Published in final edited form as:
J Natl Cancer Inst. 2002 September 18; 94(18): 1373–1380.

Screening Mammograms by Community Radiologists: Variability in False-Positive Rates

Joann G. Elmore,
Department of Medicine, University of Washington School of Medicine, Seattle

Diana L. Miglioretti,
Center for Health Studies, Group Health Cooperative of Puget Sound, and Department of Biostatistics, University of Washington School of Public Health, Seattle

Lisa M. Reisch,
Department of Medicine, University of Washington School of Medicine, Seattle

Mary B. Barton,
Department of Ambulatory Care and Prevention, Harvard Pilgrim Health Care, and Harvard Medical School, Boston

William Kreuter,
Department of Medicine, University of Washington School of Medicine, Seattle

Cindy L. Christiansen, and
Boston University School of Public Health, Health Services Department, and Center for Health Quality, Outcomes and Economic Research at Veterans Affairs Health Services Research and Development, Boston, MA

Suzanne W. Fletcher
Department of Ambulatory Care and Prevention, Harvard Pilgrim Health Care, and Harvard Medical School, Boston

Abstract

Background—Previous studies have shown that the agreement among radiologists interpreting a test set of mammograms is relatively low. However, data available from real-world settings are sparse. We studied mammographic examination interpretations by radiologists practicing in a community setting and evaluated whether the variability in false-positive rates could be explained by patient, radiologist, and/or testing characteristics.

Methods—We used medical records on randomly selected women aged 40–69 years who had had at least one screening mammographic examination in a community setting between January 1, 1985, and June 30, 1993. Twenty-four radiologists interpreted 8734 screening mammograms from 2169 women. Hierarchical logistic regression models were used to examine the impact of patient, radiologist, and testing characteristics. All statistical tests were two-sided.

Results—Radiologists varied widely in mammographic examination interpretations, with a mass noted in 0%–7.9%, calcification in 0%–21.3%, and fibrocystic changes in 1.6%–27.8% of mammograms read. False-positive rates ranged from 2.6% to 15.9%. Younger and more recently trained radiologists had higher false-positive rates. Adjustment for patient, radiologist, and testing

Correspondence to: Joann G. Elmore, M.D., M.P.H., Associate Professor, Division of General Internal Medicine, University of Washington School of Medicine, Harborview Medical Center, 325 Ninth Ave., Box 359780, Seattle, WA 98104–2499 (jelmore@u.washington.edu).

characteristics narrowed the range of false-positive rates to 3.5%–7.9%. If a woman went to two randomly selected radiologists, her odds, after adjustment, of having a false-positive reading would be 1.5 times greater for the radiologist at higher risk of a false-positive reading, compared with the radiologist at lowest risk (95% highest posterior density interval [similar to a confidence interval] = 1.17 to 2.08).

Conclusion—Community radiologists varied widely in their false-positive rates in screening mammograms; this variability range was reduced by half, but not eliminated, after statistical adjustment for patient, radiologist, and testing characteristics. These characteristics need to be considered when evaluating false-positive rates in community mammographic examination screening.

Despite many recent improvements in mammography (1), the ultimate interpretation still depends on individual physicians. The level of agreement among radiologists interpreting the same test set of mammograms is relatively low (2–6), which may delay the detection of breast cancer (7). However, recent data have shown that mammography test sets may not adequately represent actual clinical practice in a community setting (8). Few studies of variability have been done in the community setting. One study (9) found variability among radiologists' recommendations for biopsy, with radiologists in academic settings having a higher positive predictive value in their recommendations to undergo biopsy compared with community radiologists. This community-based study, however, did not control for possible differences in the patient populations or for differences among radiologists other than their affiliation with an academic institution.

In our previous work (10), we estimated that a woman's cumulative risk of experiencing at least one false-positive interpretation after 10 mammograms was approximately 50%. Several variables predicted the likelihood of a woman having a false-positive result (11). Risk ratios for a false-positive screening result increased with younger age of the woman, family history of breast cancer, use of hormone replacement therapy (HRT), time between screenings, no comparison with previous mammograms, and the radiologists' tendency to call mammograms abnormal. The single largest predictor, noted in our earlier work (11), was the radiologist's individual tendency to call mammograms abnormal.

