Performance Benchmarks for Screening Mammography[1]

Robert D. Rosenberg, MD
Bonnie C. Yankaskas, PhD
Linn A. Abraham, MS
Edward A. Sickles, MD
Constance D. Lehman, MD, PhD
Berta M. Geller, EdD
Patricia A. Carney, PhD
Karla Kerlikowske, MD
Diana S. M. Buist, PhD
Donald L. Weaver, MD
William E. Barlow, PhD
Rachel Ballard-Barbash, MD, MPH

Purpose: To retrospectively evaluate the range of performance outcomes of the radiologist in an audit of screening mammography by using a representative sample of U.S. radiologists to allow development of performance benchmarks for screening mammography.

Materials and Methods: Institutional review board approval was obtained, and study was HIPAA compliant. Informed consent was or was not obtained according to institutional review board guidelines. Data from 188 mammographic facilities and 807 radiologists obtained between 1996 and 2002 were analyzed from six registries from the Breast Cancer Surveillance Consortium (BCSC). Contributed data included demographic information, clinical findings, mammographic interpretation, and biopsy results. Measurements calculated were positive predictive values (PPVs) from screening mammography (PPV_1), biopsy recommendation (PPV_2), biopsy performed (PPV_3), recall rate, cancer detection rate, mean cancer size, and cancer stage. Radiologist performance data are presented as 50th (median), 10th, 25th, 75th, and 90th percentiles and as graphic presentations by using smoothed curves.

Results: There were 2 580 151 screening mammographic studies from 1 117 390 women (age range, <30 to ≥80 years). The respective means and ranges of performance outcomes for the middle 50% of radiologists were as follows: recall rate, 9.8% and 6.4%–13.3%; PPV_1, 4.8% and 3.4%–6.2%; and PPV_2, 24.6% and 18.8%–32.0%. Mean cancer detection rate was 4.7 per 1000, and the mean size of invasive cancers was 13 mm. The range of performance outcomes for the middle 80% of radiologists also was presented.

Conclusion: Community screening mammographic performance measurements of cancer outcomes for the majority of radiologists in the BCSC surpass performance recommendations. Recall rate for almost half of radiologists, however, is higher than the recommended rate.

Supplemental material:
radiology.rsnajnls.org/cgi/content/full/241/1/55/DC1

[1] From the Department of Radiology, University of New Mexico Health Sciences Center, MSC10 5530, 1 University of New Mexico, Albuquerque, NM 87131 (R.D.R.); Department of Radiology, University of North Carolina, Chapel Hill, NC (B.C.Y.); Group Health Center for Health Studies, Seattle, Wash (L.A.A., D.S.M.B., W.E.B.) and Cancer Research and Biostatistics (W.E.B.), Seattle, Wash; Department of Radiology, University of California San Francisco School of Medicine, San Francisco, Calif (E.A.S.); Department of Radiology, University of Washington Medical Center, Seattle, Wash (C.D.L.); Office of Health Promotion Research (B.M.G., D.L.W.), Department of Pathology (D.L.W.), and Vermont Cancer Center (B.M.G., D.L.W.), University of Vermont, Burlington, Vt; Department of Community and Family Medicine, Dartmouth Medical School, Hanover, NH (P.A.C.); General Internal Medicine Section, Departments of Veterans Affairs and Epidemiology and Biostatistics, University of California, San Francisco, Calif (K.K.); and Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, Md (R.B.). Received September 7, 2005; revision requested November 10; revision received January 7, 2006; final version accepted February 1. Supported by cooperative grants U0169976 (R.D.R.), U01CA70040 (B.C.Y.), U01CA63740 (E.A.S., K.K.), U01CA70013 (B.M.G., D.L.W.), U01CA86076 (L.A.A., W.E.B.), U01CA63731 (D.S.M.B.), and U01CA86082-01 (P.A.C.) from the National Cancer Institute. **Address correspondence to** R.D.R. (e-mail: *rrosenb@unm.edu*).

The Mammography Quality Standards Act of 1992 was created to improve patient outcomes from mammography (1). The legislation set minimum national standards for performance of mammography at the technical level and many professional requirements. Minimum training and continuing education requirements were established for the technologist, radiologist, and medical physicist. In addition, a requirement was established for an annual audit for each mammographic facility and each radiologist.

A medical audit is a compilation of specific important patient outcomes over a defined period of at least a year. This allows a radiologist and a facility to recognize areas of strength, as well as those areas that may need improvement. The medical audit is recognized as one of the best quality assurance tools (2,3). There are limited data available, however, with which the practicing radiologist and facility may compare their results. There is a lack of generalizable literature concerning the actual performance of radiologists in the United States and, thus, a limited knowledge of optimal performance targets that are achievable by general radiologists. The opinion of experienced radiologists (4, p 83) and guideline targets for performance for some parameters have been set in some countries (5, pp 4–5;6, pp 147–148). There are problems, however, with using these data for U.S. radiologists. These targets have not been quantified within the U.S. health care environment because of a lack of appropriate population-based screening data. Therefore, the value of existing guidelines is limited.

The Breast Cancer Surveillance Consortium (BCSC) is a National Cancer Institute–funded research initiative of seven population-based research sites with a Statistical Coordinating Center that collects and analyzes mammographic and pathologic data in defined populations (7). The BCSC has published data on its methods (8), confidentiality issues (9), and overall community performance (10–13). These prior results were based on a population of patients that has characteristics that are similar to the national demographic characteristics in terms of age, ethnicity, and urban or rural residence (11). For comparison purposes, use of BCSC data presents two barriers for the average radiologist: *(a)* The methods used in prior publications cannot be applied by the average community radiologist. *(b)* BCSC results represent averages rather than the distribution of the range of performance outcomes. The key methodologic limitation is that most community radiology groups do not have the ability to link their mammographic data to regional cancer registries. In addition, the large variations in measurements of performance of mammographers that have been extensively documented (14–19) underscore the importance of understanding where individual performance lies within the distribution of performance of other radiologists.

The American College of Radiology has created medical audit methods intended for use by community radiologists. The 4th edition of the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) manual (20) provides standardized terminology and assessments used in breast imaging for mammography, ultrasonography, and magnetic resonance imaging. In addition, the BI-RADS manual contains instructions on the use of these assessments to compute outcome measurements from mammographic data that are possible for many community practices. The measurements proposed by the American College of Radiology are more extensive than the minimal measurements required by the Mammography Quality Standards Act for mammographic accreditation. The purpose of our study was to retrospectively evaluate the range of individual radiologist performance outcomes in an audit of screening mammography by using a representative group of U.S. radiologists to allow development of performance benchmarks for screening mammography.

