PMC 2016 April 01.

Published in final edited form as:

AJR Am J Roentgenol. 2015 April ; 204(4): W486–W491. doi:10.2214/AJR.13.12313.

Criteria for Identifying Radiologists with Acceptable Screening Mammography Interpretive Performance based on Multiple Performance Measures

Diana L. Miglioretti, PhD[1,2], **Laura Ichikawa, MS**[2], **Robert A. Smith, PhD**[3], **Lawrence W. Bassett, MD**[4], **Stephen A. Feig, MD**[5], **Barbara Monsees, MD**[6], **Jay R. Parikh, MD**[7], **Robert D. Rosenberg, MD**[8], **Edward A. Sickles, MD**[9], and **Patricia A. Carney, PhD**[10]

[1]Division of Biostatistics, Department of Public Health Sciences, University of California Davis School of Medicine, Davis, CA 95616

[2]Group Health Research Institute, Group Health Cooperative, Seattle, WA 98101

[3]Cancer Control Science Department, American Cancer Society, Atlanta, GA 30303

[4]Department of Radiology, University of California Los Angeles, Los Angeles, CA 90095

[5]Department of Radiological Sciences, University of California Irvine, Orange, CA 92868

[6]Department of Radiology, Washington University, St. Louis, MO 63110

[7]Swedish Breast Imaging Center, Swedish Medical Center, Seattle, WA 98104

[8]Radiology Associates of Albuquerque, Albuquerque, NM 87109

[9]Department of Radiology, University of California, San Francisco, San Francisco, CA 94143

[10]Departments of Family Medicine and Public Health and Preventive Medicine, Oregon Health & Science University, Portland, OR 97239

Abstract

Objective—Using a combination of performance measures, we updated previously proposed criteria for identifying physicians whose performance interpreting screening mammograms may indicate suboptimal interpretation skills.

Materials and Methods—In this Institutional Review Board-approved, HIPAA-compliant study, six expert breast imagers used a method based on the Angoff approach to update criteria for acceptable mammography performance on the basis of combined performance measures: (Group 1) sensitivity and specificity, for facilities with complete capture of false-negative cancers; and (Group 2) cancer detection rate (CDR), recall rate, and positive predictive value of a recall (PPV_1), for facilities that cannot capture false negatives, but have reliable cancer follow-up information for positive mammograms. Decisions were informed by normative data from the Breast Cancer Surveillance Consortium (BCSC).

Correspondence: Diana L. Miglioretti, PhD, Department of Public Health Sciences, UC Davis School of Medicine, One Shields Ave., Med Sci 1C, Room 144, Davis, CA 95616, Phone: (530) 752-7168, Fax: (530) 752-3239, dmiglioretti@ucdavis.edu.

Results—Updated, combined ranges for acceptable sensitivity and specificity of screening mammography are: (1) sensitivity ≥80% and specificity ≥85% or (2) sensitivity 75–79% and specificity 88–97%. Updated ranges for CDR, recall rate, and PPV_1 are: (1) CDR ≥6/1000, recall rate 3–20%, and any PPV_1; (2) CDR 4–6/1000, recall rate 3–15%, and PPV_1 ≥3%; or (3) CDR 2.5–4/1000, recall rate 5–12%, and PPV_1 3–8%. Using the original criteria, 51% of BCSC radiologists had acceptable sensitivity and specificity; 40% had acceptable CDR, recall rate, and PPV_1. Using the combined criteria, 69% had acceptable sensitivity and specificity and 62% had acceptable CDR, recall rate, and PPV_1.

Conclusion—The combined criteria improve previous criteria by considering the inter-relationships of multiple performance measures and broaden the acceptable performance ranges compared to previous criteria based on individual measures.

INTRODUCTION

While the benefits of screening mammography in reducing breast cancer mortality have been demonstrated in randomized trials [1–3], realizing the full potential of mammography in current clinical practice requires accurate interpretations by individual radiologists. We previously established criteria for acceptable interpretive performance of screening mammography, evaluating multiple performance measures separately: specifically sensitivity, specificity, cancer detection rate (CDR), recall rate, and positive predictive value of recall (PPV_1) [4, 5]. We suggested that a radiologist not meeting the criteria for at least one measure should be advised to examine their data in the context of their specific clinical practice and consider additional, focused continuing medical education to improve performance if appropriate [5, 6]. However, considering these performance measures together is more clinically meaningful, because they are inter-related, resulting in trade-offs between recalling patients for further work-up and detecting cancer [7]. In addition, taken alone, high performance for each measure is not valued equally, e.g., higher sensitivity is valued more than higher specificity. For example, a low recall rate (or false-positive rate) alone could raise concerns that a radiologist is not recalling enough women, but this would be acceptable for physicians with a high CDR (or sensitivity). Similarly, a higher recall (or false-positive rate) might be acceptable for a physician whose CDR (or sensitivity) is higher than average.

