UC Irvine

UC Irvine Previously Published Works

Title

Diagnostic Mammography: Identifying Minimally Acceptable Interpretive Performance Criteria

Permalink

https://escholarship.org/uc/item/1z81f4f2

Journal

Radiology, 267(2)

ISSN

0033-8419

Authors

Carney, Patricia A
Parikh, Jay
Sickles, Edward A
et al.

Publication Date

2013-05-01

DOI

10.1148/radiol.12121216

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Note: This copy is for your personal non-commercial use only. To order presentation-ready copies for distribution to your colleagues or clients, contact us at www.rsna.org/rsnarights.

Diagnostic Mammography: Identifying Minimally Acceptable Interpretive Performance Criteria[1]

Patricia A. Carney, PhD
Jay Parikh, MD
Edward A. Sickles, MD
Stephen A. Feig, MD
Barbara Monsees, MD
Lawrence W. Bassett, MD
Robert A. Smith, PhD
Robert Rosenberg, MD
Laura Ichikawa, MS
James Wallace, BA
Khai Tran, MD
Diana L. Miglioretti, PhD

Purpose: To develop criteria to identify thresholds for the minimally acceptable performance of physicians interpreting diagnostic mammography studies.

Materials and Methods: In an institutional review board–approved HIPAA–compliant study, an Angoff approach was used to set criteria for identifying minimally acceptable interpretive performance for both workup after abnormal screening examinations and workup of a breast lump. Normative data from the Breast Cancer Surveillance Consortium (BCSC) was used to help the expert radiologist identify the impact of cut points. Simulations, also using data from the BCSC, were used to estimate the expected clinical impact from the recommended performance thresholds.

Results: Final cut points for workup of abnormal screening examinations were as follows: sensitivity, less than 80%; specificity, less than 80% or greater than 95%; abnormal interpretation rate, less than 8% or greater than 25%; positive predictive value (PPV) of biopsy recommendation (PPV_2), less than 15% or greater than 40%; PPV of biopsy performed (PPV_3), less than 20% or greater than 45%; and cancer diagnosis rate, less than 20 per 1000 interpretations. Final cut points for workup of a breast lump were as follows: sensitivity, less than 85%; specificity, less than 83% or greater than 95%; abnormal interpretation rate, less than 10% or greater than 25%; PPV_2, less than 25% or greater than 50%; PPV_3, less than 30% or greater than 55%; and cancer diagnosis rate, less than 40 per 1000 interpretations. If underperforming physicians moved into the acceptable range after remedial training, the expected result would be *(a)* diagnosis of an additional 86 cancers per 100 000 women undergoing workup after screening examinations, with a reduction in the number of false-positive examinations by 1067 per 100 000 women undergoing this workup, and *(b)* diagnosis of an additional 335 cancers per 100 000 women undergoing workup of a breast lump, with a reduction in the number of false-positive examinations by 634 per 100 000 women undergoing this workup.

Conclusion: Interpreting physicians who fall outside one or more of the identified cut points should be reviewed in the context of an overall assessment of all their performance measures and their specific practice setting to determine if remedial training is indicated.

©RSNA, 2013

[1] From the Departments of Family Medicine and Public Health and Preventive Medicine, Oregon Health and Science University, 3181 SW Sam Jackson Park Rd, Portland, OR 97239-3098 (P.A.C., J.W.); Swedish Breast Imaging Center, Swedish Medical Center, Seattle, Wash (J.P.); Department of Radiology, University of California–San Francisco, San Francisco, Calif (E.A.S.); Department of Radiological Sciences, University of California–Irvine, Irvine, Calif (S.A.F.); Mallinckrodt Institute of Radiology, Washington University School of Medicine, St Louis, Mo (B.M.); Department of Radiology, University of California, Los Angeles, Los Angeles, Calif (L.W.B.); Cancer Control Science Department, American Cancer Society, Atlanta, Ga (R.A.S.); Radiology Associates of Albuquerque, Albuquerque, NM (R.R.); Biostatistics Unit, Group Health Research Institute, Seattle, Wash (L.I., D.L.M.); Tacoma Radiology Associates in Medical Imaging, Carol Milgard Breast Center, Tacoma, Wash (K.T.); and Department of Biostatistics, University of Washington, Seattle, Wash (D.L.M.). Received June 13, 2012; revision requested July 17; revision received October 2; final version accepted October 19. **Address correspondence to** P.A.C. (e-mail: *carneyp@ohsu.edu*).

