Evidence-based Target Recall Rates for Screening Mammography[1]

Michael J. Schell, PhD
Bonnie C. Yankaskas, PhD
Rachel Ballard-Barbash, MD, MPH
Bahjat F. Qaqish, PhD, MD
William E. Barlow, PhD
Robert D. Rosenberg, MD
Rebecca Smith-Bindman, MD

Purpose: To retrospectively identify target recall rates for screening mammography on the basis of how sensitivity shifts with recall rate.

Materials and Methods: The study group included 1 872 687 subsequent and 171 104 first screening mammograms from 1996 to 2001 from 172 and 139 facilities, respectively, in six sites of the Breast Cancer Surveillance Consortium. Institutional review board (IRB) approval was obtained from each site. Informed consent requirements of the IRBs were followed. The study was HIPAA compliant. Recall rate was defined as the percentage of screening studies for which further work-up was recommended by the radiologist. Sensitivity was defined as the proportion of cancers that were detected at screening mammography. Piecewise linear regression was used to model sensitivity as a function of recall rate. This model allows detection of critical recall rates in which significant changes (shifts) occurred in the rates that sensitivity increased with increasing recall rate. Rates were interpreted as number of additional work-ups per additional cancer detected (AW/ACD) or, in other words, the estimated number of additional women needed to be recalled at a given rate to detect one additional cancer.

Results: For first mammograms, a single shift in the estimated AW/ACD rate occurred at a recall rate of 10.0%, with the rate jumping dramatically from 35 to 172. For subsequent mammograms, four shifts were identified. At a recall rate of 6.7%, the estimated AW/ACD increased from 80 to 132, which rendered it the highest desirable target recall rate. At a recall rate of 12.3%, the estimated AW/ACD was 304, which suggests little benefit for any higher recall rate.

Conclusion: Recall rates of 10.0% for first and 6.7% for subsequent mammograms are recommended targets on the basis of their AW/ACD rates (less than 100).

[1] From the Biostatistics Division, Department of Interdisciplinary Oncology, Moffitt Research Center, 12902 Magnolia Dr, Tampa, FL 33612-9497 (M.J.S.); Departments of Radiology (B.C.Y.) and Biostatistics (B.F.Q.), University of North Carolina, Chapel Hill, NC; Applied Research Program, Division of Cancer Control and Population Studies, National Cancer Institute, Bethesda, Md (R.B.); Department of Biostatistics, University of Washington, Seattle, Wash (W.E.B.); Department of Radiology, University of New Mexico, Health Sciences Center, Albuquerque, NM (R.D.R.); and Department of Radiology, Epidemiology and Biostatistics, University of California at San Francisco, San Francisco, Calif (R.S.). Received February 27, 2006; revision requested April 27; revision received July 28; accepted August 29; final version accepted October 4. Supported by a National Cancer Institute–funded Breast Cancer Surveillance Consortium cooperative agreement (U01CA63740, U01CA86076, U01CA86082, U01CA63736, U01CA70013, U01CA69976, U01CA63731, U01CA70040). **Address correspondence to** M.J.S. (e-mail: *michael.schell@moffitt.org*).

Considerable variation in recall rates exists between different mammographers, practices, and countries (1–4). Whereas some variation may be because of differences among the populations being screened and the ability of the radiologist, much is almost certainly because of variation in radiologist preference with regard to the importance of finding every cancer (reflected in their sensitivity) and tolerance of false-positive findings at examinations (reflected in their specificity and positive predictive value [PPV]). The recall rate in a facility is defined as the percentage of screening studies for which further work-up is recommended. Recall rates in screening programs and facilities have been reported to range from less than 1% to about 15% for screening mammography (1,5). Across screening programs, recall rate has been shown to be positively correlated with sensitivity and negatively correlated with PPV (1,5). Thus, use of a lower recall rate places a greater emphasis on maintaining a high PPV, while use of a higher recall rate places greater value on achieving high sensitivity.

Different groups have recommended different target recall rates. European guidelines recommend a target recall rate of 5%, with an acceptable rate of less than 7% for first screenings and a target recall rate of 3% (acceptable rate <5%) for subsequent screenings (6,7). The American College of Radiology and the U.S. Agency for Health Care Policy and Research both recommend an overall recall rate of less than 10% (8,9). However, to our knowledge, these targets have not been evaluated relative to their effect on sensitivity and PPV on the basis of data that reflect current mammographic screening examinations within clinical practice in the United States.

Thus, the goal of our study was to retrospectively identify target recall rates for screening mammography on the basis of how sensitivity shifts with recall rate.

Advances in Knowledge

- Evidence-based target recall rates for screening mammography are given by using the concept of additional work-ups per additional cancers detected.
- Advanced statistical modeling of the regression relationship (a concave, monotone, piecewise linear fit) between sensitivity and recall rates for screening mammography is provided.