To update the original performance criteria, we reconvened a group of experts in breast imaging to define combined criteria for two sets of measures: 1) sensitivity and specificity for facilities with complete cancer capture through linkages with cancer registries; and 2) CDR, recall rate, and PPV_1 for facilities with reliable cancer follow-up data for positive mammograms only. Expert panels have been used for decades to establish benchmarks and guidelines for breast imaging [4–6, 8].

MATERIALS AND METHODS

The Institutional Review Board (IRB) approved all activities of this Health Insurance Portability and Accountability Act (HIPAA)-compliant study. The six breast imaging experts provided written informed consent and were selected based on their expertise in breast imaging. They ranged in age from 46 to 69 years, with a mean age of 61 years. Five

of the six breast imaging experts were males, and five worked in an academic setting. All had extensive experience in interpreting breast imaging and in breast imaging education and training. One expert was fellowship trained with 14 years of experience; the remaining 5 radiologists had 25 to 38 years of experience interpreting mammography. Their annual volume of mammography interpretation ranged from 2500 to 8000. All six experts were fellows of the Society of Breast Imaging, fellows of the American College of Radiology, and had previous experience with the modified Angoff method [5, 6]. Five of the experts have been presidents of national breast societies, and four of the experts are gold medalists of the Society of Breast Imaging. A one-day meeting was held in Seattle, Washington in September 2011, and a validation conference call was held in June 2013.

Criterion Setting Approach

We used a method based on the modified-Angoff criterion setting approach in two phases: the first phase without normative data and the second phase with normative data [9]. This is the same method we used to develop individual criteria for performance measures for both screening [5] and diagnostic mammography [6]. The Angoff method is the most widely used criterion-referenced method of standard setting, and it is often used in licensing and certification examinations in medicine because it is well supported by research evidence [10–15]. A strength of the Angoff method is that a panel of expert practicing physicians, who understand the complexities and contexts of clinical practice, work together to set the criteria.

All six breast imaging experts had been involved in earlier studies using this criterion-setting approach [5, 6], and the process was reviewed immediately before the study by a facilitator with expertise in Angoff methods. In the first phase, before being shown actual, normative performance data, the experts considered the performance of a hypothetical pool of 100 interpreting physicians. They were shown previously agreed-upon definitions for screening mammography performance based on the American College of Radiology (ACR) Breast Imaging-Reporting and Data System (BI-RADS) [16] as well as previously developed individual criteria for acceptable performance [5]. Working independently, the experts identified their ranges for “minimally acceptable” performance when considering sensitivity and specificity together. They were informed that an interpreting physician whose performance was outside these ranges would be considered for additional training.

Following a round of scoring, during the second phase of the Angoff –based method, the group viewed the criteria for acceptable performance determined individually by each expert, along with normative data showing the actual performance of BCSC radiologists and the percentage these BCSC radiologists that would meet the criteria under discussion, described in the next subsection. A facilitator with expertise in Angoff methods for setting performance criteria facilitated the review, after which the experts recast their votes for acceptable performance. The process was repeated until consensus was achieved.

We conducted a validation step on a follow-up conference call with all six experts in June 2013 to ensure final agreement for the combined criteria after re-reviewing the impact the combined cut-points would have on BCSC radiologists using more recent performance data.

The experts reviewed the data and unanimously voted not to change either set of the combined criteria.

Normative data

Experts were shown normative data on performance using a community-based sample of radiologists in the Breast Cancer Surveillance Consortium (BCSC) [17]. Each BCSC registry and the Statistical Coordinating Center received IRB approval for either active or passive consenting processes or a waiver of consent to enroll participants, link data, and perform analytic studies. All procedures were HIPAA compliant and all registries and the Statistical Coordinating Center have a Federal Certificate of Confidentiality and other protection for the identities of participating women, physicians, and facilities.

Normative data were generated in real-time during scoring reviews to illustrate the percentage of practicing radiologists who would meet the criteria under discussion. The sample included both film-screen and digital screening mammograms performed from 2000–2009 on women age 18 and older without a personal history of breast cancer. To avoid misclassifying diagnostic mammograms as screening exams, we excluded mammograms from women with a breast-imaging exam within the prior 9 months. Mammogram results were linked to a state cancer registry or regional Surveillance, Epidemiology, and End Results (SEER) program and pathology databases to determine cancer status.

