Improving the Accuracy of Mammography: Volume and Outcome Relationships

Laura Esserman, Helen Cowley, Carey Eberle, Alastair Kirkpatrick, Sophia Chang, Kevin Berbaum, Alastair Gale

Background: **Countries with centralized, high-volume mammography screening programs, such as the U.K. and Sweden, emphasize high specificity (low percentage of false positives) and high sensitivity (high percentage of true positives). By contrast, the United States does not have centralized, high-volume screening programs, emphasizes high sensitivity, and has lower average specificity. We investigated whether high sensitivity can be achieved in the context of high specificity and whether the number of mammograms read per radiologist (reader volume) drives both sensitivity and specificity.** ***Methods:*** **The U.K.'s National Health Service Breast Screening Programme uses the PERFORMS 2 test as a teaching and assessment tool for radiologists. The same 60-film PERFORMS 2 test was given to 194 high-volume U.K. radiologists and to 60 U.S. radiologists, who were assigned to low-, medium-, or high-volume groups on the basis of the number of mammograms read per month. The standard binormal receiver-operating characteristic (ROC) model was fitted to the data of individual readers. Detection accuracy was measured by the sensitivity at specificity = 0.90, and differences among sensitivities were determined by analysis of variance.** ***Results:*** **The average sensitivity at specificity = 0.90 was 0.785 for U.K. radiologists, 0.756 for high-volume U.S. radiologists, 0.702 for medium-volume U.S. radiologists, and 0.648 for low-volume U.S. radiologists. At this specificity, low-volume U.S. radiologists had statistically significantly lower sensitivity than either high-volume U.S. radiologists or U.K. radiologists, and medium-volume U.S. radiologists had statistically significantly lower sensitivity than U.K. radiologists (*$P<.001$*, for all comparisons).** ***Conclusions:*** **Reader volume is an important determinant of mammogram sensitivity and specificity. High sensitivity (high cancer detection rate) can be achieved with high specificity (low false-positive rate) in high-volume centers. This study suggests that there is great potential for optimizing mammography screening. [J Natl Cancer Inst 2002;94: 369–75]**

The organization of mammography screening programs varies dramatically throughout the United States and Europe. Countries with socialized medicine, such as the U.K. and Sweden, have adopted a centrally organized approach to screening that emphasizes high specificity as well as high sensitivity, which results in an effective program that is lower in cost *(1).* By contrast, the United States has a decentralized system that is not principally organized around high-volume centers *(1).* Indeed, the minimum annual reading volume for radiologists in breast cancer screening programs varies greatly in the United States compared with the U.K. or Sweden. The minimum annual reading volume in the United States is 480 as set by the Mammography Quality Standards Act of 1992 *(2),* while the minimum set by the National Health Service Breast Screening Programme in the U.K. *(3)* is 5000 mammograms per year.

An additional difference between the United States and the U.K. or Sweden is seen in the threshold for recommending a biopsy and, thus, in the percentage of biopsy specimens that result in a cancer diagnosis (cancer-to-biopsy yield) *(4–8).* In the United States, it is commonly thought that a high cancer-to-biopsy yield would demonstrate a willingness to tolerate a high false-negative rate and a lower effectiveness of screening *(9,10).* In the U.K. and Sweden, the cancer-to-biopsy yield is considerably higher than in the United States, and the rates of additional evaluative tests for mammographic abnormalities are lower *(11,12).* For example, cancer-to-biopsy yields in the U.K. and Sweden, with the use of stereotactic techniques, are in the range of 40%–60% (fine-needle aspiration or core biopsy) *(11,13).* The cancer-to-surgical biopsy yield, however, is much higher (i.e., 88%–90%) *(13)* because of the extensive reliance on stereotactic biopsy specimens. In the largest reported series of stereotactic biopsy specimens in the United States *(9,14,15),* the cancer-to-biopsy yield was just 20%–25%. Other published stereotactic series in the United States report cancer-to-biopsy ratios of approximately 11% and intimate that aggressive biopsy rates result in improved sensitivity *(8).* More experienced mammography units in the United States *(16,17),* however, report cancer-to-stereotactic biopsy yields of 37%–40%. The bias for more intervention in the United States relative to the U.K. and Sweden has been suggested to reflect a cultural bias rather than a quality advantage *(9,15,18).*

One of the reasons cited for the acceptance of low specificity in the United States is the intent to maximize sensitivity *(9,10).* However, it has never been shown that high sensitivity necessarily requires lower specificity. In European countries, with socialized medicine and a fixed health-care budget, loss of specificity is of equal concern to loss of sensitivity. This difference between the United States and European countries provides an opportunity to explore whether high sensitivity necessarily requires low specificity or whether high sensitivity can be achieved in the context of high specificity.

We hypothesized that volume drives both sensitivity and specificity and that higher specificity is not necessarily associ-

Affiliations of authors: L. Esserman, C. Eberle (Department of Surgery), S. Chang (Institute for Health Policy Studies), University of California at San Francisco; H. Cowley, A. Gale, Institute of Behavioural Sciences, University of Derby, U.K.; A. Kirkpatrick, Scottish Breast Screening Programme, South-East Scotland Division, Edinburgh; K. Berbaum, Department of Radiology, University of Iowa, Iowa City.

