Published in final edited form as:
AJR Am J Roentgenol. 2006 July ; 187(1): 42–46. doi:10.2214/AJR.05.0455.

Reality Check: Perceived Versus Actual Performance of Community Mammographers

Joshua J. Fenton[1], **Joseph Egger**[2], **Patricia A. Carney**[3], **Gary Cutter**[4], **Carl D'Orsi**[5], **Edward A. Sickles**[6], **Jessica Fosse**[7], **Linn Abraham**[8], **Stephen H. Taplin**[8,9], **William Barlow**[10], **R. Edward Hendrick**[11], and **Joann G. Elmore**[7]

[1]Department of Family and Community Medicine, University of California, Davis, 4860 Y St., Ste 2300, Sacramento, CA 95817 [2]Disease Control and Vector Biology Unit, Department of Infectious & Tropical Diseases, London School of Hygiene & Tropical Medicine, London, England [3]Department of Family & Community Medicine, Dartmouth Medical School, Lebanon, NH [4]Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL [5]Breast Imaging Center, The Emory Clinic, Atlanta, GA [6]Department of Radiology, University of California, San Francisco, San Francisco, CA [7]Division of General Internal Medicine, University of Washington, Harborview Medical Center, Seattle, WA [8]Group Health Cooperative, Center for Health Studies, Seattle, WA [10]Cancer Research and Biostatistics, Seattle, WA [11]Northwestern Memorial Hospital, Lynn Sage Breast Cancer Center, Chicago, IL

Abstract

OBJECTIVE—Federal regulations mandate that radiologists receive regular albeit limited feedback regarding their interpretive accuracy in mammography. We sought to determine whether radiologists who regularly receive more extensive feedback can report their actual performance in screening mammography accurately.

SUBJECTS AND METHODS—Radiologists (n = 105) who routinely interpret screening mammograms in three states (Washington, Colorado, and New Hampshire) completed a mailed survey in 2001. Radiologists were asked to estimate how frequently they recommended additional diagnostic testing after screening mammography and the positive predictive value of their recommendations for biopsy (PPV_2). We then used outcomes from 336,128 screening mammography examinations interpreted by the radiologists from 1998 to 2001 to ascertain their true rates of recommendations for diagnostic testing and PPV_2.

RESULTS—Radiologists' self-reported rate of recommending immediate additional imaging (11.1%) exceeded their actual rate (9.1%) (mean difference, 1.9%; 95% confidence interval [CI], 0.9–3.0%). The mean self-reported rate of recommending short-interval follow-up was 6.2%; the true rate was 1.8% (mean difference, 4.3%; 95% CI, 3.6–5.1%). Similarly, the mean self-reported and true rates of recommending immediate biopsy or surgical evaluation were 3.2% and 0.6%, respectively (mean difference, 2.6%; 95% CI, 1.8–3.4%). Conversely, radiologists' mean self-reported PPV_2 (18.3%) was significantly less than their mean true PPV_2 (27.6%) (mean difference, −9.3%; 95% CI, −12.4% to −6.2%).

[9]Present address: Applied Research Program, National Cancer Institute, Bethesda, MD.

The opinions and assertions contained herein are those of the authors and should not be construed as official or as representing the opinions of the federal government or the National Cancer Institute.

CONCLUSION—Despite regular performance feedback, community radiologists may overestimate their true rates of recommending further evaluation after screening mammography and underestimate their true positive predictive value.

Keywords

breast cancer; breast imaging; mammography

Many physicians cannot accurately gauge their true clinical performance [1, 2]. Indeed, objective measures have sometimes belied physicians' sanguine perceptions of their own performance [1, 2]. Some have suggested that physicians' self-reports reflect either their perceived performance relative to peers [2] or their best intentions rather than their actual practice [3].