© RSNA, 2013

Diagnostic mammography is used to work up patients with abnormal findings on screening mammograms and to evaluate patients with either self-detected or clinically detected breast abnormalities (1). The Mammography Quality Standards Act (2) requires each breast imaging facility in the United States to establish a system to record medical outcomes data. The current requirements are limited to correlations between biopsies recommended after mammography and pathology outcomes from biopsy but do not discriminate between screening and diagnostic mammography. Substantially different benchmarks for performance of screening (3) and diagnostic (4) mammography have been published, and several studies have provided auditing outcomes for diagnostic mammography (5–8). Diagnostic mammography audits demonstrate higher sensitivity and lower specificity compared with screening performance measures (9), as well as higher abnormal interpretation rates, higher positive predictive values (PPVs), and higher cancer diagnosis rates (7). Combining both screening and diagnostic mammography audits concurrently requires mathematic extrapolation to assure the integrity of the audit (10). Proficiency in screening mammography does not necessarily equate to proficiency in diagnostic mammography (11).

Thresholds to identify minimally acceptable interpretive performance in terms of sensitivity, specificity, abnormal interpretation rate, PPVs, and cancer diagnosis rate have been published for screening mammography (12), but these have not yet been established for diagnostic mammography. Identifying low performers who might benefit from additional training should lead to more accurate and cost-effective diagnostic mammography. Our purpose was to develop criteria to identify thresholds for the minimally acceptable performance of physicians interpreting diagnostic mammography studies.

Advances in Knowledge

- Final cut points for workup of abnormal mammographic screening examinations were as follows: sensitivity, less than 80%; specificity, less than 80% or greater than 95%; abnormal interpretation rate, less than 8% or greater than 25%; positive predictive value (PPV) of biopsy recommendation (PPV_2), less than 15% or greater than 40%; PPV of biopsy performed (PPV_3), less than 20% or greater than 45%; and cancer diagnosis rate, less than 20 per 1000 interpretations.
- Final cut points for workup of a breast lump were as follows: sensitivity, less than 85%; specificity, less than 83% or greater than 95%; abnormal interpretation rate, less than 10% or greater than 25%; PPV_2, less than 25% or greater than 50%; PPV_3, less than 30% or greater than 55%; and cancer diagnosis rate, less than 40 per 1000 interpretations.

Materials and Methods

Recruitment of Expert Radiologists

The institutional review board at Oregon Health and Science University approved all study activities. Selection criteria for expert breast imaging radiologists included the following: *(a)* That they had interpreted mammograms for at least 10 years, *(b)* that they had devoted 75% or more of their practice to breast imaging, and *(c)* that they had either more than 15 years of experience interpreting mammograms or had completed fellowship training in breast imaging. We used professional contacts associated with the Breast Cancer Surveillance Consortium (BCSC) (13) to identify 11 radiologists who met the eligibility criteria and who were able to attend a 1-day meeting in September 2011, which was held in Seattle, Washington. Expert radiologists reviewed and signed consent forms and completed a brief survey of their demographic and practice characteristics at the 1-day meeting. Seven of the 11 expert radiologists participated in our previous study (12) in which we set criteria for low performance in interpreting screening mammography studies. Two radiologists were members of the BCSC, and none were in practice together.

Implication for Patient Care

- If underperforming physicians moved into the acceptable range after remedial training, we would expect *(a)* diagnosis of an additional 86 cancers per 100000 women undergoing workup after screening examinations, with a reduction in the number of false-positive examinations by 1067 per 100000 women undergoing this workup, and *(b)* diagnosis of an additional 335 cancers per 100000 women undergoing workup of a breast lump, with a reduction in the number of false-positive examinations by 634 per 100000 women undergoing this workup.

Published online before print
10.1148/radiol.12121216 **Content code:** BR

Radiology 2013; 267:359–367

Abbreviations:
BCSC = Breast Cancer Surveillance Consortium
BI-RADS = Breast Imaging Reporting and Data System
PPV = positive predictive value
PPV_2 = PPV of biopsy recommendation
PPV_3 = PPV of biopsy performed

Author contributions:
Guarantor of integrity of entire study, P.A.C.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; manuscript final version approval, all authors; literature research, P.A.C., J.P., S.A.F., R.A.S., J.W.; clinical studies, J.P., S.A.F., L.W.B., R.R.; experimental studies, J.P., E.A.S., B.M., L.W.B., R.R.; statistical analysis, P.A.C., L.I., D.L.M.; and manuscript editing, all authors

Funding:
This research was supported by the National Cancer Institute Breast Cancer Surveillance Consortium (grants U01CA63740, U01CA86076, U01CA86082, U01CA63736, U01CA70013, U01CA69976, U01CA63731, U01CA70040, and HHSN261201100031C).

Conflicts of interest are listed at the end of this article.