Advances in Knowledge

- There is a range of performance outcomes, covering the full complement of outcome measurements recommended in BI-RADS fourth edition, for screening mammography performed by a large sample of U.S. radiologists in a representative population of women.
- For performance outcome measurements related to cancer detection and stage at diagnosis, most radiologists exceed the Agency for Health Care Policy and Research (AHCPR) desirable goals for experts in the performance of screening mammography.
- For performance outcome measurements related to recall rate and positive predictive value, most radiologists are just at or below the AHCPR desirable goals for experts in the performance of screening mammography.

Materials and Methods

The parameters examined and the methods used for estimating performance measurements are described in the medical audit section of the 4th edition of the American College of Radiology BI-RADS manual (20). All authors were involved in decisions about the methods and in the interpretation of the results of the analysis. Analysis was per-

Published online
10.1148/radiol.2411051504

Radiology 2006; 241:55–66

Abbreviations:
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, R.D.R.; 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, R.D.R., B.C.Y., E.A.S., P.A.C., K.K.; clinical studies, K.K.; statistical analysis, L.A.A., K.K., W.E.B.; and manuscript editing, B.C.Y., L.A.A., E.A.S., C.D.L., B.M.G., P.A.C., K.K., D.S.M.B., D.L.W., W.E.B., R.B.

Authors stated no financial relationship to disclose.

formed by an individual (L.A.A.) in consultation with three other individuals (R.D.R., B.C.Y., W.E.B.) by using software (SAS; SAS Institute, Cary, NC). These methods focus on follow-up of patients with an abnormal mammographic interpretation and their associated pathologic findings. We limited follow-up for cancer to 12 months after screening mammography. Although the BCSC uses computerized matching systems to link mammograms to breast cancers, most clinical practices do not have this capacity and, therefore, sensitivity and specificity are not included in this analysis. Most guidelines for screening from other countries have defined separate targets for initial and subsequent screening mammography. With current data systems in use in many practices, however, it is not generally feasible to perform audits separately according to initial and subsequent examinations, and therefore performance outcomes for initial and subsequent mammographic examinations were combined. Ninety percent of the mammograms were from subsequent examinations.

Data Sources

Data were collected from six BCSC registries: Carolina Mammography Registry (Chapel Hill, NC), Group Health Cooperative (Seattle, Wash), New Hampshire Mammography Network (Lebanon, NH), New Mexico Mammography Project (Albuquerque, NM), Vermont Breast Cancer Surveillance System (Burlington, Vt), and San Francisco Mammography Registry (San Francisco, Calif). To determine cancer outcomes, each registry links its data to a state tumor registry or to a Surveillance Epidemiology and End Result, also known as SEER, program. Six of these registries also collect some benign pathologic results (7).

Each registry and the Statistical Coordinating Center of the BCSC have received a Federal Certificate of Confidentiality and approval from each institutional review board for the protection of human subjects to collect and send data to the Statistical Coordinating Center 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. Our study was Health Insurance Portability and Accountability Act compliant. All registries have strict procedures for deidentification of patient information and protection of confidentiality (9). Linkage procedures follow protocols specifically designed to preserve patient confidentiality.

Data Collected

Approximately 188 mammographic facilities contributed to the pooled data. This number of facilities represents about 2% of the approximately 10 000 Food and Drug Association–certified mammographic facilities in the United States in 2000. We compared the demographic makeup of the population living in the catchment areas of the six BCSC registries included in our study to that of the entire U.S. population by using 2000 census data. To describe the BCSC population, we (L.A.A., W.E.B.) included census data from all counties in which there was a participating mammographic facility.

Study Group

The study included women who had undergone at least one screening mammographic examination during the years 1996–2002. Screening mammographic examinations performed after December 2002 were excluded to ensure that there was at least 12 months following the screening examination during which cancer could be diagnosed and there was adequate time for cancer reporting. A screening mammographic examination was defined as one characterized by the interpreting radiologist as having an indication of screening.

The pooled data contain screening mammographic interpretations determined by 807 identified radiologists. A radiologist identifier was not available from some facilities but was present for 84.0% (2 166 970 of 2 580 151) of the studies in this report. Some radiologists contributed data from multiple facilities. Many radiologists also interpreted some mammograms at facilities outside of the consortium, and therefore only a subset of their interpretations would have been captured. This inclusion of only a subset of their interpretations also occurs because radiologists move between facilities or serve as temporary radiologists in a facility.

Mammographic Data Collection Procedures and Definitions

Across all BCSC registries, patients undergoing mammography complete a questionnaire at each imaging visit that requests medical history and demographic data, including date of most recent mammographic examination, family history of breast cancer, previous breast biopsy, personal history of breast cancer, and description of recent breast symptoms. Women were considered to have a family history of breast cancer if they reported having at least one female first-degree relative (mother, sister, or daughter) with breast cancer. Women were considered to have a personal history of breast cancer if they had self-reported previous breast cancer or had evidence of previous breast cancer in the cancer registry or pathology database. Each woman was considered to have a previous mammographic examination if she had a self-reported prior mammographic examination or there was indication of information about a prior mammographic examination in the BCSC database.

Generally, screening mammography is performed for women without breast symptoms, but some women with symptoms are included in all screening populations (10,21,22). In this analysis, we used mammograms identified as screening mammograms by the interpreting radiologist independent of whether or not symptoms were present at the time of the examination. We included mammograms that are variably considered screening mammograms that had been obtained because of other special cases, and these mammograms included those obtained in patients with breast implants and in patients with prior breast cancer if the mammograms were designated as from a screening study.

The mammographic registry also captures data about image interpretation, including management (imaging, biopsy, and clinical evaluations) recommendations and the BI-RADS assessment categories assigned by the interpreting radiologist for each mammographic examination (9,20). A separate assessment often is recorded for each breast. For the purpose of this study, we created an overall assessment for the entire examination by using the more serious abnormal BI-RADS assessment category according to the following hierarchy: negative (category 1), benign (category 2), probably benign (category 3), needs additional evaluation (category 0), suspicious (category 4), and highly suggestive of malignancy (category 5). A positive result was defined as one classified with BI-RADS assessment categories 0, 4, or 5, and a negative result was defined as one classified with BI-RADS assessment categories 1, 2, or 3. Results in a previously published investigation (23) showed only very small nonsignificant differences between woman-specific and breast-specific outcome data, and these results indicated that woman-specific data are sufficiently accurate measurements of interpretive performance.