Correspondence to: Laura Esserman, M.D., M.B.A., Carol Franc Buck Breast Care Center, UCSF/Mt. Zion Cancer Center, 1600 Divisadero St., 2nd Floor, San Francisco, CA 94143–1710 (e-mail: laura.esserman@ucsfmedctr.org).

See "Notes" following "References."

ated with an acceptance of lower cancer detection rates. If this hypothesis is true, changes in the organization of mammography screening could lead to substantial improvements for patients and payers alike. To test this hypothesis, we evaluated three groups of radiologists from the United States (high-, medium-, and low-volume) as well as radiologists from the U.K. and Sweden by using the PERFORMS 2 test. We used receiver-operating characteristic (ROC) curve methodology to determine whether gains in specificity come at the expense of sensitivity (i.e., movement along the same curve) or whether specificity can be improved without degrading sensitivity by performing at a different level (i.e., movement to a different curve entirely) (Fig. 1).

METHODS

Participating Radiologists

Our study sample of U.S. radiologists was chosen from a list of 1322 radiology facilities licensed to practice throughout the state of California. The list was obtained from the State Radiation Control Office, Sacramento, CA, in 1996. Because no local or state agencies track volume information for individual radiologists, we contacted all of the listed facilities by letter and telephone to confirm that the facility was still in operation, to obtain the name of the radiologist who read the highest number of mammograms each month, and to obtain the average number of mammograms read by that radiologist each month. That radiologist became the contact radiologist for that site.

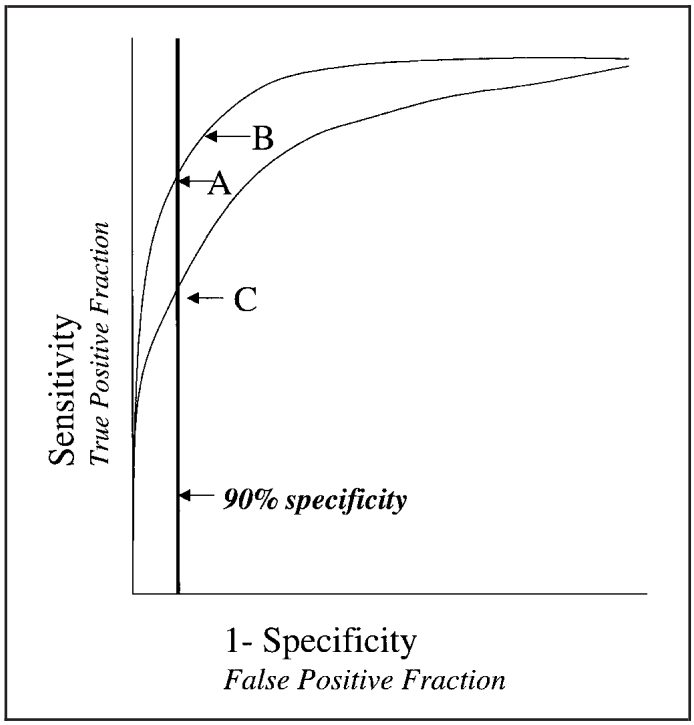

Fig. 1. A receiver-operating characteristic (ROC) curve schematic demonstrates that, when specificity of a diagnostic test increases (false-positive fraction decreases), the ROC curve dictates an attendant decrease in sensitivity. As sensitivity increases, specificity will necessarily decrease if all operators are on the same ROC curve (movement from **point A** to **point B**). Another scenario, however, is that some operators (in this case, mammographers) will not be on the optimal curve, and both sensitivity and specificity will be lower (movement from **point A** to **point C**). For a screening test, such as mammography, which requires high specificity, it is most appropriate to assess sensitivity at a specificity of 90% (**vertical line**).

Nonrespondent facilities were followed-up with additional telephone calls. Of the 1322 facilities listed, 261 (20%) were repeat listings, were no longer in practice, or had merged with another facility. This relatively high percentage of facility change or turnover reflects the dynamic nature of medical practices in California in the middle to late 1990s. Of the remaining 1061 facilities, 219 were unreachable or did not perform mammography. We successfully contacted 842 facilities (79%). All 842 facilities were certified, as of 1996, according to the Mammography Quality Standards Act of 1992.

On the basis of information from each facility regarding the volume of mammograms read by the contact radiologist (Table 1), we grouped the radiologists as low-volume radiologists (≤100 mammograms read per month), medium-volume radiologists (101–300 mammograms read per month), and high-volume radiologists (≥301 mammograms read per month), according to the volume categories described by Houn and Brown *(19)* in the National Cancer Institute's phase I of the National Survey of Mammography Facilities. The list of names in each volume group was scrambled and reordered with the use of a computer-generated randomization scheme. Letters soliciting participation in the study were sent to radiologists from each group. Radiologists at the top of each list received the first letters, and we continued to send letters until we had 60 participating radiologists. We mailed 181 letters to enroll the 60 radiologists who participated in the study—a response rate of approximately 33%.