Modified Angoff Criterion-referenced Approach

As we did when developing criteria for screening mammography (12), we used a modified Angoff approach in two phases (14–16) for diagnostic mammography performed for both workup of abnormalities found at screening examinations and workup of breast lumps. The Angoff method is a criterion-referenced method of standard setting (17,18). It is the most commonly used approach in licensing and certification examinations in medicine (17,18), and intraclass correlation coefficients have been high, at 0.81 and 0.82 (19), illustrating its robustness as a criterion-setting method. We used a modified approach, insofar as we presented normative data during the decision process of identifying performance cut points (14) for both types of diagnostic mammography, which were considered separately. The experts agreed on standard definitions for these performance measures before scoring began (Table 1). In phase I, the group of 11 expert radiologists considered the interpretive performance of a hypothetical pool of 100 interpreting physicians. Working independently, experts conveyed their cut points for achieving "minimally acceptable" performance for each measure (sensitivity, specificity, abnormal interpretation rate, PPV of biopsy recommendation [PPV_2], PPV of biopsy performed [PPV_3], and cancer diagnosis rate) by anonymously providing scores, which were immediately tallied and summarized according to the mean, median, mode, and range and were then presented to the group. The radiologists were informed that performance that fell outside the cut points would result in that interpreting physician being considered for recommendation for additional training.

A nonradiologist facilitator with relevant expertise using the Angoff approach (P.A.C.) facilitated the discussion, after which votes for minimally acceptable performance were recast. This was repeated until agreement on cut points for the hypothetical group of interpreting physicians was achieved. Each performance indicator was considered separately for phase I and then phase II before the next indicator was scored. Sensitivity and cancer diagnosis rates both involved deriving one cut point (with low performers being below it), while specificity, abnormal interpretation rate, PPV_2, and PPV_3 all involved setting upper and lower bounds, where low performance would be beyond the range between the lower bound and the upper bound cut points.

Presentation of normative data has been used about 25% of the time when Angoff methods are applied and has been shown to improve interexpert reliability (14). In phase II, the experts were shown normative data on performance from a community-based sample of interpreting physician participants of the BCSC (13). This allowed us to illustrate the impact proposed cut points might have on mammography practice (eg, how many interpreting physicians might be considered for additional training). The process of scoring the cut points was repeated in phase II as was done in phase I, with the mean, median, mode, and range presented at each round of scoring until consensus was achieved, which occurred in four rounds.

To calculate the normative statistics, we used data from the BCSC. Each BCSC registry and the Statistical Coordinating Center (at the Group Health Research Institute, Seattle, Wash) have received institutional review board approval for either active or passive consenting processes or a waiver of consent to enroll participants, link data, and perform analytic studies. All procedures are Health Insurance Portability and Accountability Act–compliant, and all registries and the Statistical Coordinating Center have received a Federal Certificate of Confidentiality and other protection for the identities of the women, physicians, and facilities that are the subjects of this research. Data from the BCSC have contributed to more than 400 publications (*http://breastscreening.cancer.gov*

Table 1

Diagnostic Mammography Definitions Used for Angoff Scoring Criteria

Examination or Performance Measure	Definition
Diagnostic mammography A	Unilateral or bilateral mammography performed in women for workup of a prior abnormal screening mammography result
Diagnostic mammography B	Unilateral or bilateral mammography performed in women for workup of a breast lump
Sensitivity	Ability of a test to find a cancer when it is present [TP/(TP + FN)]
Specificity	Ability of a test to determine that cancer is absent when a patient is cancer-free [TN/(TN + FP)]
Abnormal interpretation rate	Proportion of diagnostic mammography studies given a positive final assessment (BI-RADS category 4 or 5) [(TP + FP)/(TP + FP + TN + FN)
PPV_2	Proportion of diagnostic mammography studies with a recommendation for biopsy (BI-RADS category 4 or 5) resulting in a diagnosis of breast cancer [TP/(TP + FP)]
PPV_3	Proportion of diagnostic mammography studies with a biopsy performed (BI-RADS category 4 or 5) resulting in a diagnosis of breast cancer [TP/(TP + FP)]
Cancer diagnosis rate	Number of diagnostic mammography studies with a positive final assessment and diagnosis of breast cancer per 1000 mammograms

Note.— BI-RADS = Breast Imaging Reporting and Data System (20). FN = false-negative on the basis of final assessment at the end of imaging workup, where BI-RADS categories 1, 2, and 3 are considered negative and cancer is found within 365 days. FP = false-positive on the basis of final assessment at the end of imaging workup, where BI-RADS categories 4 and 5 are considered positive (unresolved 0s are considered missing) and no cancer is found within 365 days. TN = true-negative on the basis of final assessment at the end of imaging workup, where BI-RADS categories 1, 2, and 3 are considered negative and no cancer is found within 365 days. TP = true-positive on the basis of final assessment at the end of imaging workup, where BI-RADS categories 4 and 5 are considered positive (unresolved 0s are considered missing) and cancer is found within 365 days.

/publications/search.html). The specific data in this analysis will partially overlap with those in some of these publications.