A report about screening mammography from the BCSC (24) indicated that 10%–15% of examinations with positive (abnormal) results (BI-RADS categories 0, 4, or 5) were discordant between the BI-RADS assessment category assigned and subsequent management recommendations provided by the interpreting radiologist compared with the recommendations that the BI-RADS assessment category should inherently suggest. These nonstandard approaches tend to undercount a sizable proportion of positive mammograms (25). This undercount is caused by the common use of BI-RADS category 3 (probably benign finding) with additional imaging recommended instead of BI-RADS 0 (needs additional imaging) (24). Thus, a negative assessment (BI-RADS category 3) is used instead of a positive assessment (BI-RADS category 0). Because of the differences in how practicing radiologists implement BI-RADS, an important percentage of women with similarly abnormal mammographic findings may appear to be classified in different categories of assessment. To create comparable performance benchmarks across facilities, we made two modifications to the collected data: *(a)* If additional imaging was performed at the time of the screening, the screening mammogram was considered positive. *(b)* If a recommendation for immediate work-up was given along with an assessment that indicated a probably benign assessment (BI-RADS category 3), then the assessment was considered positive and classified as BI-RADS category 0. According to BI-RADS audit rules, any mammogram with a BI-RADS category 6 assessment (known breast cancer) was excluded from the analysis.

Patients undergoing mammography were considered to have breast cancer if a state tumor registry, Surveillance Epidemiology and End Result program registry, or pathology database indicated the diagnosis of invasive carcinoma or ductal carcinoma in situ within 12 months after a screening mammographic examination.

Table 1

Demographic Characteristics for the Study Compared with Those for the Entire U.S. Population

Characteristic	Study Population*	U.S. Population†
Total population in selected counties	11 874 535	281 421 906
Rural-urban mix (%)		
Rural	23.0	21.0
Urban	77.0	79.0
Race (%)‡		
White	82.7	84.9
African American	9.7	10.8
Other	7.5	4.3
Hispanic ethnicity (%)	6.3	7.3
No high school degree (%)§	16.0	19.6
Economic status		
Living in poverty (%)	11.2	12.4
Unemployed (%)	3.7	4.0
Median family income ($)	53 933	51 197

* Data were based on 2000 census data for all countries in which there was a mammographic facility that contributed data to this study.

† Data were based on 2000 census data for the entire U.S. population.

‡ For women 40 years of age and older.

§ For women 25 years of age and older.

Outcome Measurements and Statistical Analysis

A true-positive mammogram was defined as a screening mammographic examination with a positive interpretation that was followed by the diagnosis of invasive breast cancer or ductal carcinoma in situ within 12 months. Cancer detection rate was defined as the number of cancers following a positive mammogram divided by the total number of screening mammographic examinations. Conversely, a false-positive mammogram was defined as a screening mammographic examination with a positive interpretation and no breast cancer diagnosed within the next 12 months.

We calculated the positive predictive value (PPV) by dividing the number of true-positive examinations by the sum of true-positive and false-positive examinations. Three separate PPV calculations were performed by using BI-RADS methods: PPV_1 (probability of cancer following a positive mammographic interpretation), PPV_2 (probability of cancer following a BI-RADS assessment of 4 or 5), and PPV_3 (probabil-

ity of cancer among patients actually undergoing biopsy after a BI-RADS assessment of 4 or 5). For screening examinations with an initial BI-RADS assessment of category of 0, the final assessment was determined by looking ahead 180 days to determine whether additional imaging had been performed. Final assessment was used when PPV_2 and PPV_3 were computed. A final BI-RADS assessment of category 4 or 5 was assumed to be a biopsy recommendation. PPV_3 included the performance of any type of biopsy (fine-needle aspiration, cyst aspiration, core, or surgical biopsy). PPV_2 and PPV_3 are both important, as they are measurements of different aspects of the process; PPV_2 is a measurement of PPV for biopsy recommendations, whereas PPV_3 is a measurement for biopsies actually performed.

Because few mammographic facilities have adequate resources to estimate sensitivity or specificity, we report those calculations only on the BCSC Web site at *http://breastscreening.cancer.gov*.

Simple descriptive statistics (frequency, percentile, mean, and median values) were chosen to provide clinically relevant screening performance benchmarks. We illustrated the variability found among radiologists by using percentile values to indicate ranges that describe where the middle 50% and 80% of performance outcomes was found for specific outcome measurements. For example, the combination of 25th and 75th percentile values defines the range within which the middle 50% of performance outcomes was found, and the combination of 10th and 90th percentile values defines the range within which the middle 80% of performance outcomes was found. To reduce the amount of random statistical variation in these data, we reported outcomes from only those radiologists who contributed at least a designated, subjectively determined minimum number of mammographic examinations or cancers for each outcome displayed, as follows: recall rate and cancer detection rate, 1000 examinations; PPV_1, 100 abnormal interpretations; PPV_2, 30 biopsy recommendations; PPV_3, 30 biopsies

Table 2

Clinical Demographic Characteristics for 2 580 151 Screening Mammographic Examinations

Characteristic	No. of Examinations*
Age (y)†	
<30	3564 (0.1)
30–39	121 730 (4.7)
40–49	754 830 (29.3)
50–59	746 272 (28.9)
60–69	493 841 (19.1)
70–79	352 075 (13.6)
≥80	107 839 (4.2)
Family history of breast cancer	
Yes	323 186 (15.2)
No	1 807 081 (84.8)
Unknown	449 884 (17.4)
Personal history of breast cancer	
Yes	114 557 (6.3)
No	1 718 273 (93.7)
Unknown	747 321 (29.0)
Previous mammogram reported	
Yes	2 161 902 (89.2)
No	262 224 (10.8)
Unknown	156 025 (6.0)
Self-reported symptoms‡	
Yes	85 049 (3.6)
No	2 280 740 (96.4)
Unknown	214 362 (8.3)

* Numbers in parentheses are percentages calculated based on nonmissing values. Most of the unknown data are structurally missing, and using nonmissing data provides the best representation of each demographic characteristic.

† The mean age was 56.4 years, and the median age was 54.0 years.

‡ Self-reported symptoms included lump, discharge, and other symptoms but not pain.

Table 3

Abnormal Interpretations for 2 580 151 Screening Mammographic Examinations

Measurement and Data	Value
Recall rate (%)	9.8
No. of abnormal interpretations	253 169
Total no. of examinations	2 580 151
PPV_1, abnormal interpretations (%)*	4.8
No. of cancers	12 068
No. of abnormal interpretations	253 169
PPV_2, biopsy recommended (%)†	24.6
No. of cancers	9342
No. of abnormal interpretations	37 987
PPV_3, biopsy performed (%)‡	33.8
No. of cancers	8901
No. of abnormal interpretations	26 340

* An abnormal interpretation was based on assignment of BI-RADS category 3 (only when immediate work-up is recommended) or 0, 4, or 5 or performance of additional imaging on the same day as the screening mammographic examination.

† A classification of biopsy recommended was based on the assignment of BI-RADS category 4 or 5 at the final assessment.

‡ A classification of biopsy recommended and performed was based on assignment of BI-RADS category 4 or 5 at the final assessment and availability of biopsy results.

performed; and for cancer measurements, 15 cancers with complete information on the outcome criteria. We used graphic presentations (frequency distributions overlaid with percentile values) to display these data in an easily understandable format and present the tabular data in the BCSC Web site. More complex analytic methods, such as those designed to elucidate statistically significant interactions among the data variables collected, are beyond the scope of our study.