A total of 60 radiologists read the PERFORMS 2 test set. However, after completion of the study, data for one radiologist were omitted from the analysis because of an unrecoverable error during transmission (via e-mail to the U.K.). At the time of the administration of the PERFORMS 2 test, the 60 radiologists were asked to confirm their reader volume for final classification among volume categories. Two radiologists were reassigned from the high-volume group to the medium-volume group. Therefore, our final volume distribution was as follows: 19 low-volume radiologists (32%), 22 medium-volume radiologists (37%), and 18 high-volume radiologists (31%).

The radiologists were given $100 (U.S. dollars) for participating in the study. The study was approved by the Committee on Human Research of the University of California at San Francisco, and all of the participating radiologists signed consent forms.

Table 1. U.S. mammography facilities analyzed by volume of films read

	Total No. (%)
Mammography facilities*	1322
Questionnaires mailed/facilities contacted†	1061
Identified as mammography facilities‡	842
Low-volume contact radiologists§	212 (25)
Medium-volume contact radiologists‖	360 (43)
High-volume contact radiologists¶	270 (32)

*Number of facilities reported by the State Radiation Control Office in 1996.

†261 facilities were repeat listings, were no longer in practice, or had merged with another facility.

‡219 of the 1061 facilities contacted were unreachable or did not perform mammography. Approximately 50% responded by mail and 50% by telephone.

§Numbers of mammograms per radiologist reflect the number for the most active mammography-interpreting physician at each facility, as reported by the facility. Low-volume radiologists read 100 or fewer mammograms per month.

‖Medium-volume radiologists read 101–300 mammograms per month.

¶High-volume radiologists read 301 or more mammograms per month.

All of the U.K. mammographers who are involved in the National Screening Programme are dedicated high-volume readers. All were invited to participate in the PERFORMS 2 assessment, and more than 90% of all U.K. mammographers were evaluated—a total of 194 U.K. mammographers.

We chose two Swedish radiologists from a single high-volume program to read the PERFORMS 2 test films. The data from these two subjects were used for comparison purposes and were not included in the data analyses. In Sweden, there are five high-volume, independent screening mammography units. The two primary radiologists from the Stockholm site agreed to participate during a conference on mammography organization held in San Francisco, CA.

PERFORMS 2 Test

The PERFORMS 2 test is a teaching tool developed in the U.K. as part of the quality-assurance component of the National Health Service Breast Screening Programme *(20–22)* and is recommended by the Royal College of Radiologists and the National Health Service Breast Screening Programme.

The PERFORMS 2 test was administered to the U.S. radiologists by use of methods similar to those used in the U.K. and was described previously *(23,24).* We used a single set of 60 two-view films that contained 13 cancers. The radiologists were not told how many overall cancers were present in the test set. The 60 films were the same 60 films viewed by the U.K. radiologists. A computer system (Psion, Inc., Concord, MA) with bar codes was used to record the radiologists' answers. Radiologists had the choice of using the Psion computer or stating their answers, which were recorded on simple data sheets by the research coordinator administering the test. For each film, the radiologist was asked to record whether there were any findings, to record the location and nature of any findings, and to record whether the patient should be recalled for further tests. The radiologists classified each finding on a 5-point system: normal/benign, probably benign, indeterminate, probably malignant, and malignant. This classification system is similar to the Breast Imaging Reporting and Data System (BI-RADS) *(25).* Patient recall is considered to be appropriate except for cases that are classified as normal/benign.

Validation of the PERFORMS 2 Test

Long-term follow-up information on patients was available for all of the PERFORMS 2 test films. All films are screening films from asymptomatic women. There were no occult cancers in the set, and all of the patients had sufficient follow-up to determine the outcome by biopsy or by 3-year follow-up. For classification of lesions, an expert panel of five experienced, high-volume, nationally known U.K. mammographers evaluated the films, and a consensus was used to determine the classification of appropriate versus inappropriate recall for each film. With the use of the outcome data, the film set was validated to ensure that the cases were appropriate and consistent with the actual patient outcome (presence of cancer at 3-year follow-up). In this screening set, recall refers to the decision to recall a patient for more diagnostic studies or for a biopsy, and this is not equivalent to recommendation for biopsy.

Data Analysis

The choice of design and statistical technique for an experiment depends on whether statistical generalization to observers or to patients is more fundamental, which in turn depends on the nature of the experimental question *(26,27).* Volume of images interpreted by an observer is an attribute of the observer, not of the patient. Our application of ROC methods is designed to allow experimental results to be generalized to the population of radiologists from whom the sample was selected.