Materials and Methods

Study Data

The study group included all screening mammograms from 1996 to 2001 from six sites of the Breast Cancer Surveillance Consortium (BCSC) from which data from individual facilities were available. The BCSC is a consortium of mammographic facilities funded by the National Institutes of Health for the purpose of evaluating the performance of mammography in the community setting (10) and represents a diverse U.S. population (11). Seven community-based mammographic registries located in Vermont, New Hampshire, North Carolina, Colorado, New Mexico, California, and the state of Washington have created mammographic databases that link with population-based cancer databases. Each registry and the Statistical Coordinating Center (SCC) of the BCSC has received a federal certificate of confidentiality and approval from each institution's review board for the protection of human subjects to collect and send data (12) to the SCC and to conduct research with these data. Three of seven sites were granted a waiver of informed consent. At three of the other sites, women had the option to exclude their data from research. At one site, the patient's signature was required to allow inclusion of data for research. This study was Health Insurance Portability and Accountability Act compliant.

For our study, data from one site were not included because that site did not collect data at the facility where mammography was performed.

All data related to the screening mammographic examination were collected at the facility at the time of mammography. At mammography, patients completed a breast health survey, which included date of birth, history and date of previous mammography, and reported presence of breast signs and symptoms (lump, nipple discharge, or others, not including breast pain).

The interpreting radiologist recorded the indication for the examination, additional imaging studies performed, and date of previous mammography. In addition, breast density and mammographic assessment were recorded by using the recommended categories of the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) (13). Breast density was categorized as extremely dense, heterogeneously dense, scattered fibroglandular densities, or almost entirely fat. Mammographic assessment was performed with the following categories: 0, needs additional imaging evaluation; 1, negative finding; 2, benign finding; 3, probably benign finding; 4, suspicious abnormality; and 5, highly suggestive of malignancy.

The registries collected breast cancer information from regional Surveillance, Epidemiology, and End Results programs, state cancer registries, and pathology databases. Cancers were categorized as either invasive disease or ductal carcinoma in situ. (Lobular carcinoma in situ was considered benign for this analysis.)

Published online
10.1148/radiol.2433060372

Radiology 2007; 243:681–689

Abbreviations:
AW/ACD = additional work-ups per additional cancer detected
BCSC = Breast Cancer Surveillance Consortium
BI-RADS = Breast Imaging Reporting and Data System
PPV = positive predictive value

Author contributions:
Guarantor of integrity of entire study, M.J.S.; 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, M.J.S., B.C.Y., B.F.Q., R.S.; clinical studies, B.C.Y., R.D.R., R.S.; statistical analysis, M.J.S., B.F.Q., W.E.B., R.S.; and manuscript editing, all authors

Authors stated no financial relationship to disclose.

Mammograms were separated into first or subsequent examinations; 4.4% of the mammograms were dropped at this point, because it could not be determined to which group they belonged. For the subsequent mammograms, 23 (11.8%) of 195 facilities were excluded from the analysis. Reasons for exclusion were that 22 facilities had no cancer results (1993 mammograms) and one facility had a recall rate greater than 40% (6111 mammograms). For first mammograms, 50 (26.5%) of 189 facilities were excluded from the analysis. Reasons for exclusion were that 44 (88%) of the 50 facilities had no cancer results (7999 mammograms), five (10%) of 50 facilities had fewer than 100 mammograms (313 mammograms), and one facility had a recall rate greater than 40% (429 mammograms). After the exclusions, 1 872 687 subsequent mammograms from 172 facilities and 171 104 first mammograms from 139 facilities remained in the study for analysis. Overall, findings from 2 043 791 screening mammograms obtained by 912 radiologists in 172 facilities were included in this study. This represents an average of 2241 mammograms per radiologist.

Mammographic Assessment

All mammographic studies were assessed by radiologists with BI-RADS assessments. The follow-up period for all screening mammography was 12 months or the time to the next screening examination, if that occurred between 9 and 12 months later. For our sensitivity calculations, a screening mammographic examination was considered to yield a positive finding if the assessment was needs further evaluation, suspicious abnormality, suspicious for malignancy (BI-RADS categories 0, 4, and 5, respectively), or probably benign (category 3) when accompanied by a recommendation for immediate imaging follow-up. A screening mammographic examination was considered to yield a negative finding if the assessment was normal, benign (BI-RADS categories 1 and 2, respectively), or probably benign (BI-RADS category 3) and did not have a recommendation for immediate follow-up.

Reference Standard

A mammogram with a positive finding yielded a true-positive finding (TP) if cancer was diagnosed and a false-positive finding (FP) if cancer was not diagnosed in the follow-up period. A mammogram with a negative finding yielded a true-negative finding (TN) if no breast cancer was diagnosed and a false-negative finding (FN) if cancer was diagnosed in the follow-up period.

Sensitivity was defined as the proportion of cancers that were detected, calculated as TP/(TP + FN). Specificity was defined as the proportion of individuals without cancer correctly classified as having a negative finding at mammography, calculated as TN/(TN + FP). Recall rate was defined as the proportion of individuals recalled for additional work-up, calculated as (TP + FP)/(TP + FP + TN + FN). Cancer incidence per 1000 mammograms was calculated as $1000 \cdot (TP + FN)/(TP + FP + TN + FN)$.