Results

Demographic Factors

During the 1996–2002 study period, the six participating BCSC registries contributed 2 580 151 screening mammographic examinations for 1 117 390 women. The demographic makeup of the population living in the catchment areas of the six BCSC sites included in our study is comparable to that for the population of the entire United States (Table 1). There are only slight differences, with none greater than 5 percentage points, between our study population and the U.S. population. Our study population is slightly more rural, contains slightly fewer African Ameri-

Table 4

Cancers for 2 580 151 Screening Mammographic Examinations

Cancer Data	No. of Cancers
Cancer histologic type*	
Ductal carcinoma in situ	2603 (21.6)
All invasive	9465 (78.4)
Invasive cancer size (mm)†	
1–5	882 (10.2)
6–10	2333 (27.0)
11–15	2309 (26.7)
16–20	1293 (14.9)
>20	1839 (21.2)
Unknown	809 (8.5)
Minimal cancer‡	5818 (51.7)
Axillary lymph node status§	
Negative	7233 (79.8)
Positive	1829 (20.2)
Unknown	403 (4.3)‖
Cancer stage#	
0	2603 (25.1)
I	5446 (50.5)
II	2380 (21.1)
III	257 (2.4)
IV	98 (0.9)
Unknown	1284 (9.5)

Note.—The number of all cancers was 12 068, and the mean cancer detection rate per 1000 was 4.7.

* Numbers in parentheses are percentages of all cancers where cancer histologic type was known.

† Numbers in parentheses are percentages of invasive cancers of known size. The mean size was 16.4 mm, and the median size was 13 mm.

‡ Defined as cases of ductal carcinoma in situ or invasive cancer of 10 mm or smaller. Numbers in parentheses are percentages of cases of ductal carcinoma in situ and invasive cancers of known size.

§ Numbers in parentheses are percentages of invasive cancers of known nodal status.

‖ The denominator is 9465.

Numbers in parentheses are percentages of cases of ductal carcinoma in situ and invasive cancers of known stage.

Table 5

Performance Benchmarks for Abnormal Screening Mammographic Interpretations

Measurement and Data	Value
Recall rate (%)	9.4
No. of readers with ≥1000 examinations	344
No. of abnormal interpretations	195 697
Total no. of examinations	2 076 379
Reader performance (%)	
For 50th percentile (median)	9.7
For 10th–90th percentiles	4.4–16.8
For 25th–75th percentiles	6.4–13.3
PPV_1, abnormal interpretation (%)*	4.8
No. of readers with ≥100 abnormal interpretations	330
No. of cancers	9451
No. of abnormal interpretations	195 591
Reader performance (%)	
For 50th percentile (median)	4.5
For 10th–90th percentiles	2.6–8.6
For 25th–75th percentiles	3.4–6.2
PPV_2, biopsy recommended (%)†	25.0
No. of readers with ≥30 biopsy recommendations	256
No. of cancers	6991
No. of biopsy recommendations	27 947
Reader performance (%)	
For 50th percentile (median)	25.0
For 10th–90th percentiles	14.1–38.8
For 25th–75th percentiles	18.8–32.0
PPV_3, biopsy performed (%)‡	32.6
No. of readers with ≥30 biopsies performed	195
No. of cancers	6173
No. of biopsy recommendations	18 948
Reader performance (%)	
For 50th percentile (median)	32.3
For 10th–90th percentiles	20.6–51.0
For 25th–75th percentiles	25.0–40.5

Note.—Data include examinations for radiologists with minimum numbers of mammograms as designated; examinations for which the radiologist was unknown were excluded.

* An abnormal interpretation was defined as one for which BI-RADS category 3 (only when immediate work-up is recommended) or 0, 4, or 5 was assigned or in which additional imaging was performed on the same day as the screening mammographic examination.

† A classification of biopsy recommended was defined as one based on the assignment of BI-RADS category 4 or 5 at the final assessment.

‡ A classification of biopsy recommended and performed was defined as assignment of BI-RADS category 4 or 5 at the final assessment and availability of biopsy results.

can and Hispanic women, is slightly more highly educated, and has a slightly higher estimated median family income than the entire U.S. population. The population of patients and radiologists cannot be a random sample since the data are voluntarily provided by radiology facilities. The radiologists included in this study are thought to mirror community practice, as they come from urban, rural, large, small, and different organizational structures of practice across the broad geographic areas.

The percentage of screening mammograms obtained in patients with a report of a prior mammogram is 89.2% (2 161 902 of 2 424 126 [nonmissing data]) (Table 2). In this data set, 41.1% (459 324 of 1 117 390) of the women underwent only one screening examination, 22.5% (250 842 of 1 117 390) underwent two screening examinations, 15.6% (174 083 of 1 117 390) underwent three screening examinations, and 20.9% (233 141 of 1 117 390) underwent four or more screening examinations.

Researchers in previous reports (12,26–31) have shown that clinical outcomes for screening mammography are affected by several common demographic factors, specifically age, family history of breast cancer, personal history of breast cancer, breast density, and mammography performed previously. Because these factors vary by facility, data for these factors are presented for our study population (Table 2). A considerable percentage (based on nonmissing data) of screening mammograms were obtained in women who reported a family history of breast cancer (15.2% [323 186 of 2 130 267]), a personal history of breast cancer (6.3% [114 557 of 1 832 830]), and recent breast symptoms (3.6% [85 049 of 2 365 789]). The most common symptoms were other or not otherwise specified (1.7% [40 226 of 2 365 789]), a lump (1.6% [37 358 of 2 365 789]), and nipple discharge (0.3% [7465 of 2 365 789]). A small percentage of screening mammograms (1.0% [25 651 of 2 580 151]) were obtained in women who reported that they had breast implants. Implant information was missing from 24.4% (630 508 of 2 580 151) of the mammographic data. The mean age of women was 56.4 years, and 4.9% (125 294 of 2 580 151) of the screening mammograms were obtained in women younger than 40 years. The majority of screening examinations (77.3% [1 994 943 of 2 580 151]) were performed in women within the typical screening age range of 40–69 years.