The rating data of individual observers were fitted with the standard binormal model *(28,29)* with the use of a computer program called RSCORE4.66 (ftp://perception.radiology.uiowa.edu/rscore/r466files.zip), written by D. D. Dorfman, K. S. Berbaum, H. Abu-Dagga, and K. M. Schartz, Department of Radiology, University of Iowa, Iowa City. Because the binormal ROC curves for some observers crossed the chance line, sensitivity at specificity = 0.90 was selected to measure observer accuracy. This measure is relatively unaffected by ROC extrapolation error to regions of lower specificity. In addition, the rating data of individual observers were fitted with the contaminated binormal model *(30),* which yields only proper ROC curves that do cross the chance line. The purpose of performing this type of ROC analysis was to make certain that the conclusions did not depend on the assumption of the ROC model. Analysis of area under proper ROC curves ought to lead to the same conclusions as analysis of sensitivity at specificity = 0.90 estimated by use of the standard binormal ROC curves.

The sensitivity at specificity = 0.90 and proper areas of the four groups of observers—U.S. radiologists reading at low, medium, and high volumes and U.K. radiologists—were analyzed with the use of analysis of variance *(31)* and Scheffé's pairwise multiple comparison tests *(32)* ($P<.05$) with the use of BMDP 7D (Statistical Solutions Ltd., Cork, Ireland, 2001). In addition, to determine if there were statistically significant differences in sensitivity, specificity, and the percentage of cancers detected, we compared these data for the four groups by using analysis of variance *(31)* and Scheffé's pairwise multiple comparison tests *(32)* ($P<.05$) with the use of BMDP 7D (Statistical Solutions Ltd., 2001).

Sensitivity was defined as the percentage of cases correctly recalled for further assessment out of all of the cases known to require recall. Specificity was defined as the percentage of cases correctly classified as normal out of all of the cases known to be normal. The sensitivity at specificity = 0.90 and proper ROC areas are both measures of quality. We defined quality as the overall sensitivity and specificity.

Results

Comparisons Among Groups of Radiologists

We tested the hypothesis that volume drives both sensitivity and specificity and that higher specificity is not necessarily associated with an acceptance of lower cancer-detection rates by using the PERFORMS 2 test. For illustrative purposes, the sensitivity at specificity = 0.90 for the standard binormal model for the three groups of U.S. radiologists and for the U.K. and Swedish radiologists is shown in Fig. 2.

At specificity = 0.90, the average sensitivity was 0.648 for low-volume U.S. radiologists, 0.702 for medium-volume U.S. radiologists, 0.756 for high-volume U.S. radiologists, and 0.785 for U.K. radiologists (Fig. 2). The null hypothesis of no difference among the groups was rejected [$F(3, 249) = 12.95$; $P<.001$]. Scheffé's tests indicated that the sensitivity of low-volume U.S. radiologists was statistically significantly lower

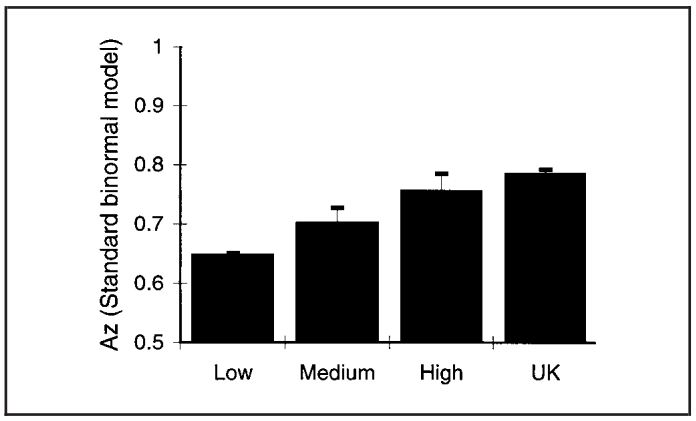

Fig. 2. Sensitivity at specificity = 0.90 with the use of the standard binormal model for different-volume mammography readers of the PERFORMS 2 test. Low-, medium-, and high-volume readers were from the United States. Low-volume radiologists read 100 or fewer mammograms per month. Medium-volume radiologists read 101–300 mammograms per month. High-volume radiologists read 301 or more mammograms per month. All U.K. radiologists were high-volume readers. Two Swedish radiologists, included as a high-volume control (data not shown), were high-volume readers who demonstrated a sensitivity of 88%. **Bars** represent the mean and standard error of the mean for the sensitivity at specificity = 0.90.

than that of high-volume U.S. radiologists and U.K. radiologists and that the sensitivity of medium-volume U.S. radiologists was statistically significantly lower than that of U.K. radiologists. For purposes of a control, we asked two highly experienced, high-volume Swedish mammographers to take the PERFORMS test, and their sensitivity at specificity = 0.90 was 88% (data not shown or included in the analysis of variance).