Although most clinicians receive little or no feedback regarding their clinical performance, many radiologists who interpret screening mammograms receive regular feedback regarding their interpretive performance. Enforced by the U.S. Food and Drug Administration (FDA), the federal Mammography Quality Standards Act (MQSA) of 1992 requires mammography facilities to collect data on cancer outcomes of women who receive recommendations for biopsy from facility radiologists [4]. The explicit goal of the act's audit requirements is to assist facilities in quality assurance and improvement efforts, although facilities are not obligated to use collected data for these purposes. Nevertheless, the FDA encourages facilities to monitor a range of mammography outcomes and to communicate audit results to radiologists [4]. Many facilities now deliver regular performance feedback to mammographers, including common metrics such as positive predictive value ("biopsy yield") and the overall proportion of women recalled for additional imaging and evaluation ("recall rate").

In 1999, the FDA began enforcing audits at the level of the individual radiologist, but little is known about how radiologists use and interpret performance feedback. We sought to determine whether mammographers who have received regular feedback about biopsy yield and recall rate can estimate their true performance in these domains accurately.

Subjects and Methods

Design, Setting, and Subjects

We conducted a cross-sectional study of radiologists who interpreted screening mammograms in 2001 within three geographically distinct mammography registries participating in the federally funded Breast Cancer Surveillance Consortium [5]. Our study combined data from a mail survey of radiologists with mammography outcome data from 1998 to 2001. Radiologists who interpreted more than 480 screening mammography examinations at consortium facilities during the study period were eligible for inclusion, consistent with MQSA accreditation requirements. Three registries participated: Group Health Cooperative Breast Cancer Surveillance System, a nonprofit health plan in the Pacific Northwest; the New Hampshire Mammography Network, which captures more than 85% of screening mammograms in New Hampshire; and the Colorado Mammography Program, which captures approximately 50% of the screening mammograms in the six-county metropolitan area of Denver. Each registry has reported biopsy yield and recall rate to radiologists at least annually since 1999. The study protocol was approved by the institutional review boards of the University of Washington, Group Health Cooperative of Puget Sound, Dartmouth College, and the Cooper Institute (Colorado).

Radiologist Survey

A committee of mammography experts and community radiologists developed a mail survey instrument, which included questions regarding radiologist demographics, experience in mammography, and frequency of various mammography recommendations. Sequential revision of the survey instrument was guided by extensive pilot testing with community mammographers. One three-part question asked radiologists to estimate the percentage of screening mammography examinations they interpreted for which they recommended immediate additional imaging (i.e., sonogram, diagnostic mammogram views); short-interval follow-up (i.e., follow-up mammogram in 3–6 months); or immediate biopsy or surgical evaluation. The subsequent question asked respondents to estimate the positive predictive value of their biopsy recommendations: “Among women whose screening mammograms you recall for additional workup and then recommend for biopsy, what percent do you think turn out to have breast cancer within one year of the screening mammogram?”

Surveys were mailed with informed consent materials, and response was encouraged by telephone follow-up when necessary. Completed survey data were double-entered into a relational database at each site.

Mammography Data

Radiologist survey data were linked with computerized mammography data for bilateral screening mammograms interpreted by responding radiologists from 1998 to 2001. Included mammograms were designated “routine screening” by the interpreting radiologists and were performed on women older than 40 years without a history of breast cancer and without breast implants. Individual mammogram records contained the date of examination and the BI-RADS assessment and recommendations [6]. Within consortium facilities, radiologists recorded BI-RADS assessment separately and independently from their follow-up recommendation [7]. Thus, in addition to a BI-RADS assessment category, each mammogram includes one of the following recommendations: normal interval follow-up, immediate additional imaging, short-interval follow-up, and immediate biopsy or surgical evaluation. Breast cancer outcomes within each registry are ascertained by regular linkage with regional cancer registries. After encrypting identifiers, mammography data were sent via file transfer protocol for central data analysis in Seattle, WA.