Figure 1

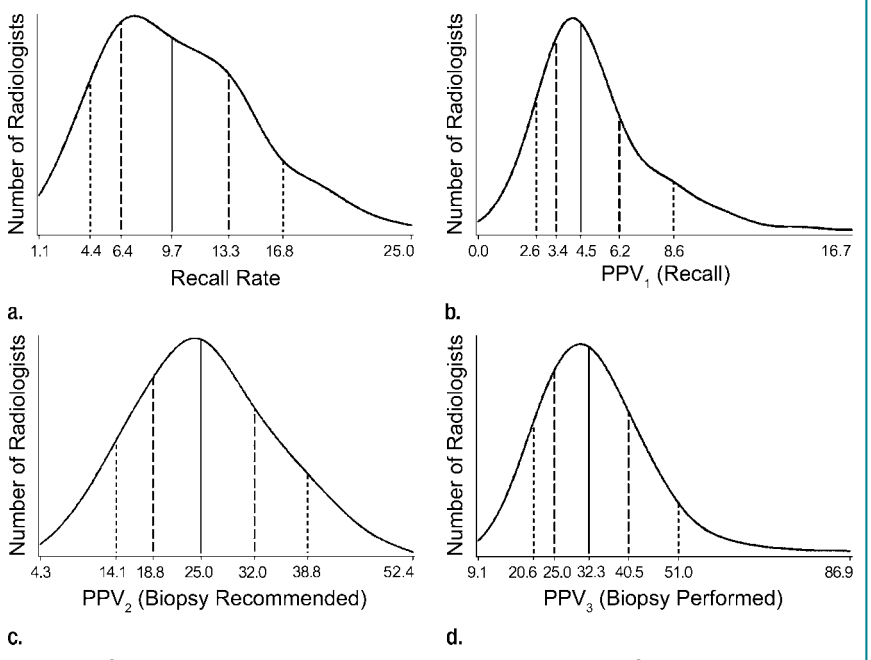

Figure 1: Smoothed plots of frequency distributions for recall rates and PPVs. Overlaid solid line indicates 50th percentile (median), paired dashed lines indicate 25th and 75th percentiles, and paired dotted lines indicate 10th and 90th percentiles. **(a)** Recall rate for 2 076 379 examinations among 344 radiologists with 1000 or more examinations. **(b)** PPV_1 for 195 591 abnormal interpretations among 330 radiologists with 100 or more abnormal interpretations. **(c)** PPV_2 for 27 947 abnormal interpretations among 256 radiologists with 30 or more biopsy recommendations. **(d)** PPV_3 for 18 948 abnormal interpretations among 195 radiologists with 30 or more biopsies performed. The corresponding mean values for all 2 580 151 examinations are in Table 3.

Mammographic Performance Measurements

The recall rate was 9.8% (253 169 of 2 580 151). PPV_1 (percentage of cancers determined after a positive screening examination) was 4.8% (12 068 of 253 169), PPV_2 (percentage of cancers determined after a BI-RADS assessment category of 4 or 5 was assigned) was 24.6% (9342 of 37 987), and PPV_3 (percentage of cancers determined after a BI-RADS assessment category of 4 or 5 was assigned and a biopsy was performed) was 33.8% (8901 of 26 340) (Table 3). Prior to reclassification of some otherwise negative or benign examinations as BI-RADS category 0, such as when additional imaging was performed at the time of the screening examination (to increase consistency of BI-RADS assessments between radiologists), the recall rate was 7.5% (193 265 of 2 580 151) and the PPV_1 was 6.0% (11 560 of 193 265).

Cancer Outcomes

Cancer stage at diagnosis is an important prognostic parameter and currently is most often described by the staging classification schema of the American Joint Committee on Cancer (32). The staging schema of the American Joint Committee on Cancer integrates clinical data on tumor size, nodal involvement, and metastases. Cancers are also characterized by these factors individually, as well as by other summary stage measurements such as "minimal cancer."

The percentage of all cancers diagnosed as ductal carcinoma in situ was 21.6% (2603 of 12 068) (Table 4). Of the invasive cancers with known size, 37.1% (3215 of 8656) were 10 mm or smaller, and 21.2% (1839 of 8656) were larger than 2 cm. The median size was 13 mm, and the mean size was 16.4 mm. The percentage of cancers considered minimal (cases of ductal carcinoma in situ or invasive cancer of 10 mm or smaller) was 51.7% (5818 of 11 259 [known size]) (Table 4).

Node-positive cancers represented 20.2% (1829 of 9062) of all invasive cancers with known nodal status. Summary stage calculation for cancers with known stage yielded 75% (8208 of 10 943) as stage 0 or I. The percentage of cancers for which information was insufficient to calculate stage was 9.3% (1125 of 12 068), primarily because, in 809 (8.5%) of 9465 invasive cancers, size was unknown (Table 4).

Performance Benchmarks

The range of recall rate of the middle 50% of radiologists was 6.4%–13.3%, and that of 80% of radiologists was 4.4%–16.8% (Table 5, Fig 1). The range of PPV_1 of the middle 50% of radiologists was 3.4%–6.2%, and that of 80% of radiologists was 2.6%–8.6%. The range of PPV_2 (BI-RADS assessment 4 and 5) was 18.8%–32.0% for the middle 50% of radiologists, and the range for 80% of radiologists was 14.1%–38.8%. Most of the smoothed curves for performance outcomes (Figs 1, 2) have clearly defined peaks, except recall rate, where there is a flattening around the average of performance outcomes.

Discussion

Background Information

As initially envisioned, the audit functioned as a teaching tool and summary for each radiologist. Subsequently, we used the opinion of experienced radiologists to create targets for performance (4, p 83;5;6, pp 147–148) by using measurements such as recall rate and cancer detection rate. Many of the early performance targets were developed on the basis of the evaluation of outcomes from small groups of radiologists with a special interest in breast imaging (4, p 83;26;33–35).

Publication of our results follows that of the recent Institute of Medicine report of 2005 (36, p 5). That report recommends adding many of the measurements reported in our study to the mammographic facility audit requirements.

Unlike European screening programs, performance targets published in the United States have not been used to enforce performance outcomes. Just recently, estimates of performance outcomes of diagnostic mammography in a large group of radiologists in the United States provided