The present study was designed to explore in more detail the extent of variability among radiologists in a community setting. Our goals were to 1) describe variability among radiologists in their specific observations, interpretations, and false-positive rates in screening mammograms; 2) evaluate the impact on variability of additional individual characteristics of the patients and the radiologists (i.e., sex, age, experience) and of additional testing characteristics (i.e., year of the mammogram, health maintenance organization [HMO] versus community facility); and 3) determine if the variability noted among radiologists would be reduced after adjusting for differences in patients, radiologists, and testing characteristics.

Methods

Setting

This retrospective cohort study was conducted on women enrolled in Harvard Pilgrim Health Care, a large HMO in New England. The study design has been previously reported and is described here in brief (10,11). The HMO has encouraged women aged 40 years and older to undergo routine breast cancer screening at both HMO and local community radiology centers. Radiologists interpreting mammograms were board-certified and worked in professional associations that contracted with the HMO.

Study Population

Female members of the HMO between 40 and 69 years of age on July 1, 1983, were potentially eligible for the study (n = 14 382). Women were excluded for the following reasons: a lapse in enrollment in the HMO between July 1, 1983, and June 30, 1995 (n = 8816); health coverage from a source other than Harvard Pilgrim Health Care or from a noncomputerized HMO center during the study period (n = 1093); and a history of breast cancer, prophylactic mastectomy or breast implants before July 1, 1983 (n = 146), or a prophylactic mastectomy or breast implants during the study period (n = 8). From the cohort of 4319 eligible women, a random sample was chosen consisting of 1200 women 40–49 years of age, 600 women 50–59 years of age, and 600 women 60–69 years of age, for a total eligible sample of 2400 women.

We excluded the data on 302 mammograms done prior to 1985, because time since the previous mammographic examination could not be calculated for most mammograms obtained in the first 18 months of the study. We note that the false-positive rate for this subset was lower than that for the remainder of the mammograms (2.0% versus 6.4%, respectively). The final study period for abstraction of screening visit data was therefore 8.5 years (January 1, 1985, to June 30, 1993) with a 2-year follow-up period for assessment of breast cancer outcomes (July 1, 1993, to June 30, 1995).

This study was approved by the Human Studies Committee of Harvard Pilgrim Health Care and the University of Washington School of Medicine.

Data Collection

Harvard Pilgrim Health Care uses computerized records for ambulatory care services (12,13). Data on demographic characteristics, breast cancer risk factors, screening mammograms, and breast cancer outcome were extracted from these records onto standardized forms. The diagnostic interpretations for mammography were classified as normal, abnormal–probably benign, abnormal–indeterminate, or abnormal–suspicious for cancer. The radiologists' recommendations for additional testing, including physical examination by the primary care provider or surgeon, diagnostic mammography within the subsequent 12 months, ultrasound examination, and biopsy were recorded.

Information on the radiologists was obtained from the Massachusetts State Medical Registry and from HMO administrative files. Data included sex, year of birth, and year of graduation from medical school. Mammography testing characteristics included the year of the mammographic examination (1985 through 1987, 1988 through 1990, 1991 through 1993), prior mammographic examination available for comparison (yes versus no/unknown), facility type where more than 50% of radiologists' clinical time occurred (HMO versus community), and time since previous mammogram. Time since previous mammographic examination was defined as ≤18 months, >18 months, or unknown/no previous mammograms. Menopausal status, if unknown, was estimated based on the median age for the cohort with known status. If no family history of breast cancer was noted, it was assumed that there was none.

Definition of Screening Mammograms and Accuracy

Screening mammograms were defined as those performed on asymptomatic women without previously noted abnormalities. Mammograms performed because of abnormalities noted by clinicians or patients or noted on previous mammograms were classified as diagnostic exams. Measures of accuracy were defined in a manner consistent with current recommendations regarding mammography audits (14–16) and with those used by others (17–20). A mammographic examination was classified as positive if the results were

indeterminate or suspicious for cancer, or if there was a recommendation for nonroutine follow-up, including physical examination, diagnostic mammographic examination within the next 12 months, ultrasound, or biopsy. A positive test was classified as true-positive if breast cancer (invasive or ductal carcinoma *in situ*) was diagnosed in the patient on the basis of pathologic findings within 1 year of the test and as false-positive otherwise.