Statistical Analysis

Sensitivity increases with recall rate, but not necessarily linearly. By using a four-step procedure, a nondecreasing, monotonic, piecewise linear fit to the data was constructed for sensitivity as a function of recall rate on the basis of facility-level data that were weighted by the number of cancers. First, isotonic regression analysis (14) was used to model sensitivity as a constant for various ranges of the recall rate. Isotonic regression provides the least-squares fit to the raw data among the class of all monotonic functions. Second, reduced monotonic regression (15) ($\alpha^* = .50$) was used to identify the recall rate groups by combining isotonic regression level sets with similar sensitivity measures. Third, the reduced monotonic regression model was adjusted for site, mean age of women, and percentage of long-interval mammographic examinations (defined for subsequent mammograms as the percentage of mammograms at the facility whose previous mammograms were more than 27 months earlier). Breast density and percentage of women with a personal history of breast biopsy were not included in these adjustments because of incomplete and/or inconsistent reporting across the facilities. Fourth, a concave, monotone, piecewise linear fit (called "concave fit" henceforth) was obtained by joining the mean recall rates for the adjusted sensitivities of the recall rate groups. A piecewise linear segment was used to span multiple groups, if a nonincreasing slope with increasing recall (the concavity requirement) was needed. Separate fits were obtained for first and subsequent mammograms.

We provide a brief explanation of the four-step modeling procedure as follows. Because of random variation, virtually no regression relationships are perfectly ordered. In our case, sensitivity does not perfectly increase with increasing recall rate. The recall rate groups provide regions of the domain (recall rate in our case) where the response (sensitivity in our case) is found to be fairly constant. Because we do not believe that sensitivity is intrinsically flat, with jump points at certain recall rates, we use the recall rate groups to construct the concave fit by using linear interpolation of points.

The slopes for the piecewise linear segments were interpreted as number of additional work-ups per additional cancer detected (AW/ACD). We defined AW/ACD as $(D_{NR})/(D_{CD})$, where D_{NR} is difference in number of patients recalled and D_{CD} is difference in number of cancers detected. $D_{CD} = O_{CR} \cdot D_S$, where O_{CR} is the overall cancer rate and D_S is difference in sensitivities. O_{CR} is the number of cancers per mammogram for the entire study. This statistic is the reciprocal of what could be called the "incremental PPV" (ie, ACD/AW), where the incremental PPV is obtained by including only women who would not have been recalled at the lower recall rate.

The 95% confidence interval for AW/ACD was obtained by substituting in the lower and upper 95% confidence limits for D_S in the AW/ACD formula. The limits were obtained by adding and subtracting 1.96 times the standard deviation for D_S, where standard deviation was obtained by using standard binomial theory. For example, the variance for the difference in sensitiv-

ity between recall rate groups 3 and 4 for subsequent mammograms (Table 1) is $0.691 \cdot (1 - 0.691)/1019 + 0.749 \cdot (1 - 0.749)/4486 = 0.000251$, so the 95% confidence interval for D_S is $0.749 - 0.691 \pm 1.96 \cdot \sqrt{0.0000251} = (0.027, 0.089)$.

All analyses were performed by an author (M.J.S.), by incorporating modeling suggestions by the authors with primary mammographic expertise (B.C.Y., R.B., R.D.R., R.S.); technical consultation and review was provided by the other two statisticians (W.E.B., B.F.Q.). Statistical software (SAS, version 8.2; SAS Institute, Cary, NC) was used for all analyses.

Results

Overall, 171 104 first and 1 872 687 subsequent mammograms obtained at 172 facilities from six BCSC sites were included. Performance measures for the study population were stratified by using demographic characteristics (Table 2). For subsequent mammograms, the mean recall rate was 8.1%, with a sensitivity of 78.2% and a PPV of 5.0%. The overall cancer rate was 5.15 cancers per 1000 mammograms (9650/1 872 687). For first mammograms, the recall rate was 13.2%, with a sensitivity of 85.9% and a PPV of 3.5%. The overall cancer rate was 5.32 cancers per 1000 mammograms (910/171 104). Sensitivity and PPV generally increased with increasing patient age for both first and subsequent mammograms. While the recall rate predominantly decreased with increasing patient age for subsequent mammograms, it increased from the younger than 40 years age group to the 50–59 years age group before decreasing for first mammograms.

Subsequent Mammograms

The association between sensitivity and recall rate for subsequent screening mammography was modeled by using a concave fit (Fig 1), which was constructed by using the seven recall rate groups obtained from the reduced monotonic regression fit (Table 1). For example, in the third recall rate group, 1019 women from 30 facilities had breast cancer at the follow-up period, with a mean recall rate of 4.3% (range, 3.2%–5.2%) and a sensitivity of 69.1%.

As the recall rate increases, so does the AW/ACD value (Fig 2). For example, increasing recall rate from 1.1% to 2.5% would require an estimated 29 additional work-ups to find one additional cancer, compared with 51 for increasing recall rates from 2.5% to 4.3%. The two groups with the lowest mean recall rates (1.1% and 2.5%) represent only six facilities and less than 4% of mammographic examinations performed. Thus, the two groups represent recall rates that most mammographers find to be unacceptably low. The shift from group 3 to group 4 is associated with an estimated AW/ACD of 80; 84.1% of screening mammograms were evaluated at facilities at or beyond recall rate group 4. Group 4, with recall rates between 5.3% and 9.2%, was the largest, representing 75 facilities that performed 47.8% of all subsequent mammograms. Recall rate groups 5 and 6 (range, 9.3%–13.6%) were associated with a single shift of AW/ACD of 132, due to the concavity requirement; 29.8% of mammograms were screened at these recall rates. An additional 6.5% of mammograms were screened at recall rates higher than 14%, associated with a very high AW/ACD of 304.