Figure 2

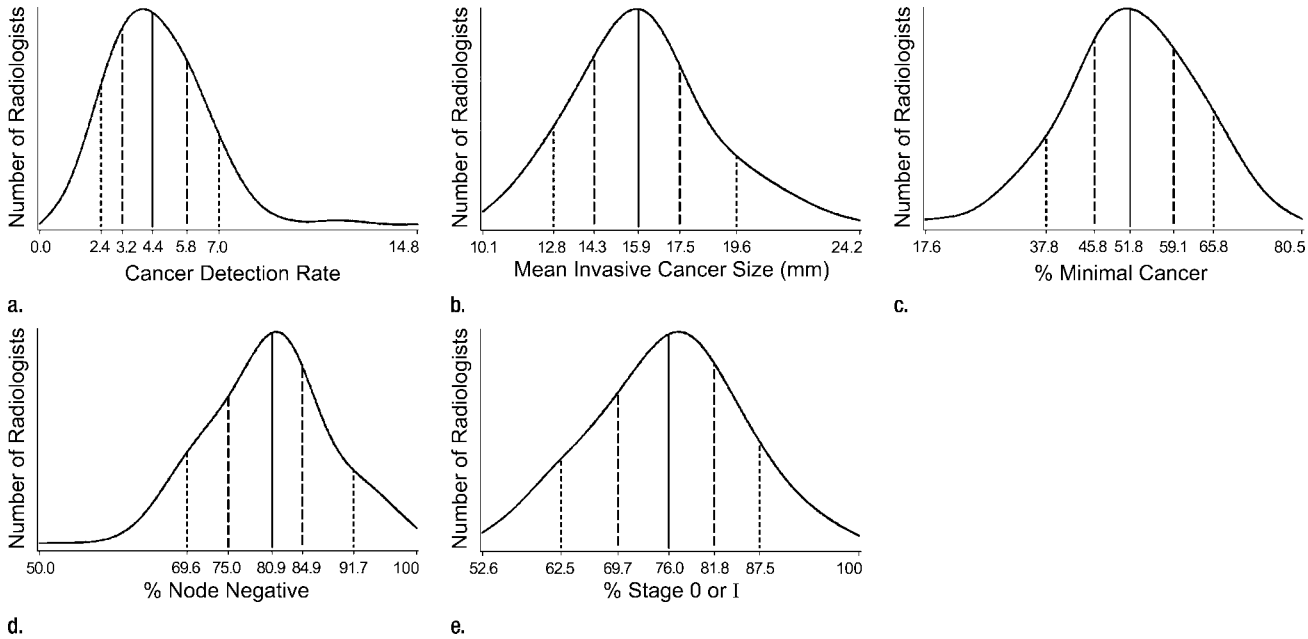

Figure 2: Smoothed plots of frequency distributions for detected cancers. Overlaid solid line indicates 50th percentile (median), paired dashed lines indicate 25th and 75th percentiles, and paired dotted lines indicate 10th and 90th percentiles. **(a)** Cancer detection rate for 2 076 379 examinations among 344 radiologists with 1000 or more examinations. **(b)** Mean cancer size for 5424 invasive cancers with known size among 159 radiologists with detection of 15 or more invasive cancers with known size. **(c)** Percentage of minimal cancer for 7689 cancers among 190 radiologists with detection of 15 or more cancers. **(d)** Percentage of node-negative cancers for 5753 invasive cancers with known nodal status among 169 radiologists with detection of 15 or more invasive cancers with known nodal status. **(e)** Percentage of stage 0 or I cancers for 7268 cancers with known stage among 182 radiologists with detection of 15 or more cancers with known stage. The corresponding mean values for all 2 580 151 examinations and 12 068 cancers are in Table 4.

empiric evidence on which to base performance targets (10,11).

Variation of mammographic outcome and accuracy measurements is well known, but the literature is generally limited to one practice or geographic area (19; 33; 34;36, p 5; 37), research methods not applicable to the community (38), or review of selected mammograms (39,40). Our study addresses these limitations by using graphic representations of outcome measurements from routine practice and from a broad base of representative radiologists and by using methods accessible to most radiologists. Thus, results of our study allow a radiologist to compare his or her outcome measurements with those of a group of radiologists who are representative of those in U.S. practice (11).

Findings in our report extend results of prior research because the range of performance outcomes in community practice for screening mammography is documented. Because only 10% of our mammograms were initial mammograms, results are best compared with results with subsequent mammograms. The average values of our results were similar to those in prior U.S. reports (8; 10;29;36, p 5). In general, the results for most radiologists are within the desirable ranges recommended for highly skilled radiologists (4, p 83; 6, pp 147–148) (Tables 6, 7), with the exception of the recall rate (median, 9.7%) and the PPV_1 (median, 4.5%).

The median recall rate for the United States is almost twice the value of the European guidelines (6, pp 147–148; 42) of less than 5% and well above the United Kingdom guidelines (5, pp 4–5) of less than 5% to 7%. This pattern of differing ranges of performance outcomes is also seen for PPV_1 but in the converse direction, with more than 50% of radiologists having values below the target value of 5%–10% suggested by organizations in the United States. The range of recall rate is primarily reflective of a pattern of care, and not random variation, as the radiologists included in this analysis all had performed more than 1000 interpretations each.

The majority of radiologists appear

Table 6

Performance Benchmarks for Cancer Detection

Cancer Data	Value
Mean cancer detection rate per 1000	4.6
No. of readers with ≥1000 examinations	344
No. of cancers	9529
Total no. of examinations	2 076 379
Reader performance (%)	
For 50th percentile (median)	4.4
For 10th–90th percentiles	2.4–7.0
For 25th–75th percentiles	3.2–5.8
Invasive cancer size	
No. of readers with ≥15 detected invasive cancers with known size	159
No. of cancers	5424
Reader performance for mean tumor size (%)	
For 50th percentile (median)	15.9
For 10th–90th percentiles	12.8–19.6
For 25th–75th percentiles	14.3–17.5
Reader performance for median tumor size (%)	
For 50th percentile (median)	13.0
For 10th–90th percentiles	10.0–15.0
For 25th–75th percentiles	12.0–14.5
Percentage of minimal cancer*	52.9
No. of readers with ≥15 detected cancers†	190
No. of minimal cancers	4066
Total no. of cancers†	7689
Reader performance (%)	
For 50th percentile (median)	51.8
For 10th–90th percentiles	37.8–65.8
For 25th–75th percentiles	45.8–59.1
Percentage of node-negative cancers	80.2
No. of readers with ≥15 invasive cancers‡	169
No. of node-negative cancers	4613
Total no. of cancers‡	5753
Reader performance (%)	
For 50th percentile (median)	80.9
For 10th–90th percentiles	69.6–91.7
For 25th–75th percentiles	75.0–84.9
Percentage of Stage 0 or I	76.0
Readers with ≥15 cancers‡	182
No. of stage 0 or I cancers	5521
Total no. of cancers‡	7268
Reader performance (%)	
For 50th percentile (median)	76.0
For 10th–90th percentiles	62.5–87.5
For 25th–75th percentiles	69.7–81.8

Note.—Data include examinations for radiologists with minimum numbers of mammograms as designated; examinations for which the radiologist was unknown were excluded.

* Included cases of ductal carcinoma in situ or invasive cancer of 10 mm or smaller.

† Ductal carcinoma in situ or invasive cancer with known size.

‡ Known nodal status.

§ Known stage.

to practice within the Agency for Health Care Policy and Research and European minimal guidelines for cancer detection rate and ductal carcinoma in situ detection, and the percentage of cancers that were classified as early stage appeared to substantially exceed the guidelines. These favorable outcomes are perhaps the result of the higher recall rates and/or shorter screening intervals (median of 18 months in the United States vs 36 months in the United Kingdom) (42), which are commonly observed in practice in the United States.