The conclusions based on the analysis of the area under the proper ROC curves were the same as those from the analysis of sensitivity. The average area under the proper contaminated binormal ROC curve was 0.832 for low-volume U.S. radiologists, 0.856 for medium-volume U.S. radiologists, 0.891 for high-volume U.S. radiologists, and 0.902 for U.K. radiologists. Levene's test *(33)* for equality of group variability was statistically significant [$F(3, 249) = 4.42$; $P = .0048$]. Therefore, Brown–Forsythe F tests *(34,35)* and separate variance pairwise t tests that do not assume homogeneity of variance were performed on proper ROC area. The Bonferroni inequality was used to correct the α level for multiple t tests: $\alpha = 0.05$ divided by the number of tests *(36)*. The null hypothesis of no difference among the groups was rejected [$F(3, 60) = 9.96$; $P<.001$]. Separate variance pairwise t tests indicated that, on average, the proper ROC area of low-volume U.S. radiologists was statistically significantly lower than that of high-volume U.S. radiologists and of U.K. radiologists and that the proper ROC area of medium-volume U.S. radiologists was statistically significantly less than that of U.K. radiologists.

The results from the two types of ROC analysis, therefore, suggest that volume affects diagnostic accuracy.

Differences in Sensitivity, Specificity, and Missed Malignancies

The average sensitivity was 70.3% for low-volume U.S. radiologists, 69.7% for medium-volume U.S. radiologists, 77.0% for high-volume U.S. radiologists, and 79.3% for U.K. radiologists. Levene's test *(33)* for equality of group variability was statistically significant [$F(3, 249) = 4.69$; $P = .0033$]. Therefore, Brown–Forsythe F tests *(34,35)* and separate variance pairwise t tests that do not assume homogeneity of variance were performed on sensitivity data. The α level reported was corrected for the six tests performed with the use of the Bonferroni inequality *(36)*. The null hypothesis of no difference among the groups was rejected [$F(3, 55) = 6.99$; $P<.001$]. Separate variance pairwise t tests indicated that, on average, the sensitivity of low-volume and medium-volume U.S. radiologists was statistically significantly lower than that of U.K. radiologists ($P<.05$).

The average specificity was 83.6% for low-volume U.S. radiologists, 88.2% for medium-volume U.S. radiologists, 88.0% for high-volume U.S. radiologists, and 88.0% for U.K. radiologists. Levene's test *(33)* for equality of group variability was not statistically significant, and the null hypothesis of no difference among the groups was not rejected [$F(3, 249) = 1.92$; $P = .126$].

The average percentage of cancers detected was 71.5% for low-volume U.S. radiologists, 69.0% for medium-volume U.S. radiologists, 78.6% for high-volume U.S. radiologists, and 83.5% for U.K. radiologists (Fig. 3). Levene's test *(33)* for equality of group variability was statistically significant [$F(3, 249) = 7.35$; $P<.001$]. Therefore, Brown–Forsythe F tests *(34,35)* and separate variance pairwise t tests that do not assume homogeneity of variance were performed on the data for the percentage of cancers detected. The α level reported was corrected for the six tests performed with the use of Bonferroni inequality *(36)*. The null hypothesis of no difference among the groups was rejected [$F(3, 57) = 9.99$; $P<.001$]. Separate variance pairwise t tests indicated that, on average, the percentage of cancers detected by low-volume and by medium-volume U.S. radiologists was statistically significantly lower than that detected by U.K. radiologists ($P<.01$).

Discussion

In this study, we sought to determine whether higher volume of mammograms interpreted by radiologists is associated with

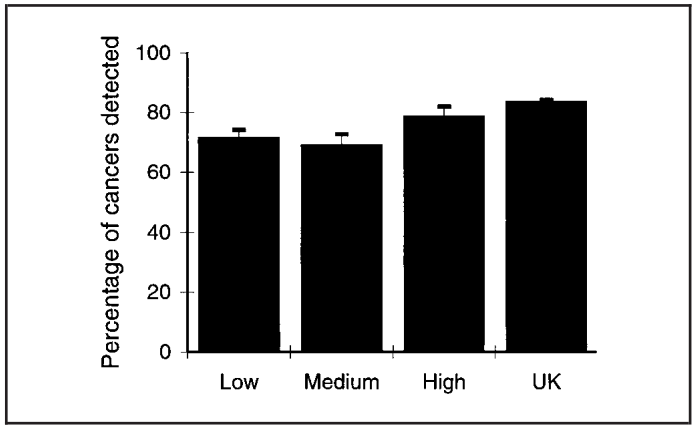

Fig. 3. Percentage of malignant cases detected among the groups of radiologists. Low-, medium-, and high-volume readers were from the United States. Low-volume radiologists read 100 or fewer mammograms per month. Medium-volume radiologists read 101–300 mammograms per month. High-volume radiologists read 301 or more mammograms per month. All U.K. radiologists were high-volume readers. Two Swedish radiologists, included as a high-volume control (data not shown), were high-volume readers who identified 100% of malignant cases. **Bars** represent the mean and standard error of the mean for the percentage of cancers detected by radiologists in each group.

an increase in diagnostic accuracy or is a simple tradeoff of sensitivity for specificity. We hoped that testing this hypothesis would shed light on the striking difference in the threshold for biopsy (both stereotactic and surgical) in the United States and in the U.K. Our findings demonstrate clearly that the volume of mammograms interpreted is a determinant of diagnostic accuracy. This conclusion is in line with a recent study by Kan et al. *(37),* who used standardized abnormal interpretation ratios and standardized cancer detection ratios to conclude that a minimum of 2500 interpretations per year is associated with lower abnormal interpretation rates and average or better cancer detection rates. In our series, the length of time that a radiologist has been reading mammograms did not affect the quality of the reading, probably because 82% of the U.S. readers had read mammograms for more than 10 years and 88% had read them for at least 5 years. We did not have data on the duration of reading of 10% of the radiologists. Experience, which is a combination of both volume of studies read and years spent as a reader, was not specifically tested, although they are clearly related.