Recall rate group 4, with a mean recall rate of 6.7%, represents the best choice for mammographers who wish to maximize their sensitivity while keeping the estimated AW/ACD less than 100. Group 4 facilities accounted for 47.8% of mammograms evaluated, with 15.9% and 36.3% of women being screened at facilities with lower and higher recall rates, respectively (Table 1). Estimated performance measures for recall rates ranging from 3% to 12% were obtained from the concave fit (Table 3); this information might be useful to mammographers contemplating shifts in their individual recall rates.

First Mammograms

The association between sensitivity and recall rate for first screening mammography was modeled with a concave fit, which had a single shift in AW/ACD estimates (Fig 3), on the basis of four

Table 1

Association between Recall Rate and Sensitivity for Subsequent Mammographic Screenings

Recall Rate Group	No. of Facilities	No. of Mammograms*	No. of Cancers	Recall Rate (%)		Sensitivity†	Cancer Detection Rate‡
				Range	Mean		
1	3	26 903 (1.4)	110	0.3–1.5	1.1	53.0	2.73
2	3	45 922 (2.5)	230	2.3–3.1	2.5	62.3	3.21
3	30	224 844 (12.0)	1019	3.2–5.2	4.3	69.1	3.56
4	75	894 982 (47.8)	4486	5.3–9.2	6.7	74.9	3.86
5	27	337 221 (18.0)	1742	9.3–11.5	10.2	78.5	4.04
6	20	220 362 (11.8)	1260	11.6–13.6	12.3	83.0	4.27
7	14	122 453 (6.5)	803	14.0–20.4	15.4	85.0	4.38
Overall	172	1 872 687 (100)	9650	0.3–20.4	8.1	78.2	4.03

* Data in parentheses are percentages.

† Adjusted for site, mean age of woman, and percentage of long-interval mammograms.

‡ Number of cancers detected per 1000 mammograms, assuming 5.15 cancers per 1000 mammograms, which was the rate observed in our study for subsequent screenings.

Table 2

Performance according to Demographic Characteristics

Characteristic	Subsequent Mammography					First Mammography				
	No. of Examinations*	Mean Recall Rate	Sensitivity (%)	PPV (%)	No. of Cancers	No. of Examinations*	Mean Recall Rate	Sensitivity (%)	PPV (%)	No. of Cancers
Age group (y)										
<40	46 987 (2.5)	9.0	68.2	1.4	88	60 472 (35.3)	11.2	71.3	1.4	129
40–49	545 684 (29.1)	9.2	70.7	2.3	1640	70 634 (41.3)	14.3	82.1	2.0	240
50–59	574 505 (30.7)	8.3	78.1	4.5	2779	18 068 (10.6)	15.6	92.1	4.9	151
60–69	367 301 (19.6)	7.4	79.7	7.0	2381	11 245 (6.6)	13.5	89.3	8.9	150
≥70	338 210 (18.1)	6.6	81.9	10.2	2762	10 685 (6.2)	12.4	91.7	16.6	240
Race										
Black	107 722 (5.8)	6.4	77.3	5.4	484	17 754 (10.4)	12.6	80.8	4.5	125
White	1 267 170 (67.7)	8.1	78.1	5.1	6648	105 085 (61.4)	13.2	85.2	3.2	520
Other	111 280 (5.9)	6.7	80.7	5.3	483	18 684 (10.9)	10.9	88.3	4.1	94
Unknown	386 515 (20.6)	8.9	78.2	4.6	2035	29 581 (17.3)	14.9	90.6	3.5	171
Family history of breast cancer										
Yes	245 660 (13.1)	8.8	76.8	6.0	1689	12 556 (7.3)	14.7	86.9	4.0	84
No	1 343 531 (71.7)	8.6	80.0	4.6	6615	135 166 (79.0)	13.7	87.3	3.3	691
Unknown	283 496 (15.1)	5.1	71.5	6.6	1346	23 382 (13.7)	9.2	78.5	4.9	135
History of breast biopsy or surgery										
Yes	341 606 (18.2)	10.3	75.7	5.1	2372	6097 (3.6)	17.3	79.2	3.6	48
No	1 383 100 (73.9)	7.7	79.3	4.9	6494	154 381 (90.2)	13.1	85.4	3.3	785
Unknown	147 981 (7.9)	7.0	76.9	5.8	784	10 626 (6.2)	11.7	96.1	5.9	77
Current breast problem										
Yes	27 974 (1.5)	16.9	76.1	6.5	402	6513 (3.8)	22.4	88.6	7.5	123
No	1 362 689 (72.8)	8.1	78.0	4.7	6592	120 755 (70.6)	12.5	84.8	3.0	545
Unknown	482 024 (25.8)	7.6	79.2	5.7	2656	43 836 (25.6)	13.5	87.2	3.5	242
Time since last mammography (m)										
≤27	1 414 723 (75.5)	7.5	76.3	5.0	6937					
>27	302 470 (16.2)	9.6	83.9	5.1	1785					
Unknown	155 494 (8.3)	10.6	81.6	4.6	928					
Parenchymal density										
Extremely dense	112 195 (6.0)	9.9	64.1	3.6	618	14 499 (8.5)	11.8	70.5	2.5	61
Heterogeneously dense	572 461 (30.6)	9.5	75.3	4.6	3311	55 869 (32.7)	14.0	81.6	2.8	267
Scattered fibroglandular densities	686 148 (36.6)	7.2	82.4	5.5	3304	58 288 (34.1)	13.3	90.3	4.4	373
Almost entirely fat	129 132 (6.9)	3.9	89.1	6.4	357	10 393 (6.1)	8.6	88.5	5.2	52
Unknown	372 751 (19.9)	8.4	78.6	5.1	2060	32 055 (18.8)	13.5	87.9	3.1	157
Overall	1 872 687 (100.0)	8.1	78.2	5.0	9650	171 104 (100.0)	13.2	85.9	3.5	910