For normative data, we included diagnostic mammography studies that were performed either for workup after an abnormal screening examination or for evaluation of a breast lump and that were interpreted at a BCSC facility from 2003 to 2007. Mammograms were linked to cancer registries and pathology databases to determine cancer status (ductal carcinoma in situ or invasive carcinoma) within 1 year of the mammogram. We calculated the percentile distributions across radiologists for each performance measure. We restricted the analysis to those who contributed at least a subjectively determined minimum number of cases for each performance measure. For sensitivity, we included those who had interpreted a minimum of 10 mammographic studies associated with a cancer diagnosis. Radiologists who had interpreted at least 100 mammographic studies that did not show cancer contributed to the analysis of specificity. To contribute to the analysis of PPV_2, radiologists needed to have recalled patients after a positive result for at least 10 diagnostic mammography studies, and for $PPV_{3,}$ radiologists needed to have had biopsies performed after a positive mammogram for at least 10 studies. Radiologists needed to have interpreted at least 10 diagnostic mammography studies for abnormal interpretation rate and at least 100 diagnostic mammography studies for cancer diagnosis rate. We displayed the frequency distributions overlaid with percentile values to display these data in an easily understandable format.

We also used BCSC data to determine the percentage of radiologists and facilities that would be affected by the cut points identified. Here, the BCSC data were similar to those used to calculate normative statistics, except the most recent 5 years of data (which varied according to registry) were used instead of 2003–2007 data for all registries. This was done to include 5 years of data for each registry where available data varied by cancer registry ascertainment and years of participation in the BCSC. The earliest 5 years for a registry were 2000–2004, and the latest 5 years were 2004–2008. For some measures, we also required higher volume criteria for a radiologist than the normative statistics to provide stable results. The minimum number of diagnostic mammography studies after which patients were recalled for PPV_2 increased from 10 to 30. For PPV_3, the minimum number of biopsies after a positive mammography study increased from 10 to 30 for workup of an abnormal screening examination and from 10 to 20 biopsies for evaluation of a breast lump. The number of diagnostic mammography studies required for abnormal interpretation rate increased from 10 to 100. These were calculated separately for each specific cut point.

Determining Interrelationships among Performance Measures

The performance measures we examined are interrelated, and these interrelationships must be considered when setting cut points. For example, the cut points for abnormal interpretation rate and specificity and were determined together, because these measures are very closely related, as the majority of women undergoing diagnostic mammography do not turn out to have cancer. The difference between an interpreting physician's false-positive rate (1 − specificity) and abnormal interpretation rate is bounded by the cancer rate, which is relatively low, even in a population undergoing diagnostic mammography (approximately 30 cancers per 1000 examinations). Similarly, a given abnormal interpretation rate and PPV result in a specific cancer diagnosis rate, so the cut points need to be considered together.

Simulation Analysis

Using BCSC normative data, we conducted a statistical simulation to examine the impact of moving lower-performing physicians' performance measures into the acceptable range. We created a simulated cohort of 1 million women undergoing diagnostic mammography for workup of an abnormal screening examination and a cancer status for each woman based on a prevalence of 46 cancers per 1000 women (3,4). Similarly, we created a simulated cohort of 1 million women undergoing diagnostic mammography for workup of a breast lump with cancer status based on a prevalence of 67 cancers per 1000 women. For each simulated woman, we chose one of the actual BCSC study interpreting physicians, with the selection probability proportional to the physician's interpretive volume, and then randomly generated a mammographic result given the simulated woman's cancer status and the physician's own observed diagnostic performance measures. We also simultaneously generated a second mammographic result for each woman associated with the same chosen interpreter to simulate the retraining of low performers. If the interpreter's observed performance measure was in the acceptable range, then the second mammographic result was identical to the first. If not, the relevant performance measure was replaced with a value from a randomly chosen interpreter with an observed value in the acceptable range. We tabled both test results against cancer status for our simulated cohorts and compared the number of true-positive and false-positive tests. The simulation was performed by using the R statistical software package (21).

Results

The demographic and practice characteristics of the 11 breast imaging experts are shown in Table 2. The mean age of the expert group was 55 years, and 55% of the experts were men. They worked in relatively large medical practices and spent, on average, 95% of their time in breast imaging. Forty-five percent were fellowship trained. Importantly, equal percentages of experts represented academic and community-based practices.

The number of rounds of scoring needed to come to agreement ranged from two to three in phase I and was one or two in phase II. Summary scores for each of the two phases are illustrated in Table 3. The ranges of cut point scores in phase I were lower and generally wider than in phase II. The ranges were highest for sensitivity, specificity, and PPV and were lowest for abnormal interpretation and cancer diagnosis rates in phase I and phase II.

Table 4 illustrates normative performance data for diagnostic mammography performed for workup after abnormal screening examinations and that performed to evaluate breast lumps for between 91 and 459 radiologist participants in the BCSC who interpreted diagnostic mammography studies between 2003 and 2007 and who met or exceeded case volume criteria for that particular performance measure. Benchmarks for abnormal interpretation rate are based on 119 851 diagnostic mammography studies performed for workup after an abnormal screening examination and 46 682 mammography studies performed for evaluation of a breast lump among radiologists who interpreted at least 10 mammography studies. Benchmarks for sensitivity are based on 4110 cancers diagnosed within 1 year of diagnostic mammography performed for workup after an abnormal screening examination and 2258 cancers diagnosed within 1 year after mammography performed for evaluation of a breast lump.