For each BCSC radiologist in the sample, we examined characteristics of their patients and calculated observed performance measures from their most recent five years of BCSC data, based on the agreed upon definitions [16]. The number of years and specific years of data for each radiologist varied depending on each registry's years of participation in the BCSC and years of complete cancer ascertainment. The earliest five years for registry data were 2000–2004 and the latest were 2005–2009. To reduce the number of radiologists with zero observed "events" (e.g., no recalls, no cancers diagnosed, etc.) and to increase precision of the estimates, we restricted analysis to radiologists who contributed at least 1000 screening mammograms. In addition, inclusion in sensitivity and specificity calculations required at least 10 mammograms associated with a cancer diagnosis. We displayed the distribution of radiologists' performance measures on the receiver operating characteristic (ROC) space for sensitivity and specificity (e.g., Figure 1) and with "PPV-referral diagrams" [18], which plot PPV_1 against recall rate with CDR "isobars" (e.g., Figure 2).

Performance measures were calculated using standard BCSC definitions, which are based on ACR BI-RADS definitions [16, 19]. A mammogram was positive if it had a BI-RADS assessment of 0 (need additional imaging evaluation), 4 (suspicious abnormality), or 5 (highly suggestive of malignancy). A mammogram was negative if it had an assessment of 1 (negative) or 2 (benign finding). A BI-RADS assessment of 3 (probably benign finding) with a recommendation for immediate follow-up was recoded to 0 and considered positive. Otherwise, a BI-RADS assessment of 3 was considered negative. A woman was considered to have breast cancer if she was diagnosed with ductal carcinoma *in situ* or invasive carcinoma within one year of the mammogram and before the next screening mammogram. Analyses were performed with SAS 9.2 software (SAS Institute, Cary, NC).

RESULTS

Of the 486 BCSC radiologists included in the normative data, 13% contributed between 1000–1499 mammograms, 22% contributed 1500–2499 mammograms, 27% contributed 2500–4999 mammograms, and 38% contributed 5000 or more mammograms. Table 1 shows variability in the patient population characteristics of these radiologists. On average, 5% of mammograms were first mammograms, ranging from 0 to 12% across radiologists (median 5%); 65% of mammograms were annual screens (range 21% to 90%, median 67%). On average, 32% of mammograms were performed on women under 50 years of age (range 12% to 69%, median 31%). The racial and ethnic distributions of the radiologists' patient populations varied greatly. The original and updated criteria for sensitivity and specificity are in Table 2 and Figure 1. The original criteria required a sensitivity of at least 75% and a specificity of 88–95% for acceptable performance. Of the 350 BCSC radiologists with at least 10 mammograms associated with a cancer diagnosis, 51% met the original criteria for both sensitivity and specificity (Figure 1, Zone A).

The criteria were refined to consider sensitivity and specificity in combination using normative data from 350 BCSC radiologists who interpreted at least 10 mammograms associated with a cancer diagnosis. The experts first considered the 44 BCSC radiologists with specificity above 95%, 50% of whom met the sensitivity criterion of at least 75%. The experts agreed these individuals had acceptable performance, so they increased the upper cut point for specificity to 97%, which captured an additional 6% of BCSC radiologists (Figure 1, **Zone B**). The experts also considered radiologists with sensitivity of at least 80% and agreed that given a high sensitivity, a specificity between 97–100% was acceptable, although this captured <1% of additional BCSC radiologists (Figure 1, **Zone C**). Last, the experts agreed that a slightly lower specificity was acceptable for radiologists with a sensitivity of at least 80%, so they decreased the lower boundary for specificity to 85% in this case, capturing an additional 12% of BCSC radiologists (Figure 1, **Zone D**). Thus, the combined criteria require a specificity of at least 85% if sensitivity is 80% or higher, or a specificity of 88–97% if sensitivity is 75–80% (Table 2). In the BCSC sample, 69% of radiologists met these combined criteria, up from 51% based on the original individual-measure criteria (Table 2).

The criteria were also refined to consider for CDR, recall rate, and PPV_1 in combination using normative data from all 486 BCSC radiologists who interpreted at least 1000 mammograms (Table 3). The original criteria required a CDR of at least 2.5/1000, a recall rate of 5–12%, and a PPV_1 of 3–8%. In the BCSC, 40% of radiologists met these criteria for all three measures (Figure 2, **Zone A**). Similar to arguments considered above for sensitivity and specificity, for radiologists with a high CDR of at least 6/1000, the experts agreed a broader range of recall rates from 3–20% was acceptable, including an additional 9% of BCSC radiologists (Figure 2, **Zone B**). PPV_1 did not need to be considered for this group, because PPV_1 must be at least 3% to achieve CDR and recall rate values within the range under consideration. For a CDR of least 4/1000 but less than 6/1000, the experts agreed that a recall rate of 3–15% and PPV_1 of at least 3% is acceptable, including an additional 13% of BCSC radiologists (Figure 2, **Zone C**). Overall, 62% of the 486 BCSC radiologists met these combined criteria up from 40% based on the original criteria (Table 3).