Our result that reader volume affects cancer-detection accuracy is not very surprising: The volume of procedures or patients has been demonstrated repeatedly to be a strong determinant of quality in medical procedures *(38,39).* Volume–outcome studies clearly show that mortality from surgical procedures, including cardiac *(40–42),* gastrointestinal *(43–45),* and transplantation *(46)* procedures, decreases dramatically when critical threshold volumes are reached. Therefore, it is not unexpected that the outcome of other high-volume procedures, such as breast imaging, is also improved as volume increases. Indeed, the data from the Swedish population-based screening studies, in which mammography is performed by experts in high-volume centers *(1),* provide the foundation from which evidence-based recommendations for mammography screening are derived *(47–50).*

Many factors could affect the cross-cultural differences in cancer-to-biopsy yield and thresholds for intervention. One explanation for lower cancer-to-biopsy yields in the United States is that the possibility of litigation from missed cancers *(15,18)* is much greater in the United States than in European countries *(51,52)* and may cause U.S. radiologists to lower the threshold for biopsy.

Although high sensitivity is viewed in the United States as a primary goal of mammography, there is a substantial cost for a lower threshold to biopsy. Elmore et al. *(53)* found that, during a 10-year period, one third of women screened had an abnormal mammogram that required an additional evaluation, even though no breast cancer was present. Furthermore, many women undergo biopsies for benign findings, which causes them great emotional distress *(54,55).* The controversy regarding screening of women aged 40–49 years is driven in part by their higher rates of false-positive screens *(56).* The cost of potentially unnecessary biopsies for the United States as a whole is more than $1 billion annually *(4).* The cost of disproving false-positive tests, in fact, drives a substantial part of the total cost of screening *(4).* Thus, reducing mammography interventions for benign disease without increasing missed cancers would be of enormous benefit *(4).*

The controversy over mammography is often focused on whether or not it should be used as a screening tool *(57).* But another equally important issue, given its widespread use, is the optimization of mammography. Substantial evidence supports the use of mammography as a screening tool, and it is currently considered to be part of the standard of care, at least for women over 50 *(58).* Considerable effort should, therefore, be devoted to determining how to make mammography as effective as it can be and to reduce the tremendous variation in interpretation and biopsy rates *(6,7).* These efforts are likely to make mammography screening more cost-effective, enabling better use of resources.

Our finding that higher volume improves diagnostic performance suggests that there may be an opportunity to improve quality and efficiency by re-engineering the organization of U.S. mammography screening programs. Higher quality (improved diagnostic performance) does not need to come at the price of more interventions. Mammography is one of many examples in medicine where we need to focus on rewarding quality outcomes rather than just on payment for procedures performed. Tracking and reporting critical outcome measures, such as sensitivity, specificity, size and stage of tumors detected, interval cancer rates, and time to recall and diagnosis, have been used in many countries to improve screening performance *(3,13).* Digital mammography is one technologic advance that may support the establishment of high-volume centers of excellence, permitting mammogram interpretations to be made by high-volume experienced and dedicated radiologists at distant locations. This would also enhance comparison of films from year to year because films stored electronically are more readily accessible. The net effect of this kind of reorganization should be the reduction in recall rates and false-positive biopsy specimens as well as a decrease in both economic and social costs.

Comparison of international practice styles in mammography has provided insight into organizational strategies that might very well improve the frequency, sensitivity, and specificity of screening, as well as decrease the overall cost of mammography in the United States. What is necessary are creative and new approaches to the implementation of mammography screening that include a better understanding of the factors that affect cost and quality, the tracking of quality, and strategies for developing centers of excellence for breast screening in a changing healthcare environment.

REFERENCES

(1) Shapiro S, Coleman EA, Broeders M, Codd M, de Koning H, Fracheboud J, et al. Breast cancer screening programmes in 22 countries: current policies, administration and guidelines. International Breast Cancer Screening Network (ISBN) and the European Network of Pilot Projects for Breast Cancer Screening. Int J Epidemiol 1998;27:735–42.

(2) Food and Drug Administration, U.S. Department of Health and Human Services. Federal Register: October 28, 1997 (Vol. 62, No. 208). Quality mammography standards; final rule. 21 CFR, Parts 16 and 900. p. 55852.

(3) National Health Service (NHS) Breast Screening Radiologists Quality Assurance Committee. Quality assurance guidelines for radiologists. National Health Service Breast Screening Programme (NHSBSP) Publ No. 15. Sheffield (U.K.): NHSBSP Publications; revised May 1997.