Study Limitations

Although we believe a major strength of our study is its large sample of clinical practices drawn from a diverse geographic area, features that allow it to be a mirror of the typical U.S. community practice, limitations in this analysis should be noted. We did not evaluate how characteristics, such as physician volume or the practice of double reading, influenced observed outcomes, nor did we evaluate how the outcome parameters interact (eg, an examination of the relationship of recall rate to sensitivity or cancer detection rate). These types of issues are crucial to the creation of new guidelines and are being evaluated in other research efforts in which BCSC data are used.

Use of BCSC Performance Benchmarks Data by Radiologists

For these data to be valuable to community radiologists, these radiologists need to collect the necessary data about their own practice. Efficient audit systems of screening mammographic practices are needed for collection of prospective, long-term, standardized, and high-quality data as recommended by the recent Institute of Medicine report (36). In many practices, maybe most, radiologists do not collect much of these data and cannot therefore evaluate their performance outcomes relative to the BCSC data about performance benchmarks for mammographic screening. Evaluation of the quality of care delivery is not an easy process and is best performed with specialized software, knowledgeable personnel, and access to pathologic data. There is no reimbursement for the substantial costs involved.

The collection of cancer data is likely to be incomplete in community radiology, and acquisition of complete follow-up information for outcomes for all biopsies is a greater challenge when no centralized registry exists.

Another important limitation is the small number of cancers present for each radiologist on screening mammograms performed during 1 year. A difference of even one or two cancers has a major effect on the resulting performance measurement. This imprecision may be offset by aggregation of several years of data and/or data from an entire group of radiologists (Appendix E1 *[radiology.rsnajnls.org/cgi/content/full/241/1/55/DC1]*).

Patient population differences also will alter these performance measurements. Patients' demographic characteristics such as age, race and ethnicity, family history of breast cancer, and prior mammography will affect all outcomes, especially cancer detection rate and recall rate (10–12,29,30,40).

The benchmark data displayed in our study represent the current range of community performance outcomes. It should not be inferred that the average community performance outcome is necessarily the recommended target for performance. Desirable goals, targets, or guidelines are created by a panel of experienced radiologists by using an evidence-based process that includes a thorough review of peer-reviewed literature and their own experiences. Examples of evidence-based review processes are those commissioned by the Agency for Healthcare Research and Quality–managed U.S. Preventive Services Task Force or a panel of experienced radiologists provided by a committee of the American College of Radiology. Benchmark data are useful to create these

Table 7

Comparison of Outcome Recommendations and Results with Current Performance Measurements

Outcome Measurement	Desirable AHCPR Guidelines, 1994*	European and United Kingdom Guidelines, 2001 and 2005†		British Columbia Study Results‡	BCSC Values, 1996–2002 (%)§
		Minimal	Desirable		
Recall rate	≤10 (all)	· · ·	· · ·	· · ·	9.7 (4.4–16.8)
Initial screening	· · ·	<7	≤5	9.8	12.3 (6.1–23.5)
Subsequent screening	· · ·	<5	≤3	4.4	8.8 (3.8–15.6)
Cancer detection rate	2–10‖	· · ·	· · ·	· · ·	4.4 (2.4–7.0)
Initial screening	6–10‖	3‖	>3‖	5.0‖	4.4 (2.2–7.9)
Subsequent screening	2–10‖	1.5‖	>1.5‖	2.8‖	4.3 (2.2–6.9)
PPV_1	5–10	NA	NA	· · ·	4.5 (2.6–8.6)
PPV_2	25–40	· · ·	· · ·	· · ·	25.0 (14.1–38.8)
Node-positive cancer	<25	· · ·	· · ·	19	18.8 (12.5–37.5)
Ductal carcinoma in situ cases	· · ·	10	10–20	20	21.6#
Minimal cancer**	>30	· · ·	NA	· · ·	51.8 (37.8–65.8)
Stage 0 or I	>50	· · ·	· · ·	· · ·	76.0 (62.5–87.5)

Note.—Values are percentages except where otherwise indicated. NA = not applicable.

* Reference 4, p 83. AHCPR = Agency for Health Care Policy and Research.

† References 5, pp 4–5; 6, pp 147–148.

‡ Reference 41.

§ Data are from Tables 5 and 6 of the current study; initial versus subsequent recall rate data are not shown elsewhere. Data are for medians, and numbers in parentheses are values for the 10th–90th percentiles.

‖ Per 1000.

The 10th–90th percentiles were not calculated.

** Defined as cases of ductal carcinoma in situ or invasive cancer of 10 mm or smaller.

goals or guidelines, but they are useful to the practicing radiologist only if he or she has data from a similar population that are calculated by using the same methods. Our study findings, therefore, should be helpful in updating any future guidelines, as they demonstrate the variation present in community practice.

The wide variation in recall rate likely represents a lack of consensus among practicing radiologists concerning performance targets. A wide variation is of concern because the literature suggests that wide variation in the processes of care delivery may be associated with lower quality or worse outcomes from care delivery (43, p 88). These benchmark data demonstrate the need for improvement in screening mammographic performance in the United States if the guidelines recommended by the Agency for Healthcare Research and Quality are to be met.

A major caution in the use of screening mammographic guidelines is that a single performance measurement in isolation may not be meaningful. In particular, the acceptability of a value for recall rate also depends on the parallel acceptability of values for PPV_1, cancer detection rate, and size of invasive cancer detected. For example, a radiologist with a higher than average recall rate may need those additional recalled patients to achieve a sufficiently high cancer detection rate and a sufficiently high detection rate for invasive cancers of small size. Investigators in another BCSC study are examining how the parameters of recall rate, cancer detection rate, and sensitivity interact in clinical practice.

Fortunately, efforts have already begun to evaluate screening mammographic performance through an integrated assessment of recall rate, cancer detection rate, and PPV. One such innovative approach has been developed in the United Kingdom where facilities receive feedback in a single graphic display that indicates whether they are operating within an acceptable range of three measurements: recall rate, cancer detection rate, and PPV_1 (44). This feedback allows identification of facilities or radiologists who are operating outside of the acceptable range for these parameters. With on-site review, experienced radiologists can then confirm problems and assist in processes necessary to improve performance outcome (5, pp 4–5; 45).

Conclusion

Our study findings indicate the range of performance benchmarks for screening mammography performed by community radiologists in the United States and should be useful as comparative data for individual radiologists and for establishment of outcome guidelines.

Although most measurements of the performance of radiologists for screening examinations are similar to published recommendations, for many radiologists, recall rate is higher and PPV_1 is lower than the recommendations. Variability in all of these measurements for any practice may relate in part to methods used but primarily reflects actual differences. Additional research and involvement by panels of experienced radiologists will be required to better define the optimal performance targets appropriate to the U.S. health care environment.