DISCUSSION

We updated criteria for acceptable screening mammography performance by defining combined ranges that consider two or three interrelated performance measures. We developed one set of criteria based on sensitivity and specificity for facilities that have complete capture of cancer outcomes including false-negative mammograms, and a separate set of criteria based on CDR, recall rate, and PPV_1 for facilities that only have cancer outcome data for positive mammograms. We used this approach because neither sensitivity nor specificity can be accurately calculated at sites without complete cancer capture. For example, incomplete capture of false-negative cancers can occur when women are diagnosed at facilities other than the screening mammography site, which results in an overestimation of sensitivity. Few facilities outside the BCSC are currently able to link with cancer registries for complete cancer capture; thus, most are unable to evaluate sensitivity and specificity.

The ranges for acceptable specificity and recall rates changed when the experts considered these criteria together with sensitivity and CDR. In addition to identifying physicians with low specificity (<88%) or high recall rates (>12%), the original criteria identified physicians with specificities greater than 95% or recall rates lower than 5% as potentially having unacceptable performance because these physicians might be recalling too few women and thus missing cancers. The combined criteria allow a broader range of specificity values, i.e., ≥ 85% for physicians with sensitivities of at least 80%; for physicians with sensitivity between 75–80%, a specificity of 88–97% is required. Similarly, for physicians with a high CDR of at least 6/1000, the experts agreed that recall rates of 3–20% would be acceptable, and for physicians with a CDR of 4–6/1000, a recall rate of 3–15% would be acceptable. However, for those with a lower CDR of 2–4/1000, a recall rate of 5–12% would be required for acceptable combined performance, similar to the original criteria. Because finding cancers at an early stage is critical for screening mammography to reduce breast cancer mortality, the experts put less emphasis on specificity and recall rate for physicians with a high sensitivity or CDR. However, avoiding excessive rates of false positives is also regarded as an important goal of screening, so considering these measures together is important.

Applying these criteria in clinical practice presents some challenges, because the detection of breast cancer is a relatively rare event in the context of regular screening. In addition, the precision of observed performance estimates is affected by low interpretive volume. To be included in our normative data, we required BCSC radiologists to have a minimum of 1000 screening mammography interpretations, and for evaluating sensitivity and specificity, we additionally required at least 10 cancer cases. This ensured a minimum level of precision in the observed performance estimates. However, these choices were somewhat ad hoc and influenced by the overall low annual interpretive volume of the average radiologist. When applying these combined criteria in practice, especially the sensitivity and CDR criteria, judgments based on small numbers of mammograms should be made with caution. Combining audit data across multiple years will increase precision of estimated performance measures, but these estimates might not reflect current performance. We plan to address this issue in future research to develop probabilistic methods that account for the degree of

uncertainty in observed performance metrics. As a practical matter, greater exposure to a higher volume of cancer cases—either through increased interpretive volume, participation in review of all positive cases within a practice, or simulations (e.g., test sets)—could increase both the validity and reliability of measures and improve interpretative performance.

When applying these combined criteria to clinical practice, the context of the specific practice setting must also be considered. For example, breast cancer incidence is known to vary geographically in the United States [20], which would influence the CDR. Patient risk factors such as age, family history, race, and ethnicity as well as screening interval influence performance indices [21–24]. For example, CDR and sensitivity are lower in populations that are more frequently screened [24]. BCSC radiologists represent a wide range of practice types with highly variable patient populations, so these sources of variability are represented in the normative data. Subdividing audits or criteria by patient factors would create even more instability in observed performance estimates, because of low cancer incidence and interpretative volumes. While a radiologist who falls outside of the acceptable range should attempt to determine if their patient population characteristics make meeting acceptable performance criteria more challenging, it is unlikely that a unique patient mix will account for much of the performance outside of the acceptable range. However, it is also important to consider that radiologists who recently completed training often have a higher recall rate for several years before establishing a stable practice pattern [25]. Last, recent data demonstrate that improved performance appears to be related to a combination of screening and diagnostic mammography interpretation [26]. Thus, in practices with no diagnostic interpretation, optimal screening performance may be more difficult to achieve.