Definitions of Actual Recommendation Rates and Positive Predictive Value

For each radiologist, we determined the proportion of screening mammograms with the following recommendations (1998–2001): immediate additional imaging, short-interval follow-up, and immediate biopsy or surgical evaluation. Because our survey question asked radiologists about the positive predictive value of their biopsy recommendations (PPV_2), we defined a screening mammogram as positive if it contained a recommendation for immediate biopsy and had a BI-RADS assessment of 3, 0, 4, or 5. Although a departure from BI-RADS, radiologists occasionally recommend biopsy alongside a BI-RADS assessment of 3 or 0 [7], so we initially included mammograms with these BI-RADS assessments. We calculated the positive predictive value of a biopsy recommendation (PPV_2) for each radiologist from 1998 to 2001 as the proportion of women who were diagnosed with breast cancer (including ductal carcinoma in situ) within 1 year of an initial positive screening mammogram. We obtained essentially identical PPV_2 estimates after including only mammograms with biopsy recommendations and a BI-RADS assessment of 4 or 5, so we report here the results of the initial calculation.

Data Analyses

We first performed analyses to assure the absence of time trends in recommendation rates or PPV_2 during the study period. Generalized estimating equations were used to examine the association between recall rate (the probability of a BI-RADS assessment of 0, 4, 5, or 3 with a recommendation for immediate follow-up) and screening year [8]. The association between PPV_2 and year was investigated by selecting screenings with a BI-RADS assessment of 3, 0, 4, or 5 with biopsy recommendation and fitting a similar model in which the outcome was the probability of a cancer diagnosis during follow-up. Both models included the year of screening mammogram as the main covariate of interest (1998, 1999, 2000, or 2001), adjusted for mammography registry, and accounted for the correlation within a radiologist using an independent correlation structure. We found no statistically significant association between study year and either recall rate or PPV_2.

We first computed the mean differences (and 95% confidence intervals [CIs]) between radiologists' self-reported and actual rates of each recommendation and PPV_2. Mean recommendation rates were weighted by the number of screening mammograms interpreted by each radiologist during the study period; means of PPV_2 were weighted by the total number of positive mammograms with biopsy recommendations for each radiologist. We used weighted means so that results would not be unduly affected by radiologists who made relatively few recommendations during the study period. General linear models were used to study whether these mean differences were associated with radiologist characteristics (e.g., demographics, academic affiliation, breast imaging experience). Lastly, we tested whether statistically significant correlation existed between the radiologists' self-reported and actual recommendation rates and PPV_2. Statistical tests were two-sided with an alpha level of 0.05.

Results

Radiologist Sample

Of 181 potentially eligible radiologists, 139 (77%) returned the survey with consent to link responses to mammography outcomes. Responders and nonresponders did not significantly differ with regard to sex, years since medical school graduation, recall rate, or accuracy of interpretation (either sensitivity or specificity). Of these, 105 (76%) had complete data on perceived and actual performance rates and had interpreted 480 or more screening mammograms from 1998 to 2001, altogether including 366,128 screening mammograms at 65 facilities in three U.S. states. The radiologists were predominantly male and with more than 10 years of experience interpreting mammograms, although most (88%) spent less than 40% of their work time in breast imaging (Table 1).

Actual Versus Perceived Performance

Radiologists' mean perceived rate of recommending immediate additional imaging (11.1%) was slightly higher than their actual mean rate (9.1%) (mean difference, 1.9%; 95% CI, 0.9%–3.0%) (Table 2). However, radiologists' perceived rate of recommending short-interval follow-up exceeded the actual rate by threefold, and radiologists' perceived recommendation rate for immediate biopsy or surgical evaluation exceeded the actual rate by fivefold (Table 2). In contrast, the mean perceived PPV_2 was significantly less than the actual PPV_2 (mean difference, −9.3%; 95% CI, −12.4% to −6.2%).

In general, radiologists overestimated their recall rates and underestimated their PPV_2, regardless of their demographic characteristics, full-time versus part-time status, academic affiliation, experience in breast imaging, or recent volume of mammography (data not shown). In bivariate analyses, radiologists with a primary academic affiliation, who were fellowship trained in breast imaging, or who interpreted a lower volume of mammograms,

overestimated their rate of recommending immediate additional imaging to a significantly greater degree ($p < 0.05$) (data not shown).