* Data in parentheses are percentages.

recall rate groups (Table 4). Recall rate group 2 had the greatest number of mammograms, at 53.7%, and a recall rate range of 6.1%–13.2%, with an average of 10%. Recall rate group 3, with recall rates ranging from 13.3% to 23.1%, reflects the practices where 40.4% of mammograms were evaluated. Recall rate groups 1 and 4 represent exceptionally low (2.4%–6.0%) and high (23.2%–27.9%) practice patterns, seen in a total of 10 practices (5.9% of patients).

These four recall rate groups give rise to a single shift in AW/ACD. Below 10%, the estimated value was 35, compared with 172 for higher recall rates. Thus, a target recall of 10% is the best choice for mammographers who wish to maximize their sensitivity while keeping the estimated AW/ACD below either 50 or 100 (Fig 4). Group 2 facilities, whose recall rates include the target rate of 10%, accounted for 53.7% of cancers detected, with 3.4% and 42.9% of women being screened at facilities with lower and higher recall rates, respectively (Table 4).

Discussion

For a given mammographer, sensitivity clearly increases with recall rate, because recalling additional women from a given cohort could not decrease the true-positive rate. Indeed, if all women were recalled, very few cancers would be missed. Consequently, establishing a target recall rate should not be based on

maximizing sensitivity alone. Judgment is needed to settle on a recall rate at which the additional yield of cancers detected is not worth the additional number of recalls. Such a decision is difficult, because it represents a trade-off between the benefit of finding additional cancers and the increased number of procedures for noncancers and the associated anxiety and monetary cost that women experience as they undergo further work-ups (16–18). It seems reasonable that values of AW/ACD should be below the prevailing cancer rate.

We believe that the best choices for recall rate targets are those that represent the mean recall rates for one of the recall rate groups established in the Results section. By using the metric of AW/ACD, a detriment (additional work-up) to benefit measure of increasing or decreasing recall rate, individual mammographers may be able to gauge the effect that changing their recall rate by 1% would have on performance.

While individual mammographers and informed patients may have different ideas regarding the optimal trade-off, we suggest using an AW/ACD of, at most, 100. This benchmark level seems to be a good choice. In our study, a level larger than 132 would lead to a target recall rate of 12.3%, which is higher than that recommended by American

Table 3

Estimated Performance Measures for Selected Recall Rates for Subsequent Screenings

Recall Rate (%)	Sensitivity (%)	Work-ups per Cancers Detected*	PPV†	AW/ACD‡	Incremental PPV§
3	64.1	9.1	.110	. . .	. . .
4	67.9	11.4	.087	51	.020
5	70.7	13.7	.073	68	.015
6	73.1	15.9	.063	82	.012
7	75.3	18.0	.055	94	.011
8	76.8	20.2	.049	132	.008
9	78.2	22.3	.045	132	.008
10	79.7	24.4	.041	132	.008
11	81.1	26.3	.038	132	.008
12	82.6	28.2	.035	132	.008

* Work-ups per cancers detected = recall rate/(5.15/1000 × sensitivity), assuming 5.15 cancers/1000 mammograms.

† PPV = cancers detected per work-up.

‡ AW/ACD = (difference in number of patients recalled)/(difference in number of cancers detected), where the difference is with the previous row.

§ Incremental PPV = ACD/AW.