A strength of our study is that we developed two sets of criteria, one for facilities able to identify false negatives through linkage with cancer registries or other sources, and another for facilities that have reliable cancer follow-up data only on positive mammograms. In addition, our criteria were developed by five breast imaging experts from academic practices and one from a community-based practice using an approach based on the Angoff criterion-referenced method and informed by normative data developed from the BCSC. Our analysis was based on performance measures calculated from high quality audit data using rigorous methods, which might not be available at a typical radiology practice. A limitation of our study is that our methods do not account for variability in the observed measures due to small sample sizes and rare events.[27] We included both digital and film mammography exams in our normative data; however, most mammograms in the US are now digital. We do not think this influenced our criteria, because interpretative performance of digital mammography is similar to that of film mammography except in some subsets of women [28–30]. In addition, our combined criteria are for screening mammography only. We did not consider diagnostic mammography performance, for which previously developed criteria are based on evaluating each performance measure separately [6]. It is important to separate screening from diagnostic exams for mammography audits [16], because of differences in the patient populations and cancer prevalence, definitions used for calculating performance measures, and values for acceptable performance. However, focusing only on screening interpretation still is important because interpretation of screening is the overwhelmingly dominant clinical activity in breast imaging.

In conclusion, the combined criteria considering the inter-relationships of multiple performance measures broadens the acceptable performance ranges compared to previous criteria based on individual measures. These criteria are intended to be guidelines, i.e., a constructive tool to indicate whether a radiologist is reading at an acceptable performance level, as judged by expert opinion, or whether the radiologist should further review their individual data in the context of their specific clinical practice to determine whether further training may be warranted. Reviews of individual data should consider the demographic characteristics of the physician's patient population and be conducted with the understanding that the observed performance measures for some physicians with acceptable performance may sometimes fall outside the "acceptable" range. The use of performance criteria such as those we propose support the common goal of radiologists, patients, and society to increase the accuracy of screening mammography interpretation.

Acknowledgements

Chris Tachibana from Group Health Research Institute provided scientific editing. This work was supported the American Cancer Society and made possible by a generous donation from the Longaberger Company's Horizon of Hope® Campaign (SIRSG-11-249-01 and SIRSG-11-247-01). Collection of mammography performance data was supported by the 3 National Cancer Institute Breast Cancer Surveillance Consortium (BCSC; P01CA154292 and 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 U.S.; For a full description of these sources, please see: http://breastscreening.cancer.gov/work/acknowledgement.html. The authors had full responsibility in the design of the study, the collection of the data, the analysis and interpretation of the data, the decision to submit the manuscript for publication, and the writing of the manuscript. 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/.

REFERENCES

1. Smith RA, Duffy SW, Tabar L. Breast cancer screening: the evolving evidence. Oncology (Williston Park, NY). 2012; 26:471–475. 479–481, 485–476.
2. Nelson HD, Tyne K, Naik A, Bougatsos C, Chan BK, Humphrey L. Screening for breast cancer: an update for the U.S. Preveventive Services Task Force. Ann Intern Med. 2009; 151:727–737. [PubMed: 19920273]
3. Humphrey LL, Helfand M, Chan BK, Woolf SH. Breast cancer screening: a summary of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med. 2002; 137:347–360. [PubMed: 12204020]
4. Bassett, LW.; Hendrick, RE.; Bassford, TL., et al. Quality Determinants of Mammography. Clinical Practice Guideline No. 13, AHCPR Publication No. 95-0632 ed. Rockville, MD: Agency for Health Care Policy and Research, Public Health Service, U.S. Department of Health and Human Services; 1994.
5. Carney PA, Sickles EA, Monsees BS, et al. Identifying minimally acceptable interpretive performance criteria for screening mammography. Radiology. 2010; 255:354–361. [PubMed: 20413750]
6. Carney PA, Parikh J, Sickles EA, et al. Diagnostic Mammography: Identifying Minimally Acceptable Interpretive Performance Criteria. Radiology. 2013; 267:359–367. [PubMed: 23297329]
7. Doyle GP, Onysko J, Pogany L, et al. Limitations of minimally acceptable interpretive performance criteria for screening mammography. Radiology. 2011; 258:960–961. [PubMed: 21339358]
8. Saslow D, Boetes C, Burke W, et al. American Cancer Society guidelines for breast screening with MRI as an adjunct to mammography. CA Cancer J Clin. 2007; 57:75–89. [PubMed: 17392385]
9. Ricker KL. Setting Cut-Scores: A Critical Review of the Angoff and Modified Angoff Methods. The Alberta Journal of Educational Research. 2006; 52:53–64.