A moderate correlation was found between radiologists' perceived and actual rates of recommendations for immediate additional imaging ($r = 0.36$, $p < 0.001$; Fig. 1). Radiologists' perceived and actual recommendations for short-interval follow-up were similarly correlated ($r = 0.41$, $p < 0.001$). Perceived and actual rates of recommendation for immediate biopsy or surgical evaluation were weakly correlated ($r = 0.17$, $p = 0.09$). No significant correlation was found between radiologists' perceived and actual PPV_2 ($r = 0.10$, $p = 0.31$; Fig. 2).

Discussion

We found significant discrepancies between radiologists' actual rates of recommending further evaluation after screening mammography and their perception of these rates. Radiologists in this survey substantially underestimated their true PPV_2, believing that far fewer women had cancer after a biopsy recommendation than actually did. We found a similar pattern of findings regardless of radiologists' demographic characteristics, academic affiliation, and experience in breast imaging. Although the radiologists in the study had received at least annual performance reports from their respective mammography registries since roughly 1997, their perceptions of their performance only moderately reflected their true performance in clinical practice.

One explanation for our findings is that most radiologists simply do not review or remember the results from their past outcome audits. Although little is known about how radiologists use data from outcome audits to modify their interpretive practice, studies in other clinical settings suggest that feedback has the greatest effect on the minority of clinicians who deviate substantially from the practice norm and has a comparatively small effect on most clinicians who may view themselves as within the norm [9]. Similarly, most radiologists might judge from audit reports whether their interpretive performance is within the norm, and if so, they may quickly forget their audit results. If radiologists use audit reports in this manner, the principal effect of the MQSA audit requirements may be the encouragement of normative interpretive behavior among U.S. mammographers.

Radiologists overestimated the frequency with which they recommend further evaluation after screening mammography and underestimated their PPV_2. Together, these findings suggest that radiologists in the study tended to believe their false-positive rate is higher than it actually is. In other words, the radiologists tended to underestimate their specificity. Why might U.S. mammographers overestimate their true false-positive rate? Recall rates are known to be higher in the United States compared with programs in other countries, which may be partly attributable to fears of malpractice among U.S. mammographers [10, 11]. Indeed, media reports [12, 13] have emphasized that the relatively high recall rate in the United States has not substantially increased the cancer detection rate compared with screening programs in other countries. Mammographers in our sample may have developed an exaggerated impression of their own false-positive rate if their self-perceptions were influenced by media reports suggesting that U.S. recall rates are unnecessarily high.

Our study has several important limitations. First, we compared self-reported recommendation rates and PPV_2 in 2001 to actual rates computed from 1998 to 2001, which allowed more precise estimates of individual radiologist performance. Although we found no evidence of temporal trends in recommendation rates or PPV_2 during the study period, it is possible that recall rate or PPV_2 could have changed over time for individual radiologists. In the absence of temporal trends across our study population, we nevertheless are confident

in the validity of our principal findings. Second, although we piloted our mail survey extensively to reduce the likelihood of misinterpretation, some radiologists could have misunderstood the survey questions regarding follow-up recommendations or PPV_2. Finally, although the radiologists in the study may not be representative of the entire U.S. population of mammographers, our sample includes both community-based and academic radiologists practicing in diverse facilities in three distinct U.S. geographic regions.

We conclude that the radiologists within three U.S. mammography registries tended to overestimate their true frequency of recommending further evaluation after screening mammography. The same radiologists tend to underestimate their PPV_2, despite receiving at least annual feedback from their facilities regarding these specific aspects of their interpretive performance. Our findings suggest that many radiologists may not state accurately the results of outcome audits previously reported to them. This calls into question the potential value of future federal regulations that might require reporting of specific outcomes as a means of feedback to encourage improved clinical performance. Research is needed to characterize how radiologists interpret and use feedback to modify their interpretive practice.

Acknowledgments

Supported by the Agency for Health Research and Quality and the National Cancer Institute (grants HS10591, U01 CA63731, 5 U01 CA 63736–09, and 1 U01 CA86082-01).

We appreciate the dedication of the participating radiologists and project support staff.