(4) Burnside E, Belkora JK, Esserman LJ. The impact of alternative practices on the cost and quality of mammographic screening in the United States. Clin Breast Cancer 2001;2:145–52.

(5) Ellis IO, Gale MH, Locker A, Roebuck EJ, Elston CW, Blamey RW, et al. Early experience in breast cancer screening: emphasis on development of protocols for triple assessment. The Breast 1993;2:148–53.

(6) 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;156:209–13.

(7) Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med 1994;331: 1493–9.

(8) Margolin FR, Leung JW, Jacobs RP, Denny SR. Percutaneous imaging-guided core breast biopsy: 5 years' experience in a community hospital. AJR Am J Roentgenol 2001;177:559–64.
(9) Stacey-Clear A, McCarthy KA, Hall DA, Pile-Spellman E, White G, Hulka C, et al. Breast cancer survival among women under age 50: is mammography detrimental? Lancet 1992;340:991–4.
(10) Moskowitz M. Guidelines for screening for breast cancer. Is a revision in order? Radiol Clin North Am 1992;30:221–33.
(11) Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994;191:241–4.
(12) Ikeda DM, Andersson I, Wattsgard C, Janzon L, Linell F. Interval carcinomas in the Malmo Mammographic Screening Trial: radiographic appearance and prognostic considerations. AJR Am J Roentgenol 1992;159: 287–94.
(13) Breast Screening Programme 2000. Reducing the risk. National Health Service (NHS) Cancer Screening Programmes. Sheffield (U.K.): National Health Service Breast Screening Programme Publications; 2000.
(14) Parker SH, Burbank F, Jackman RJ, Aucreman CJ, Cardenosa G, Cink TM, et al. Percutaneous large-core breast biopsy: a multi-institutional study. Radiology 1994;193:359–64.
(15) Brenner RJ. Surgical malignancy rate in women who have undergone needle core biopsy [letter]. Radiology 1996;200:283–4.
(16) 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;270:2444–50.
(17) Bassett LW, Butler DL. Mammography and early breast cancer detection. Am Fam Physician 1991;43:547–57.
(18) Mondor M. Failure to diagnose rising to number one on malpractice charts. Healthcare Review 1999;9:18.
(19) Houn F, Brown ML. Current practice of screening mammography in the United States: data from the National Survey of Mammography Facilities. Radiology 1994;190:209–15.
(20) Gale AG, Cowley HC. Breast cancer screening: comparison of radiologists' performance in a self-assessment scheme and in actual breast screening. In: Krupinski EA, editor. Medical imaging 1999: image perception and performance. Proc SPIE 1999;3663:157–68.
(21) Gale AG, Cowley HC, Wilson AR. Mammographic training sets for improving breast cancer detection. In: Kundel HL, editor. Medical imaging 1996: image perception. Proc SPIE 1996;2712:102–12.
(22) Gale AG, Cowley HC. Analysis of breast cancer screening results. In: Doi K, Giger ML, Nishikawa RM, Schmist RA, editors. Digital mammography '96. Amsterdam (The Netherlands): Elsevier Science; 1996. p. 27–31.
(23) Gale AG, Savage CJ, Pawley EF, Wilson AR, Roebuck EJ. Breast screening: visual search and observer performance. In: Kundel HL, editor. Medical imaging 1994: image perception. Proc SPIE 1994;2166:66–75.
(24) Gale AG, Wilson AR, Roebuck EJ. Mammographic screening: radiological performance as a precursor to image processing. In: Acharya RS, Goldof DB, editors. Biomedical image processing and biomedical visualisation. Proc SPIE 1993;1905:458–64.
(25) Baker JA, Kornguth PJ, Floyd CE Jr. Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description. AJR Am J Roentgenol 1996;166:773–8.
(26) Hanley JA. Alternative approaches to receiver operating characteristic analyses [editorial]. Radiology 1988;168:568–70.
(27) Berbaum KS, Dorfman DD, Franken EA Jr. Measuring observer performance by ROC analysis. Indications and complications. Invest Radiol 1989;24:228–33.
(28) Dorfman DD, Alf E Jr. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals-rating-method data. J Math Psychol 1969;6:487–96.
(29) Dorfman DD, Berbaum KS. Degeneracy and discrete receiver operating characteristic rating data. Acad Radiol 1995;2:907–15.
(30) Dorfman DD, Berbaum KS. A contaminated binormal model for ROC data: Part II. A formal model. Acad Radiol 2000;7:427–37.
(31) Kirk RE. Experimental design. 2nd ed. Belmont (CA): Wadsworth; 1982. p. 49–170.
(32) Scheffe H. The analysis of variance. New York (NY): Wiley; 1949.
(33) Levene H. Robust tests for equality of variance. In: Olkin I, editor. Contributions to probability and statistics. Palo Alto (CA): Stanford University Press; 1960.
(34) Brown MB, Forsythe AB. The small sample behavior of some statistics which test the equality of several means. Technometrics 1974;16: 129–32.