10. Boursicot K, Roberts T. Setting standards in a professional higher education course: defining the concept of the minimally competent student in performance based assessment at the level of graduation from medical school. Higher Educ Q. 2006; 60:74–90.
11. Talente G, Haist SA, Wilson JF. A model for setting performance standards for standardized patient examinations. Eval Health Prof. 2003; 26:427–446. [PubMed: 14631613]
12. Alston GL, Haltom WR. Reliability of a minimal competency score for an annual skills mastery assessment. American journal of pharmaceutical education. 2013; 77:211. [PubMed: 24371335]
13. Ben-David M. AMEE Guide No. 18: Standard setting in student assessment. Medical Teacher. 2000; 22:120–130.
14. Jackson, N.; Jamieson, A.; Khan, Ae. Assessment in Medical Education and Training: A Practical Guide. Radcliffe Publishing; 2007.
15. Norcini JJ. Setting standards on educational tests. Med Educ. 2003; 37:464–469. [PubMed: 12709190]
16. American College of Radiology. American College of Radiology (ACR) Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas). 4th edition ed. Reston, VA: Am Coll Radiol; 2003.
17. 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; 169:1001–1008. [PubMed: 9308451]
18. Blanks RG, Moss SM, Wallis MG. Monitoring and evaluating the UK National Health Service Breast Screening Programme: evaluating the variation in radiological performance between individual programmes using PPV-referral diagrams. J Med Screen. 2001; 8:24–28. [PubMed: 11373846]
19. National Cancer Institute. Breast Cancer Surveillance Consortium, BCSC Glossary of Terms. BCSC; 2010.
20. American Cancer Society. Cancer Facts & Figures 2013. Atlanta: American Cancer Society; 2013.
21. Carney PA, Miglioretti DL, Yankaskas BC, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. Ann Intern Med. 2003; 138:168–175. [PubMed: 12558355]
22. Kerlikowske K, Carney PA, Geller B, et al. Performance of screening mammography among women with and without a first-degree relative with breast cancer. Ann Intern Med. 2000; 133:855–863. [PubMed: 11103055]
23. Cook AJ, Elmore JG, Miglioretti DL, et al. Decreased accuracy in interpretation of community-based screening mammography for women with multiple clinical risk factors. J Clin Epidemiol. 2010; 63:441–451. [PubMed: 19744825]
24. Yankaskas BC, Taplin SH, Ichikawa L, et al. Association between mammography timing and measures of screening performance in the United States. Radiology. 2005; 234:363–373. [PubMed: 15670994]
25. Miglioretti DL, Gard CC, Carney PA, et al. When radiologists perform best: The learning curve in screening mammography interpretation. Radiology. 2009; 253:632–640. [PubMed: 19789234]
26. Buist DS, Anderson ML, Haneuse SJ, et al. Influence of annual interpretive volume on screening mammography performance in the United States. Radiology. 2011; 259:72–84. [PubMed: 21343539]
27. Burnside ES, Lin Y, Munoz Del Rio A, et al. Addressing the challenge of assessing physician-level screening performance: mammography as an example. PLoS One. 2014; 9:e89418. [PubMed: 24586763]
28. Kerlikowske K, Hubbard RA, Miglioretti DL, et al. Comparative effectiveness of digital versus film-screen mammography in community practice in the United States: A cohort study. Ann Intern Med. 2011; 155:493–502. [PubMed: 22007043]
29. Pisano ED, Hendrick RE, Yaffe MJ, et al. Diagnostic accuracy of digital versus film mammography: exploratory analysis of selected population subgroups in DMIST. Radiology. 2008; 246:376–383. [PubMed: 18227537]
30. Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. N Engl J Med. 2005; 353:1773–1783. [PubMed: 16169887]