For diagnostic mammography performed for workup after abnormal screening examinations, normative data for performance measures of sensitivity ranged from 30% to 100%, with 90% as the median. Specificity ranged from 66.4% to 98.4%, with a median of 89.7%. Abnormal interpretation rate ranged from 0% to 63.6%, with a median of 13.7%. For diagnostic mammography performed for workup of a breast lump, normative data for the performance measure of sensitivity ranged from 46.2% to 100%, with 90% as the median. Specificity ranged from 66.8% to 98.7%, with a median of 90.1%. Abnormal interpretation rate ranged from 0% to 50%, with a median of 15.3%.

Table 5 lists the final cut points for low performance derived from the study's two phases, which included, for diagnostic mammography performed for workup after abnormal screening examinations, sensitivity less than 80%, specificity less than 80% or greater than 95%, abnormal interpretation rate less than 8% or greater than 25%, PPV_2 less than 15% or greater than 40%, PPV_3 less than 20% or greater than 45%, and cancer diagnosis rate less than 20 per 1000 interpretations. Final cut points to identify low performance for diagnostic mammography performed for workup of a breast lump were sensitivity less than 85%, specificity less than 83% or greater than 95%, abnormal interpretation rate less than 10% or greater than 25%, PPV_2 less than 25% or greater than 50%, PPV_3 less than 30% or greater than 55%, and cancer diagnosis rate less than 40 per 1000 interpretations. The selected cut points for performance measures would likely result in 16%–34% of interpreting physicians and 11%–24% of facilities being considered for additional training in diagnostic mammography following abnormal screening examinations and 21%–42% of radiologists and 14%–54% of facilities being considered for additional training in diagnostic mammography performed to evaluate a breast lump.

Last, in our simulated cohort of 1 million women undergoing diagnostic mammography for workup after an abnormal screening examination, 45 439 had breast cancer. The number of cancers we estimated that would be correctly recalled for biopsy increased from 40 772 to 41 636 in the simulated cohort as a result of the effect of additional training for low-performing interpreters that improved their performance to acceptable levels; likewise, false-positive biopsy recommendations decreased from 114 481 to 103 809. On the basis of these estimates, if we could move currently underperforming interpreters into the acceptable range, we would expect the earlier detection of approximately 86 cancers per 100 000 women and a reduction in the number of false-positive examinations by 1067 per 100 000 women. In our simulated cohort of 1 million women undergoing diagnostic mammography for workup of a breast lump, 66 795 had breast cancer. The number of cancers we estimated that would be correctly recalled for biopsy increased from 59 207 to 62 555 as a result of effective additional training for low-performing interpreters that brought their performance to acceptable levels; likewise, false-positive biopsy recommendations decreased from 102 077 to 95 738. On the basis of these estimates, if we could effectively shift currently underperforming interpreters into the acceptable range, we would expect the earlier detection of

Table 2

Characteristics of 11 Radiologists Involved in Setting Criteria

Characteristic	Datum
Sex	
Male	55
Female	45
Age (y)*	55.3 ± 10.1 (43–69)
Practice setting	
University	45
Community based	45
Community based with university affiliation	9
No. of radiologists in practice group*	37.5 ± 22.4 (5–70)
No. of radiologists in practice group who interpret breast imaging studies*	7.8 ± 5 (3–16)
Completed fellowship training in breast imaging	45
Estimated percentage of clinical time spent in breast imaging	95.5% ± 9.0 (70%–100%)

Note.—Unless otherwise specified, data are percentages.

* Data are means ± standard deviations, with ranges in parentheses.

approximately 335 cancers per 100 000 women and a reduction in the number of false-positive biopsy recommendations by 634 per 100 000 women.

Table 3

Scoring Summary for Phase I and Phase II

	After Abnormal Screening Mammography Study		Workup of Breast Lump	
Measure	Phase I (Before Normative Data Presented)	Phase II (After Normative Data Presented)	Phase I (Before Normative Data Presented)	Phase II (After Normative Data Presented)
Sensitivity				
Mean	81.5	80.9	84.1	86.8
Mode	80.0	80.0	85.0	85.0
Range	78–88	80–82	80–91	83–92
Specificity				
Upper bound				
Mean	90.5	93.2	93.6	95.1
Mode	95.0	95.0	95.0	95.0
Range	80–95	90–95	85–96	95–96
Lower bound				
Mean	79.1	80.9	83.5	85.0
Mode	80.0	80.0	85.0	85.0
Range	70–90	75–85	80–85	85–85
Abnormal interpretation rate				
Upper bound				
Mean	23.9	22.3	28.2	28.2
Mode	20.0	20.0	30.0	30.0
Range	18–40	20–25	20–30	20–30
Lower bound				
Mean	9.0	9.7	10.0	10.0
Mode	10.0	10.0	10.0	10.0
Range	5–12	5–15	10–10	10–10
PPV_2				
Upper bound				
Mean	39.4	39.5	48.4	48.2
Mode	40.0	40.0	45.0	50.0
Range	20.1–50	35–40	40–65	45–55
Lower bound				
Mean	17.6	15.5	25.5	25.0
Mode	15.0	15.0	25.0	25.0
Range	15–25	15–20	17–35	25–30
PPV_3				
Upper bound				
Mean	41.8	44.7	54.5	55.0
Mode	40.0	45.0	55.0	55.0
Range	35–45	42–45	50–55	50–60
Lower bound				
Mean	18.5	19.3	29.5	28.5
Mode	20.0	20.0	30.0	30.0
Range	12–20	18–20	25–30	25–30
Cancer diagnosis				
Lower bound				
Mean	23.9	21.1	41.8	39.5
Mode	25.0	20.0	45.0	40.0
Range	20–28	20–25	25–50	30–45