Statistical Analysis

A total of 93 radiologists interpreted screening mammograms for the women included in this study. The number of mammograms interpreted by each radiologist ranged from 1 to 2036. Estimates of accuracy by individual radiologists may be unreliable for radiologists reading a small number of study films. Therefore, only the 24 radiologists who each read more than 50 screening mammograms were included in this analysis. These 24 radiologists interpreted screening mammograms on 2169 of the 2400 eligible women in the initial cohort; 45 of the 2169 women were subsequently diagnosed with breast cancer. Because estimates of sensitivity and true-positive rates may be unreliable, given the relatively low number of breast cancer cases in this study, only the false-positive rates were determined for each radiologist. Among the 24 radiologists, the percentage of mammograms with specific observations, diagnostic interpretations, and recommendations were noted, and results were presented for the median and range.

We estimated the effects of patient, radiologist, and testing characteristics on false-positive rates by using hierarchical logistic regression. The outcome of interest was the probability of a false-positive reading (versus a true-negative reading). Hierarchical logistic regression is similar to standard logistic regression except that we included woman-specific and radiologist-specific effects to account for the correlation between multiple readings within the same woman and by the same radiologist. We fit separate hierarchical logistic regression models for each variable of interest and for two multivariable models. The first multivariable model adjusted for patient and testing characteristics only; the second full model additionally adjusted for radiologist characteristics.

Through the inclusion of radiologist-specific effects, the models estimated adjusted false-positive rates for each radiologist by taking a weighted average of the radiologist's adjusted rate (given the covariates) and the overall mean rate. The weight given to the radiologist's rate depends on the number of mammograms read by that radiologist—more film readings resulted in more weight being given to the radiologist's rate and less weight being given to the overall average. In this way, the model indirectly adjusted for the number of films read by each radiologist. The false-positive rates are adjusted for the covariates included in the model. For example, if a radiologist tended to see many women with risk factors associated with higher false-positive rates, then that radiologist's adjusted false-positive rate would be lower than his or her observed false-positive rate.

Hierarchical models gave direct estimates of subject-specific (conditional) means and odds ratios (ORs), which measured the expected value for an individual woman; however, in this study, we were interested in population (marginal) averages, which estimated the average across a population of women. [For a discussion on the differences between subject-specific and population averages, *see* (21–23)]. Therefore, we estimated the population average false-positive rates and ORs from the conditional estimates by using Monte Carlo integration. The radiologist-specific effects may also be used to examine the heterogeneity between false-positive rates among radiologists. To quantify this heterogeneity, we calculated the average OR between any two radiologists, comparing the one having a higher false-positive rate with the one having a lower false-positive rate (24).

The hierarchical logistic regression models were fit using Bayesian Inference Using Gibbs Sampling (BUGS) (25). The regression coefficients were taken to be Gaussian with zero mean and precision of 1×10^{-6}. The population variances were taken to be gamma (0.01, 0.01). Following procedures that are commonly used with Gibbs sampling, we ran the single-variable logistic regressions for 25 000 iterations, discarded the first 5000 iterations, and kept every 20^{th} iteration of the remaining 20 000, for a total of 1000 samples from the posterior distribution. The full models were run for 100 000 iterations with 10 000 burn-in iterations, thinned by 90 iterations. Convergence of the Gibbs samplers was assessed by examining the trace plots. For the statistics of interest, we report the posterior mode of the population averages and the 95% highest posterior density intervals (95% HPD), which are similar to classical 95% confidence intervals (CIs).

We included the following patient variables in the analyses: patients' age at the time of the mammogram, menopausal status, HRT use (current, previous, or never), family history of breast cancer (yes versus no/unknown), history of breast aspirate or biopsy (none since the start of study versus one or more), body mass index (BMI) at the time of the mammographic examination (BMI ≤25 kg/m^2 versus >25 kg/m^2) and race (white, black, other, or unknown). Radiologists' characteristics included age of the radiologist, the number of years since graduation from medical school, and sex (male versus female). Testing characteristics included the year of the mammographic examination in three categories for parsimony in the full model (1985 through 1987, 1988 through 1990, and 1991 through 1993), whether the radiologist indicated that a prior mammographic examination was available for comparison (no/unknown versus yes), time since previous mammographic examination (≤18 months, >18 months, or unknown/no previous mammograms), and facility type (HMO versus community).

Race was not included in the full models because there were 747 women with unknown race, and using a missing category in multiple regression can bias results (26,27); however, there were no differences in the false-positive rate by race in the unadjusted model. The 149 women with missing BMI were excluded from the full models.