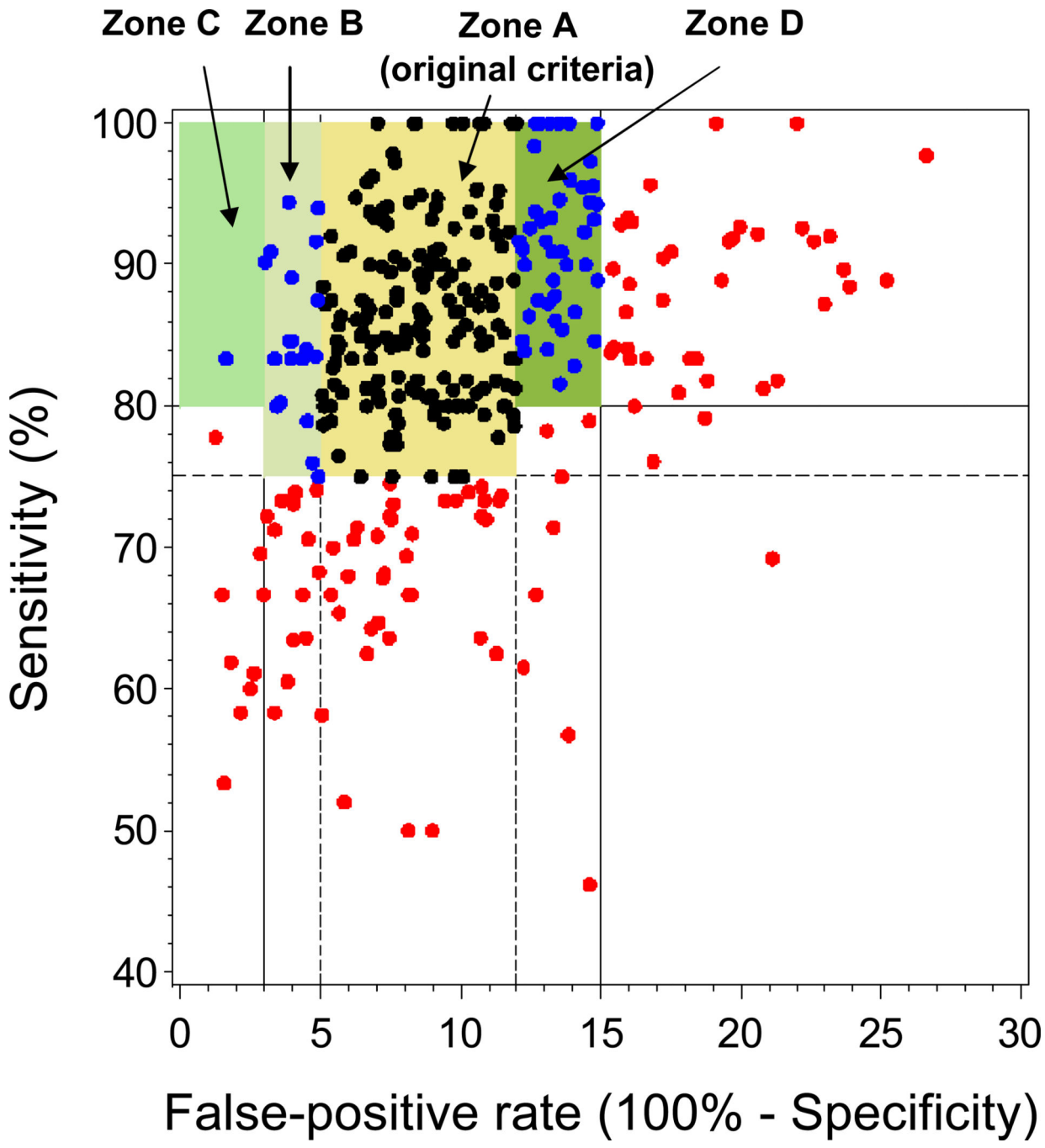

Figure 1.

Sensitivity versus false-positive rate for original criteria based on separate evaluation of these measures and for the updated criteria evaluating these measures in combination. Dots represent observed values for 350 BCSC radiologists with at least 1000 screening mammograms and at least 10 screening mammograms associated with cancer: black, acceptable under original and updated criteria; blue, acceptable under updated criteria but not original criteria; red, acceptable under neither. Zone A, original criteria for acceptable performance. Zones B, C, and D, areas of acceptable performance added with updated

criteria. Zone B, upper specificity threshold increased to 97%. Zone C, upper specificity threshold increased to 100% for radiologists with at least 80% sensitivity. Zone D, lower specificity threshold decreased to 85% for radiologists with at least 80% sensitivity. Note: the y-axis (sensitivity) starts at 40% instead of 0%.