Discussion

The final criterion set for "low performers" was less than 80% sensitivity for follow-up of abnormal screening examinations and less than 85% for follow-up of a breast lump. Published median benchmarks for diagnostic mammography (4) also involve abnormal interpretation rate (8.0%), PPV_2 (31.5%), PPV_3 (39.5%) and cancer diagnosis rate (25.3 per 1000 examinations). The criteria set for most measures included both an upper and a lower bound. Lower bounds for abnormal interpretation rate were similar to median benchmarks but were substantially lower for PPV_2, PPV_3, and cancer diagnosis rate for workup after abnormal screening examinations compared with diagnostic benchmarks. The upper bounds were much higher for abnormal diagnostic interpretation rate for both workup after abnormal screening examinations and workup of breast lumps. For PPV_2, PPV_3, and cancer diagnosis rate, the ranges between the lower and upper bounds indicate that the cut points were chosen to define the limits of acceptable performance rather than the measures of average performance.

A critical caveat is that performance outside one cut point should be considered within the specific practice setting and overall assessment of all performance measures (22). For example, certain combinations of performance outcomes, such as high cancer diagnosis rate combined with a below-lower-bound abnormal interpretation rate might not warrant a recommendation for additional training. Also, new interpreters may have higher abnormal interpretation rates for a number of years before they establish a stable practice pattern (23). Emerging research shows that improved performance appears to be related more to the combination of both screening and diagnostic mammography study interpretation than to the interpretive volume alone (24). This suggests that to identify low

Table 4

BCSC Normative Performance Data at 10th, 25th, 50th, 75th, and 90th Percentiles (2003–2007)

Type of Mammography and Measure	No. of Radiologists	No. of Mammography Studies	No. of Cancers	Normative Performance Data						
				Minimum	10th Percentile	25th Percentile	50th Median	75th Percentile	90th Percentile	Maximum
Diagnostic mammography for workup of prior abnormal screening results										
Sensitivity*	150	3645 True-positives	4110	30.0	70.3	82.4	90.0	94.4	100	100
Specificity†	246	92 781 True-negatives	105 663 Noncancers	66.4	77.9	84.0	89.7	92.5	94.9	98.4
Abnormal interpretation rate‡	459	119 851	N/A	0.0	6.2	9.7	13.7	19.0	26.3	63.6
PPV_2	288	17 651 (Biopsies recommended)	4249	0.0	11.1	16.7	24.2	33.0	41.5	70.0
PPV_3	246	13 673 (Biopsies performed)	3853	0.0	13.6	20.3	27.1	37.5	47.8	100
CDR	253	110 983	4104 True-positives	0.0	15.4	22.2	31.9	44.6	61.0	106
Diagnostic mammography for workup of a breast lump										
Sensitivity*	91	2012 True-positive	2258	46.2	76.9	83.3	90.0	95.0	100	100
Specificity†	114	29 651 True-negative	33 649 Noncancers	66.8	81.6	87.5	90.1	93.5	95.7	98.7
Abnormal interpretation rate‡	370	46 682	N/A	0.0	7.1	10.9	15.3	21.0	28.0	50.0
PPV_2	181	7080 (Biopsies recommended)	2468	0.0	18.2	27.3	38.7	50.0	61.3	83.3
PPV_3	134	4910 (Biopsies performed)	2025	10.0	26.7	34.4	49.5	60.0	72.7	100
CDR	123	36 931	2172 True-positives	8.3	31.3	44.0	55.2	70.0	91.7	230

Note.—CDR = cancer diagnosis rate per 1000 mammography studies among radiologists with at least 100 diagnostic mammography studies. PPV_2 = PPV from diagnostic mammography for biopsies recommended (among radiologists with at least 10 biopsies recommended). PPV_3 = PPV from diagnostic mammography for biopsies performed (among radiologists with at least 10 biopsies performed).

* Among radiologists with at least 10 breast cancers found within 365 days of diagnostic mammography.

† Among radiologists with at least 100 diagnostic mammography studies without a cancer diagnosis within 365 days.