References

1. Monsees BS. The Mammography Quality Standards Act: an overview of the regulations and guidance. Radiol Clin North Am 2000;38:759–772.
2. Spring DB, Kimbrell-Wilmot K. Evaluating the success of mammography at the local level: how to conduct an audit of your practice. Radiol Clin North Am 1987;25:983–992.
3. Murphy WA Jr, Destouet JM, Monsees BS. Professional quality assurance for mammography screening programs. Radiology 1990; 175:319–320.
4. Bassett LW, Hendrick RE, Bassford TL, et al. Quality determinants of mammography. In: Clinical practice guideline no. 13: AHCPR publication no. 95–0632. Rockville, Md: Agency for Health Care Policy and Research, Public Health Service, U.S. Department of Health and Human Services, October 1994.
5. Liston J, Wilson R, eds. Quality assurance guidelines for breast cancer screening radiology. NHS Breast Screening Programmes publication no. 59. Sheffield, England: NHS Cancer Screening Programmes, January 2005.
6. Perry N, Broeders M, deWolf C, Törnberg S, eds. European guidelines for quality assurance in mammography screening. 3rd ed. Luxembourg: European Commission, 2001.
7. 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.
8. Rosenberg RD, Yankaskas BC, Hunt WC, et al. Effect of variations in operational definitions on performance estimates for screening mammography. Acad Radiol 2000;7: 1058–1068.
9. 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.
10. Yankaskas BC, Taplin SH, Ichikawa L, et al. Association between mammography timing and measures of screening performances in the United States. Radiology 2005;234:363–373.
11. Sickles EA, Miglioretti DL, Ballard-Barbash R, et al. Performance benchmarks for diagnostic mammography. Radiology 2005;235: 775–790.
12. Kerlikowske K, Carney PA, Geller B, et al. Performance of screening mammography among women with and without a first-degree relative with breast cancer. Ann Intern Med 2000;133:855–863.
13. Kerlikowske K, Smith-Bindman R, Abraham LA, et al. Breast cancer yield for screening mammographic examinations with recommendation for short-interval follow-up. Radiology 2005;234:684–692.
14. Gur D, Sumkin JH, Hardesty LA, et al. Recall and detection rates in screening mammography. Cancer 2004;100:1590–1594.
15. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists: findings from a national sample. Arch Intern Med 1996;156: 209–213.
16. Elmore JG, Wells CK, Howard DH. Does diagnostic accuracy in mammography depend on radiologists' experience? J Womens Health 1998;7:443–449.
17. Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. Radiology 2002;224:861–869.
18. Wagner RF, Beam CA, Beiden SV. Reader variability in mammography and its implications for expected utility over the population

of readers and cases. Med Decis Making 2004;24:561–572.

19. Smith-Bindman R, Chu P, Miglioretti DL, et al. Physician predictors of mammographic accuracy. J Natl Cancer Inst 2005;97:358–367.
20. D'Orsi CJ, Bassett LW, Berg WA, et al. Follow-up and outcome monitoring. In: Breast imaging reporting and data system: ACR BI-RADS. 4th ed. Reston, Va: American College of Radiology, 2003;229–251.
21. Sickles EA. Auditing your practice. In: Kopans DB, Mendelson EB, eds. Syllabus: a categorical course in breast imaging. Oak Brook, Ill: Radiological Society of North America, 1995; 81–91.
22. Aiello EJ, Buist DS, White E, Seger D, Taplin SH. Rate of breast cancer diagnoses among post-menopausal women with self-reported breast symptoms. J Am Board Fam Pract 2004;17:408–415.
23. Heinzen MT, Yankaskas BC, Kwok RK. Comparison of woman-specific versus breast specific data for reporting screening mammography performance. Acad Radiol 2000; 7:232–236.
24. 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.
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; 222:536–542.
26. Sickles EA, Ominsky SH, Sollitto RA, Galvin HB, Monticciolo DL. Medical audit of a rapid-throughput mammography screening practice: methodology and results of 27,114 examinations. Radiology 1990;175:323–327.
27. Tabár L, Fagerberg G, Duffy SW, Day NE, Gad A, Gröntoft O. Update of the Swedish two-county program of mammographic screening for breast cancer. Radiol Clin North Am 1992;30:187–210.
28. Sickles EA. Quality assurance: how to audit your own mammography practice. Radiol Clin North Am 1992;30:265–275.
29. Kerlikowske K, Grady D, Barclay J, Sickles EA, Eaton A, Ernster V. Positive predictive value of screening mammography by age and family history of breast cancer. JAMA 1993; 270:2444–2450.
30. Frankel SD, Sickles EA, Curpen BN, Sollitto RA, Ominsky SH, Galvin HB. Initial versus subsequent screening mammography: comparison of findings and their prognostic significance. AJR Am J Roentgenol 1995;164: 1107–1109.
31. Kerlikowske K, Grady D, Barclay J, Sickles EA, Ernster V. Effect of age, breast density, and family history on the sensitivity of first screening mammography. JAMA 1996;276: 33–38.
32. American Joint Committee on Cancer. Manual for staging of cancer. 5th ed. Philadelphia, Pa: Lippincott, 1997.
33. Linver MN, Paster SB, Rosenberg RD, Key CR, Stidley CA, King WV. Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2 year medical audit of 38,633 cases. Radiology 1992;184:39–43. [Published correction appears in Radiology 1992;184(3):878.]
34. Rosenberg RD, Lando JF, Hunt WC, et al. The New Mexico Mammography Project: screening mammography performance in Albuquerque, New Mexico, 1991 to 1993. Cancer 1996;78:1731–1739.
35. Poplack SP, Tosteson AN, Grove MR, Wells WA, Carney PA. Mammography in 53,803 women from the New Hampshire Mammography Network. Radiology 2000;217:832–840.
36. Nass S, Ball J, eds. Improving breast imaging quality standards. Committee on Improving Mammography Quality Standards, National Research Council. Washington, DC: National Academies Press, 2005.
37. 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.
38. 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.
39. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med 1994;331:1493–1499.
40. Kerlikowske K, Grady DG, Barclay J, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS). J Natl Cancer Inst 1998;90:1801–1809.
41. Olivotto IA, Kan L, d'Yachkova Y, et al. Ten years of breast screening in the Screening Mammography Program of British Columbia, 1988-97. J Med Screen 2000;7:152–159.
42. 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.
43. Hewitt M, Simone JV, eds. Ensuring the quality of cancer care. National Cancer Policy Board, Institute of Medicine and National Research Council. Washington, DC: National Academies Press, 1999.
44. Blanks RG, Moss SM, Wallis MG. Monitoring and evaluating the UK National Health Service Breast Screening Programme: evaluating the variation in radiological performance between individual programmes using PPV-referral diagrams. J Med Screen 2001;8:24–28.
45. Perry NM. Interpretive skills in the National Health Service Breast Screening Programme: performance indicators and remedial measures. Semin Breast Dis 2003;6:108–113.