Figures 1, 2

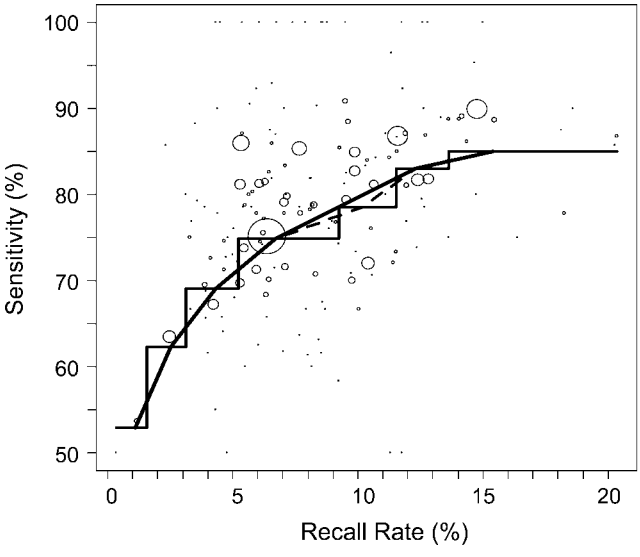

Figure 1: Graph of sensitivity as function of recall rate for subsequent screenings at 172 facilities, with concave (solid curved line) and reduced monotonic regression (step function) fits. Size of circles depicts relative number of cancers from each facility. (Data from three facilities with sensitivities less than 50% are not shown.) Dashed lines show where the piecewise linear fit obtained by joining mean recall rates of adjacent groups differs from concave fit. ○ = individual facility (weighted by number of cancers from that facility).

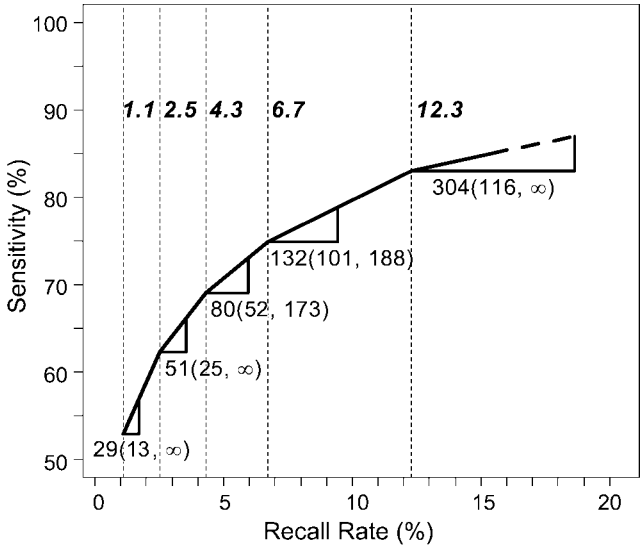

Figure 2: Graph of concave fit for subsequent screenings. Vertical dashed lines show where model changes slope. Corresponding estimated AW/ACD values are shown (represented as base of triangle for a fixed gain in sensitivity), along with 95% confidence intervals.

Figures 3, 4

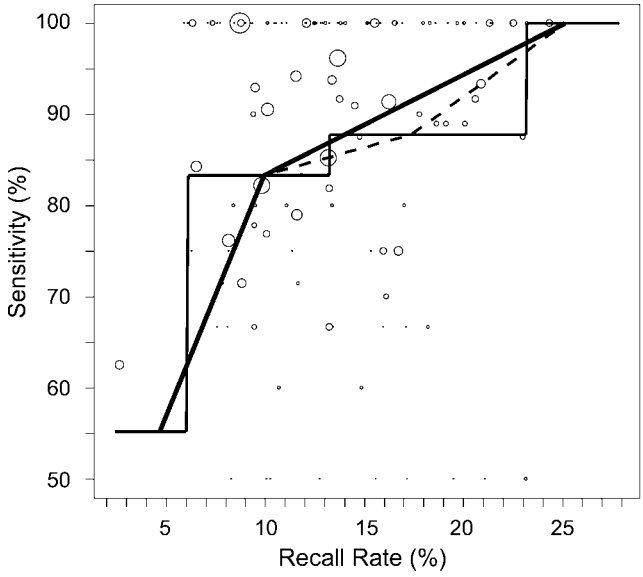

Figure 3: Graph of sensitivity as function of recall rate for first screenings at 139 facilities, with concave (solid curved line) and reduced monotonic regression (step function) fits. Size of circles depicts relative number of cancers from each facility. (Data from six facilities with sensitivities less than 50% are not shown.) Dashed lines show where the piecewise linear fit obtained by joining mean recall rates of adjacent groups differs from concave fit. ○ = individual facility (weighted by number of cancers from that facility).

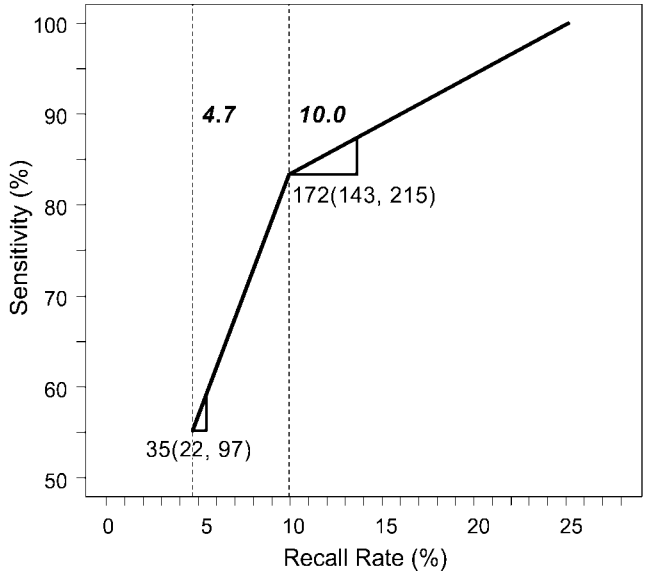

Figure 4: Graph of concave fit for first screenings. Vertical dashed lines show where model changes slope. Corresponding estimated AW/ACD values are shown (represented as base of triangle for a fixed gain in sensitivity), along with 95% confidence intervals.