‡ Among radiologists with at least 10 diagnostic mammography studies.

performers, it will be important to consider low performers in both screening and diagnostic mammography.

Performance measures may be affected by many factors, such as differences in patient populations and a low number of cancers diagnosed. In phase II of this study, we set volume criteria that increased the stability of the rates for performance measures in the normative data. For example, for sensitivity, we included radiologists who interpreted a minimum of 10 mammography studies associated with a cancer diagnosis. Performance measures also may be affected if interpreting physicians use BI-RADS assessment categories in a manner inconsistent with the guidance provided in the BI-RADS atlas (20). Such unintended use has been documented (25).

For outcome measures that reflect “true-positive performance” (cancer diagnosis rate and sensitivity), we did not select an upper bound for acceptability because the primary goal of diagnostic breast imaging is to identify as many cases of breast cancer as possible. However, for the outcome measures that reflect false-positives (abnormal interpretation rate, specificity) and the PPVs, we selected both an upper and a lower bound for acceptability. This is because too high an abnormal interpretation rate, which typically results in a low PPV and specificity, may indicate an excessive number of abnormal assessments, resulting in increased false-positives and a low probability of diagnosing cancer among women with a positive assessment. Similarly, too low an abnormal interpretation rate, typically resulting in a high PPV and specificity, may indicate infrequent abnormal assessments, resulting in too low a cancer diagnosis rate and too high a PPV among women with a positive assessment. Some false-positive examinations are necessary in diagnostic mammography because the mammographic features of early breast cancer may overlap with benign mammographic changes.

A substantial percentage (16%–42%) of BCSC interpreting physicians appear to fall outside one of our six derived cut points, and these physicians are

Table 5

Final Performance Cutoff Points according to Type of Diagnostic Mammography

Type of Diagnostic Mammography and Measure	Low Performance Range	Percentage of BCSC Radiologists in Low Performance Range
Diagnostic mammography for workup of prior abnormal screening results		
Sensitivity	<80	16.1
Specificity	<80 Or >95	22.3
Abnormal interpretation rate	<8 Or >25	21.5
PPV_2	<15 Or >40	25.7
PPV_3	<20 Or >45	34.1
CDR	<20 Per 1000	21.5
Diagnostic mammography for workup of a breast lump		
Sensitivity	<85	32.0
Specificity	<83 Or >95	26.6
Abnormal interpretation rate	<10 Or >25	20.7
PPV_2	<25 Or >50	23.8
PPV_3	<30 Or >55	41.8
CDR	<40 Per 1000	21.5

Note.—CDR = cancer diagnosis rate per 1000 mammography studies among radiologists with at least 100 diagnostic mammography studies.

believed to be representative of all interpreting physicians in the United States (3,4). It may be unrealistic for remedial training to be recommended for all out-of-range physicians; it is likely that only a small percentage of those physicians initially flagged for review will be recommended to have additional training.

Most interpreting physicians do not have access to data for accurate calculations of sensitivity and specificity. However, the combination of abnormal interpretation rate, PPV, and cancer diagnosis rate, especially if supplemented by the early diagnosis parameters of tumor size, lymph node status, and cancer stage, do provide sufficient information from which to determine whether diagnostic mammography performance results in the detection of early stage breast cancer. In fact, the Institute of Medicine has recommended expanding the Mammography Quality Standards Act–mandated mammography audit to provide feedback on several readily acquired, clinically important performance indexes (24).

A limitation of our study was that we did not develop cut points for periodic surveillance of BI-RADS category 3 mammography cases initially recommended for short interval follow-up. Another limitation was that we did not examine how cut points for multiple performance measures could be applied simultaneously to individual interpreting physicians to better identify individual physicians who would benefit from additional training, although we are currently working on methods to do this accurately. A limitation of the simulation model was the assumption that remedial training would be effective for all low performers. Last, our use of 11 experts may not have resulted in a cohort that is representative of all expert mammographers in the United States, and this may have affected the cut points selected.

In conclusion, we have identified minimally acceptable performance levels for physicians interpreting two important types of diagnostic mammography. We recognize that a combination of performance measures must be assessed for any individual interpreter; however, those who fall outside the identified cut points should be reviewed in the context of their practice setting and all available parameters to be considered for additional training.

Acknowledgments: This work was supported the American Cancer Society and made possible by a generous donation from the Longaberger Company's Horizon of Hope Campaign (grants SIRGS-07-271-01, SIRGS-07-272-01, SIRGS-07-274-01, SIRGS-07-275-01, SIRGS-06-281-01, and ACS A1-07-362). 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*. 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/*.