References

1. Saver BG, Taylor TR, Treadwell JR, Cole WG. Do physicians do as they say? The case of mammography. Arch Fam Med. 1997; 6:543–548. [PubMed: 9371047]
2. Streja DA, Rabkin SW. Factors associated with implementation of preventive care measures in patients with diabetes mellitus. Arch Intern Med. 1999; 159:294–302. [PubMed: 9989542]
3. Myers RE, Hyslop T, Gerrity M, et al. Physician intention to recommend complete diagnostic evaluation in colorectal cancer screening. Cancer Epidemiol Biomarkers Prev. 1999; 8:587–593. [PubMed: 10428195]
4. United States Food and Drug Administration. [Accessed January 12, 2005] MQSA Final Regulations. Available at: www.fda.gov/cdrh/mammography/robohelp/START.HTM
5. Ballard-Barbash R, Taplin SH, Yankaskas BC, et al. Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database. AJR. 1997; 169:1001–1008. [PubMed: 9308451]
6. D'Orsi, CJ.; Bassett, LW.; Feig, S., et al. Breast imaging reporting and data system (BI-RADS). 3. Reston, VA: American College of Radiology; 1998. American College of Radiology (ACR).
7. Taplin SH, Ichikawa LE, Kerlikowske K, et al. Concordance of breast imaging reporting and data system assessments and management recommendations in screening mammography. Radiology. 2002; 222:529–535. [PubMed: 11818624]
8. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 4:695–702.
9. Jamtvedt G, Young JM, Kristoffersen DT, Thomson O'Brien MA, Oxman AD. Audit and feedback: effects on professional practice and health care outcomes. Cochrane Database Syst Rev. 2003:CD000259. [PubMed: 12917891]
10. Smith-Bindman R, Chu PW, Miglioretti DL, et al. Comparison of screening mammography in the United States and the United Kingdom. JAMA. 2003; 290:2129–2137. [PubMed: 14570948]
11. Elmore JG, Nakano CY, Koepsell TD, Desnick LM, D'Orsi CJ, Ransohoff DF. International variation in screening mammography interpretations in community-based programs. J Natl Cancer Inst. 2003; 95:1384–1393. [PubMed: 13130114]

12. Tanner, L. US mammogram results lag UK. Chicago Tribune; October 29. 2003 p. 7
13. Maugh, TH. Study warns of mammogram false alarms. Los Angeles Times; April 16. 1998 p. 1

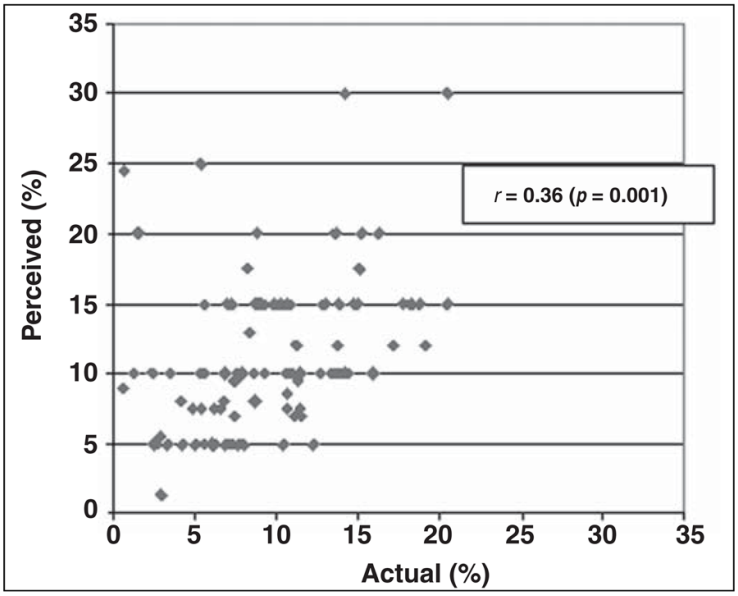

Fig. 1.
Scatterplot of radiologists' self-reported positive predictive value (PPV_2) compared with their actual PPV_2. Positive mammograms included recommendation for biopsy, fine-needle aspiration, or surgery.