Table 4

Association between Recall Rate and Sensitivity for First Mammographic Screenings

Recall Rate Group	No. of Facilities	No. of Mammograms*	No. of Cancers	Recall Rate (%)		Sensitivity†	Cancer Detection Rate‡
				Range	Mean		
1	4	5876 (3.4)	31	2.4–6.0	4.7	55.2	2.96
2	67	91 921 (53.7)	471	6.1–13.2	10.0	83.3	4.47
3	62	69 108 (40.4)	383	13.3–23.1	17.4	87.8	4.71
4	6	4199 (2.5)	25	23.2–27.9	25.1	100.0	5.37
Overall	139	171 104 (100)	910	2.4–27.9	13.1	85.9	4.61

* Data in parentheses are percentages.

† Adjusted for site and mean age of woman.

‡ Number of cancers detected per 1000 mammograms, assuming 5.32 cancers per 1000 mammograms, which was the rate observed in our study for first screenings.

College of Radiology guidelines (8,9). A level below 80 would result in a target recall rate of 4.3%. The latter rate is close to European guidelines (6,7), which call for rates below 5%. However, less than 15% of U.S. patients underwent mammography at facilities with rates at or below that rate. The resulting target recall rates are 10.0% for first mammograms and 6.7% for subsequent mammograms. Notably, the recall rate groups with the largest percentages of mammograms evaluated include these target rates, with a small percentage below these rates and a sizable percentage above—42.9% of first and 36.3% of subsequent mammograms. European guidelines of less than 5% for subsequent screenings correspond to a lower tolerance for AW/ACD (6,7). Interpreted according to our concave fit, their recommendations would suggest use of a maximum AW/ACD between 51 and 80, which would lead to a recall rate group with a 4.3% mean recall rate and 69% sensitivity. Their guideline for less than 7% recalls after first screenings would place them in the lower end of the 6.1%–13.2% recall rate group given in Table

4, which corresponds to an 83.3% sensitivity.

Yankaskas et al (19) showed that these differences do exist when comparing international screening recall rates. In a meta-analysis of 24 mammographic screening programs, Elmore et al (1) concluded that "the percentage of mammograms judged to be abnormal in North American programs was 2–4 percentage points higher than it was in programs from other countries without evident benefit in the yield of cancers detected per 1000 women screened, although an increase was noted in DCIS [ductal carcinoma in situ] detection." This may be exaggerated because they included BI-RADS category 3 as a recall, which is not done internationally or in this study.

Yankaskas et al (5) concluded that facilities with recall rates between 4.9% and 5.5% achieved the best trade-off of sensitivity and PPV. Their range of recall rates was obtained by performing reduced monotonic regressions on the relationships of sensitivity and PPV with recall rate and by identifying the range in which both sensitivity and PPV were maintained at high levels. Our study is nearly 10 times larger than theirs and, to our knowledge, is the largest study to date to examine this issue in the United States. We believe that it improves on that study by splitting mammograms into first and subsequent screenings and by using the concave fit approach, with its accompanying AW/ACD concept. Their target rate is between the reasonable targets of 4.3% and 6.7% presented by us.

There is some concern about the establishment of recall rate guidelines in mammography. Gur et al (20) examined the correlation between recall and cancer detection rates in a group of 10 radiologists from a single academic institution. Noting an increase in the presumed linear relationship, where the recall rates ranged from 7.7% to 17.2%, they concluded that "the performance level of a radiologist in the screening environment is a complex, multifactorial issue that cannot and should not be simplified. Reducing recall rates by "decree" (through the enforcement of recommended practice guidelines) may result in a corresponding reduction in the detection rates." Their study results, however, focused on detection rates, which lack the additional information on missed cancers that is available with sensitivity. They also modeled the relationship between the recall and detection rates as linear. Consequently, their evidence cannot suggest that any level of recall rate is too high. Inspection of their data suggests that beyond about a 12% recall rate, little or no gain in the detection rate occurred. In a recent commentary, Hardesty et al (21) suggested usage of three nonlinear models to determine the sensitivity–recall rate relationship, including a concave fit, which they erroneously called convex. We used a concave fit in our analysis, which we believe makes sense clinically because discrimination between cancer and noncancer as the recall rate increases becomes increasingly difficult.

In a re-review of missed cancers from a Dutch breast screening program that has the lowest recall rate worldwide, Otten et al (22) examined the effect of increasing the recall rate. In their article, 15 mammographers rereviewed 495 screening mammograms with negative findings for subsequent screenings, which included 245 missed cancers, from which they extrapolated their outcome measures to 500 000 subsequent screenings by using an equivalent concept to AW/ACD. For recall rates between 4% and 7%, their AW/ACD values were 30%–70% higher than ours. For recall rates between 8% and 10%, their AW/ACD values were within 14% of ours. It is unclear whether the higher values they obtained are because of differences in mammographic evaluation between the United States and the Netherlands, different screening intervals, or the relatively few cancers on which they based their estimates.