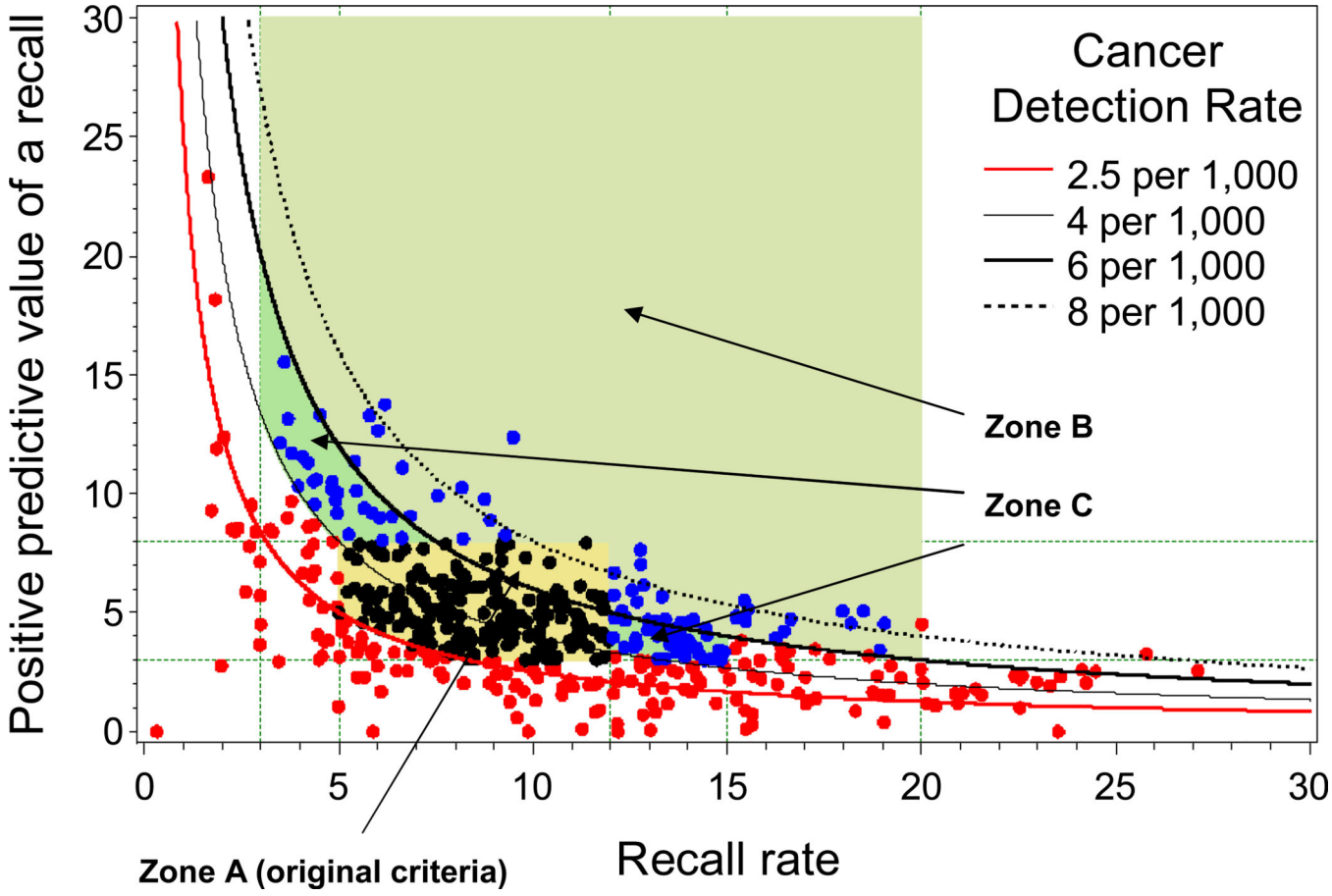

Figure 2.

Cancer detection rate (CDR), recall rate, and positive predictive value of a recall (PPV_1) for original criteria based on separate evaluation of these measures and for the updated criteria based on evaluating these measures in combination. Dots represent observed values for 486 BCSC radiologists with at least 1000 screening mammograms: black, acceptable under original and updated criteria; blue acceptable under updated criteria but not original criteria; red, acceptable under neither. Zone A, acceptable under original criteria. Zones B and C, areas of acceptable performance added with updated criteria. Zone B, 3–20% recall rate for radiologists with CDR at least 6/1000. Zone C, 3–15% recall rate and PPV_1 of at least 3% for radiologists with CDR at least 4/1000 but less than 6/1000.

Table 1

Characteristics of patient populations of 486 BCSC radiologists* contributing normative screening mammography data

Patient characteristics	Average percentage	Range of percentages across BCSC radiologists
First mammograms	5%	0% – 12%
Annual mammograms†	65%	21% – 90%
Biennial mammograms§	18%	5% – 67%
Age <50 years	32%	12% – 68%
White, non-Hispanic	76%	0% – 99%
Black, non-Hispanic	6%	0% – 53%
Hispanic	8%	<1% – 78%
Asian	8%	1% – 99%
Family history of breast cancer	15%	4% – 23%

BCSC, Breast Cancer Surveillance Consortium.

*Restricted to radiologists with <10% missing data for each patient characteristic.

†Previous mammogram in the last 9–18 months.

§Previous mammogram in the last 19–30 months.

Table 2

Original criteria and updated, combined criteria for identifying radiologists with acceptable sensitivity and specificity

Criteria		Sensitivity		Specificity	Percentage of BCSC radiologists who met criteria (*N*=350)
Original		≥75%		88–95%	51%
Updated		≥80%	and	≥85%	62%
	or	75–80%	and	88–97%	7%

BCSC, Breast Cancer Surveillance Consortium

Table 3

Original criteria and updated, combined criteria for identifying radiologists with acceptable cancer detection rate, recall rate, and positive predictive value of a recall

Criteria		Cancer detection rate		Recall rate		Positive predictive value of a recall	Percentage of BCSC radiologists who met criteria (*N*=486)
Original		≥2.5/1000		5–12%		3–8%	40%
Updated		≥6/1000	and	3–20%			13%
	or	≥4–6/1000	and	3–15%	and	≥3%	31%
	or	2.5–4/1000	and	5–12%	and	3–8%	18%

BCSC, Breast Cancer Surveillance Consortium