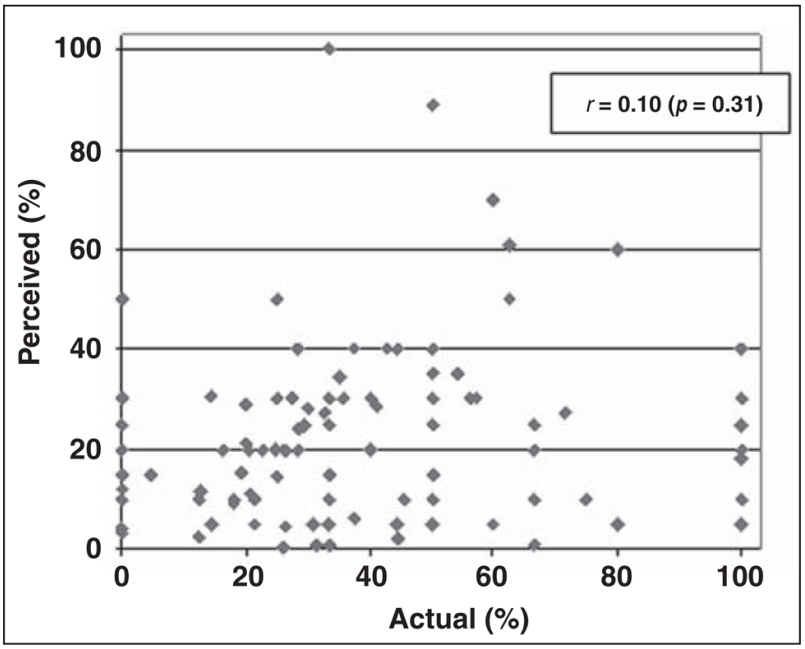

Fig. 2.
Scatterplot of radiologists' perceived and actual rates of recommending immediate additional imaging.

TABLE 1

Characteristics of 105 Responding Radiologists

Characteristic	No.	(%)
Demographics		
Sex		
Male	79	75.2
Female	26	24.8
Age (y)		
34–44	31	29.5
45–54	42	40.0
55+	32	30.5
History of breast cancer		
None	16	15.2
Colleague or friend	48	45.7
Self, spouse or partner, or relative	41	39.0
Practice		
Work full-time		
Yes	75	71.4
No	30	28.6
Academic affiliation		
Yes, primary	5	4.8
Adjunct or affiliate	10	9.5
No	90	85.7
General experience in breast imaging		
Fellowship training in breast imaging		
Yes	4	3.8
No	101	96.2
Number of years interpreting mammograms		
< 10	23	21.9
10–19	51	48.6
≥20	31	29.5
Time working in breast imaging		
< 20%	43	41.0
20–39%	49	46.7
≥40%	13	12.4
Work reported in 2001		
Number of mammograms interpreted		
< 1,000	23	22.1
1,001–2,000	40	38.5
> 2,000	41	39.4
Percentage that were screening mammograms		
0–50	9	8.6

Characteristic	No.	(%)
51–75	47	44.8
76–100	49	46.7

TABLE 2

Perceived and Actual Recommendation Rates and Positive Predictive Values of 105 Radiologists Who Performed 336,128 Screening Mammograms

Item Rated	Self-Reported Rate (Mean %)	Actual Rate (Mean %)	Difference (Self-Reported—Actual)	
			Mean %	95% CI
Recommendations[a]				
Immediate additional imaging	11.1	9.1	1.9	(0.9–3.0)
Short-interval follow-up	6.2	1.8	4.3	(3.6–5.1)
Immediate biopsy or surgical evaluation	3.2	0.6	2.6	(1.8–3.4)
Positive predictive value[b]				
Proportion of women recommended for biopsy diagnosed with breast cancer within 1 y	18.3	27.6	−9.3	(−12.4 to −6.2)

Note—CI = confidence interval.

[a] Means for recommendations are weighted by the number of screening mammograms interpreted by each radiologist from 1998 to 2001.

[b] Equivalent to American College of Radiology definition of PPV_2. Positive mammograms included screening mammograms with a BI-RADS assessment of 3, 0, 4, or 5 that also had a recommendation for biopsy. Means for positive predictive value are weighted by the total number of positive screening mammograms (including recommendations for biopsy) per radiologist from 1998 to 2001.