Disclosures of Conflicts of Interest: P.A.C. No relevant conflicts of interest to disclose. **J.P.** Financial activities related to the present article: none to disclose. Financial activities not related to the present article: is on the Scientific Advisory Board of Hologic. Other relationships: none to disclose. **E.A.S.** No relevant conflicts of interest to disclose. **S.A.F.** No relevant conflicts of interest to disclose. **B.M.** No relevant conflicts of interest to disclose. **L.W.B.** No relevant conflicts of interest to disclose. **R.A.S.** No relevant conflicts of interest to disclose. **R.R.** No relevant conflicts of interest to disclose. **L.I.** No relevant conflicts of interest to disclose. **J.W.** No relevant conflicts of interest to disclose. **K.T.** No relevant conflicts of interest to disclose. **D.L.M.** No relevant conflicts of interest to disclose.

References

1. American College of Radiology. ACR practice guideline for the performance of screening and diagnostic mammography. Reston, Va: American College of Radiology, 2011.

2. U.S. Food and Drug Administration. Radiation emitting products: Mammography Quality Standards Act. http://www.fda.gov/RadiationEmittingProducts/Mammography-QualityStandardsActandProgram/default.htm. Accessed May 31, 2012.

3. Rosenberg RD, Yankaskas BC, Abraham LA, et al. Performance benchmarks for screening mammography. Radiology 2006;241(1):55–66.

4. Sickles EA, Miglioretti DL, Ballard-Barbash R, et al. Performance benchmarks for diagnostic mammography. Radiology 2005;235(3):775–790.

5. Leung JW, Margolin FR, Dee KE, Jacobs RP, Denny SR, Schrumpf JD. Performance parameters for screening and diagnostic

mammography in a community practice: are there differences between specialists and general radiologists? AJR Am J Roentgenol 2007;188(1):236–241.

6. Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. Radiology 2002;224(3):861–869.
7. Dee KE, Sickles EA. Medical audit of diagnostic mammography examinations: comparison with screening outcomes obtained concurrently. AJR Am J Roentgenol 2001;176(3):729–733.
8. Miglioretti DL, Smith-Bindman R, Abraham LA, et al. Radiologist characteristics associated with interpretive performance of diagnostic mammography. J Natl Cancer Inst 2007;99(24):1854–1863.
9. Barlow WE, Lehman CD, Zheng Y, et al. Performance of diagnostic mammography for women with signs or symptoms of breast cancer. J Natl Cancer Inst 2002;94(15):1151–1159.
10. Sohlich RE, Sickles EA, Burnside ES, Dee KE. Interpreting data from audits when screening and diagnostic mammography outcomes are combined. AJR Am J Roentgenol 2002;178(3):681–686.
11. Beam CA, Conant EF, Sickles EA. Correlation of radiologist rank as a measure of skill in screening and diagnostic interpretation of mammograms. Radiology 2006;238(2):446–453.
12. Carney PA, Sickles EA, Monsees BS, et al. Identifying minimally acceptable interpretive performance criteria for screening mammography. Radiology 2010;255(2):354–361.
13. National Cancer Institute. Breast Cancer Surveillance Consortium. http://breastscreening.cancer.gov/. Accessed June 22, 2009.
14. Ricker KL. Setting cut scores: critical review of Angoff and Modified-Angoff methods. Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, Alberta, Canada. http://www.google.com/search?q=angoff+method&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-US:official&client=firefox-a. Accessed November 22, 2011.
15. Cope RT. A generalizability study of the Angoff method applied to setting cutoff scores of professional certification tests. Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987). http://www.eric.ed.gov/ERICWebPortal/custom/portlets/recordDetails/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED282921&ERICExtSearch_SearchType_0=no&accno=ED282921. Accessed November 22, 2011.
16. Arrasmith DG, Hambleton RK. Steps for setting standards with the Angoff method. http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED299326&ERICExtSearch_SearchType_0=no&accno=ED299326. Accessed November 22, 2011.
17. 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(1):74–90.
18. Talente G, Haist SA, Wilson JF. A model for setting performance standards for standardized patient examinations. Eval Health Prof 2003;26(4):427–446.
19. George S, Haque MS, Oyebode F. Standard setting: comparison of two methods. BMC Med Educ 2006;6:46.
20. American College of Radiology. BI-RADS (Breast Imaging Reporting and Data System) atlas. http://www.acr.org/Quality-Safety/Resources/BIRADS. Accessed October 1, 2012.
21. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2008.
22. Lemmers O, Broeders M, Verbeek A, Heeten G, Holland R, Borm GF. League tables of breast cancer screening units: worst-case and best-case scenario ratings helped in exposing real differences between performance ratings. J Med Screen 2009;16(2):67–72.
23. Miglioretti DL, Gard CC, Carney PA, et al. When radiologists perform best: the learning curve in screening mammogram interpretation. Radiology 2009;253(3):632–640.
24. Buist DS, Anderson ML, Haneuse SJ, et al. Influence of annual interpretive volume on screening mammography performance in the United States. Radiology 2011;259(1):72–84.
25. Geller BM, Barlow WE, Ballard-Barbash R, et al. Use of the American College of Radiology BI-RADS to report on the mammographic evaluation of women with signs and symptoms of breast disease. Radiology 2002;222(2):536–542.