Our study had limitations. The relationship between recall rate and sensitivity was assessed by using seven community-based registries in the United States. Extension of these results to other U.S. states and to other regions of the world cannot be assumed from statistical principles but must rather be based on subjective judgment. Because the women included in the BCSC are largely representative of women undergoing screening mammography within the United States, however, we believe it is likely that most radiologists' patients will be similar to those within the BCSC (11). Sensitivity differences were found between the states. We believe that this is most likely because of differences in completeness in identifying cancers between the various state registries. This site difference was adjusted for in obtaining our models, with North Carolina as the referent group because it provided the largest number of mammograms to this study. Thus, the actual sensitivities will differ by state. However, the locations where the AW/ACD rates shifted were modeled as being the same for all sites. As further data are accrued, additional covariates affecting the sensitivities or, perhaps, the location of AW/ACD rate shifts may be found. However, to our knowledge, this study represents the largest collective evidence on the recall rate–sensitivity relationship published to date and incorporates the data from the second-largest study (5).

The data and analysis presented demonstrate a wide variation in recall rate–sensitivity pairs in this convenience sample of U.S. radiologists. This variation likely represents both different preferences and different abilities among radiologists. Clustering of the facilities' results, together with analysis of the gains from additional work-up for these radiologists, strongly suggests ranges of targeted performance. We recommend operating at a target recall rate of approximately 6.7% for subsequent mammograms and 10.0% for first mammograms, because these rates keep the estimated number of AW/ACD lower than 100.

Acknowledgments: We acknowledge the statistical analysis and graphic presentation assistance of Li Lin, MS, and the secretarial assistance of Jane Beley.

References

1. Elmore JG, Nakano CY, Koepsell TD, Desnick LM, D'Orsi CJ, Ransohoff DF. International variation in screening mammogra-

phy interpretations in community-based programs. J Natl Cancer Inst 2003;95:1384–1393.

2. 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.

3. Smith-Bindman R, Ballard-Barbash R, Miglioretti DL, Patnick J, Kerlikowske K. Comparing the performance of mammography screening in the USA and the UK. J Med Screen 2005;12:50–54.

4. Barlow WE, Chi C, Carney PA, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst 2004;96:1840–1850.

5. Yankaskas BC, Cleveland RJ, Schell MJ, Kozar R. Association of recall rates with sensitivity and positive predictive values of screening mammography. AJR Am J Roentgenol 2001;177:543–549.

6. Recommendations on cancer screening in the European union. Advisory Committee on Cancer Prevention. Eur J Cancer 2000;36:1473–1478.

7. Roselli del Turco M, Hendriks JH, Perry NM. Radiological guidelines. In: Perry NM, Broeder M, de Wolf CJM, Tornberg S, eds. European guidelines for quality assurance in mammography screening. Luxembourg: Office for Official Publications of the European Communities, 2001; 366.

8. Quality determinants of mammography. Quality Determinants of Mammography Guidelines Panel. Rockville, Md: United States Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research, 1994; 78–86.

9. Feig SA, D'Orsi CJ, Hendrick RE, et al. American College of Radiology guidelines for breast cancer screening. AJR Am J Roentgenol 1998;171:29–33.

10. 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.

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

12. Carney PA, Geller BM, Moffett H, et al. Current medicolegal and confidentiality issues in large, multicenter research programs. Am J Epidemiol 2000;152:371–378.

13. American College of Radiology. ACR BI-RADS mammography. In: ACR Breast Imaging Reporting and Database System, Breast Imaging Atlas. 4th ed. Reston, Va: American College of Radiology, 2003.

14. Robertson T, Wright FT, Dykstra RL. Order restricted statistical inference. New York, NY: Wiley, 1988.

15. Schell MJ, Singh B. The reduced regression method. J Am Stat Assoc 1997;92:128–135.

16. Pinckney RG, Geller BM, Burman M, Littenberg B. Effect of false-positive mammograms on return for subsequent screening mammography. Am J Med 2003;114:120–125.

17. Barton MB, Moore S, Polk S, Shtatland E, Elmore JG, Fletcher SW. Increased patient concern after false-positive mammograms: clinician documentation and subsequent ambulatory visits. J Gen Med Intern 2001;16:150–156.

18. Brett J, Bankhead C, Henderson B, Watson E, Austoker J. The psychological impact of mammographic screening: a systematic review. Psychooncology 2005;14:917–938.

19. Yankaskas BC, Klabunde CN, Ancelle-Park R, et al. International comparison of performance measures for screening mammography: can it be done? J Med Screen 2004;11:187–193.

20. Gur D, Sumkin JH, Hardesty LA, et al. Recall and detection rates in screening mammography: a review of clinical experience—implications for practice guidelines. Cancer 2004;100:1590–1594.

21. Hardesty LA, Klym AH, Shindel EE, Chough DM, Sumkin JH, Gur D. Is maximum positive predictive value a good indicator of an optimal screening mammography practice? AJR Am J Roentgenol 2005;184:1505–1507.

22. Otten JD, Karssemeijer N, Hendriks JH, et al. Effect of recall rate on earlier screen detection of breast cancers based on the Dutch performance indicators. J Natl Cancer Inst 2005;97:748–754.