(35) Brown MB, Forsythe AB. Robust tests for the equality of variances. J Am Stat Assoc 1974;69:364–7.
(36) Hayes WL. Statistics. 3rd ed. New York (NY): Holt, Rinehart, and Winston; 1981. p. 435, 437.
(37) Kan L, Olivotto IA, Warren Burhenne LJ, Sickles EA, Coldman AJ. Standardized abnormal interpretation and cancer detection ratios to assess reading volume and reader performance in a breast screening program. Radiology 2000;215:563–7.
(38) Wennberg JE, Cooper MM, editors, The Dartmouth atlas of health care. Chicago (IL): American Hospital Association Press; 1999.
(39) Herzlinger R. Market-driven health care: who wins, who loses in the transformation of America's largest service industry. Cambridge (MA): Perseus Publishing; 1997.
(40) Thiemann DR, Coresh J, Oetgen WJ, Powe NR. The association between hospital volume and survival after acute myocardial infarction in elderly patients. N Engl J Med 1999;340:1640–8.
(41) Hannan EL, Racz M, Ryan TJ, McCallister BD, Johnson LW, Arani DT, et al. Coronary angioplasty volume–outcome relationships for hospitals and cardiologists. JAMA 1997;277:892–8.
(42) Luft HS, Bunker JP, Enthoven AC. Should operations be regionalized? The empirical relation between surgical volume and mortality. N Engl J Med 1979;301:1364–9.
(43) Sosa JA, Bowman HM, Gordon TA, Bass EB, Yeo CJ, Lillemoe KD, et al. Importance of hospital volume in the overall management of pancreatic cancer. Ann Surg 1998;228:429–38.
(44) Gordon TA, Bowman HM, Bass EB, Lillemoe KD, Yeo CJ, Heitmiller RF, et al. Complex gastrointestinal surgery: impact of provider experience on clinical and economic outcomes. J Am Coll Surg 1999;189:46–56.
(45) Gordon TA, Burleyson GP, Tielsch JM, Cameron JL. The effects of regionalization on cost and outcome for one general high-risk surgical procedure. Ann Surg 1995;221:43–9.
(46) Edwards EB, Roberts JP, McBride MA, Schulak JA, Hunsicker LG. The effect of the volume of procedures at transplantation centers on mortality after liver transplantation. N Engl J Med 1999;341:2049–53.
(47) Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et al. Breast cancer screening with mammography: overview of Swedish randomised trials. Lancet 1993;342:973–8.
(48) Lidbrink EK, Tornberg SA, Azavedo EM, Frisell JO, Hjalmar ML, Leifland KS, et al. The general mammography screening program in Stockholm. Organisation and first-round results. Acta Oncol 1994;33: 353–8.
(49) Tabar L, Fagerberg G, Duffy SW, Day NE. The Swedish two county trial of mammographic screening for breast cancer: recent results and calculation of benefit. J Epidemiol Community Health 1989;43:107–14.
(50) Frisell J, Lidbrink E, Hellstrom L, Rutqvist LE. Followup after 11 years—update of mortality results in the Stockholm mammographic screening trial. Breast Cancer Res Treat 1997;45:263–70.
(51) Cyrlak D. Induced costs of low-cost screening mammography. Radiology 1988;168:661–3.
(52) McLelland R, Pisano ED. The politics of mammography. Radiol Clin North Am 1992;30:235–41.
(53) Elmore JG, Barton MB, Moceri VM, Polk S, Arena PJ, Fletcher SW. Ten-year risk of false positive screening mammograms and clinical breast examinations. N Engl J Med 1998;338:1089–96.
(54) Olsson P, Armelius K, Nordahl G, Lenner P, Westman G. Women with false positive screening mammograms: how do they cope? J Med Screen 1999;6:89–93.
(55) Gram IT, Lund E, Slenker SE. Quality of life following a false positive mammogram. Br J Cancer 1990;62:1018–22.
(56) Salzmann P, Kerlikowske K, Phillips K. Cost-effectiveness of extending screening mammography guidelines to include women 40 to 49 years of age. Ann Intern Med 1997;127:955–65.
(57) Olsen O, Gotzsche PC. Cochrane review on screening for breast cancer with mammography. Lancet 2001;358:1340–2.
(58) Breast Cancer Screening. NIH Consensus Statement. 1977;1:5–8.

NOTES

The U.S. and U.K. teams contributed equally to the work presented in this article.

Supported by a grant from the California Breast Cancer Research Program.

We acknowledge the hard work and dedication of Dr. Jan Patterson; research assistants Abby Sokoloff and Katherine Himes; Swedish radiologists Drs. Edward Azavedo and Gunilla Svane (Chief, Section of Breast Imaging), Karolinska Hospital, Stockholm; and Dr. Ingvar Andersson, Chief of Radiology, Lund University, Malmo. We also thank Dr. Robin Wilson, Nottingham Breast Screening Training Center, U.K., for helping us to coordinate the use of the PERFORMS 2 dataset and Alex McMillan for giving us the biostatistical support on survey analysis. We are also grateful to all of the California radiologists who offered us their time in completing the PERFORMS 2 test.

Manuscript received March 6, 2000; revised December 31, 2001; accepted January 7, 2002.