Results

Characteristics of Women Screened

Over the 8.5-year study period, the 24 radiologists interpreted 8734 screening mammograms obtained on 2169 women. The median number of mammograms per woman was 4 (range = 1–9). Most of the women (78.9%) were white; 10.0% were black, 2.5% were of other races, and 8.6% were of unknown race. A family history of breast cancer was recorded for 19.7% of the women. Current HRT use was reported at some time during the study period for 12.1% of the women, and previous use was reported for 7.5%. Forty-eight percent of the women were overweight or obese (BMI >25 kg/m^2).

Breast cancer was diagnosed in 45 women during the study period: local disease was present in 38 women and regional disease was present in seven women. The mean age at diagnosis of the women with breast cancer was 60 years (range = 45–76). Ductal carcinoma *in situ* was diagnosed in seven women. In 35 women, breast cancer was diagnosed as a result of an abnormality first noted on a screening mammogram.

Characteristics of the Radiologists

The 24 radiologists worked at nine different radiology facilities, consisting of two community and seven HMO sites. The median number of mammograms interpreted per radiologist was 111 (range = 59–1990). Four radiologists each interpreted more than 1000 mammograms; one radiologist interpreted 620 mammograms and the others interpreted

between 59 and 292 mammograms. The median age of the 24 radiologists at the time of interpreting their first screening mammographic examination in the cohort was 48 years (range = 31–70). Four radiologists were women. The mean number of years between graduating from medical school and interpreting their first mammographic examination for the members of the study cohort was 23 years (range = 5–46).

Variability in Observations, Interpretations, and Management Recommendations

The observations, diagnostic interpretations, and specific recommendations for management made by the radiologists are shown in Table 1. A mass was reported by the 24 radiologists in a median of 2.3% of films interpreted, with the range being from 0% of cases interpreted by one radiologist to 7.9% of cases interpreted by another radiologist. There were wide ranges among radiologists in their notation of the presence of calcifications, fibrocystic changes, and other abnormalities. For example, one of the 24 radiologists did not observe calcifications in any film, whereas another radiologist noted the presence of calcifications in 21.3% of the films read. A wide range was also noted in the diagnostic interpretation categories of normal (range = 55.1%–83.6%) and abnormal benign (range = 6.0%–39.3%), although there was much less variability in the abnormal category suggestive of cancer (range = 0.5%–2.7%). The largest variability in recommendations was in suggesting that additional mammographic views be ordered (1.1% for one radiologist to 11.0% for another radiologist).

Predictors of False-Positive Rate

The observed false-positive rates of the 24 radiologists ranged from 2.6% (95% CI = 0.3% to 9.0%) to 15.9% (95% CI = 8.7% to 25.6%) and are shown graphically in Fig. 1. While the 95% CIs for the two extreme radiologists do overlap, the 95% CIs for false-positive rates from other radiologists who read more films do not overlap. For example, for the radiologist with a false-positive rate of 2.7%, the 95% CI was 1.2% to 5.3% and for the radiologist with a false-positive rate of 15.9% the 95% CI was 8.7% to 25.6%.

Table 2 shows the association of patient, radiologist, and testing characteristics with false-positive interpretations for the unadjusted and full hierarchical logistic regression models. Those women who were younger, premenopausal, using HRT at the time of the mammogram, had a positive family history of breast cancer, or had a history of previous biopsy were more likely to have a false-positive screening test result. Films interpreted by younger radiologists and by radiologists who graduated from medical school within the past 15 years were also more likely to have a false-positive result.

A secular trend was noted, with women who had mammograms in the 1990s being more likely to have a false-positive result than women who had mammograms in the 1980s. Screening mammograms for which radiologists noted a prior film available for comparison had a false-positive rate of 5.4% compared with a 9.0% false-positive rate for screens without prior films. Women who had mammograms within 18 months of previous mammograms were less likely to have a false-positive result compared with those waiting longer between screens or not having any prior screens.

Fig. 2 shows the observed and adjusted false-positive rates for the 24 radiologists in the study. The mean unadjusted ORs for all possible pairwise comparisons among radiologists (comparing the radiologist at higher risk of a false-positive reading with the radiologist at lower risk) is 2.05 (Fig. 2, line A). In other words, if a woman went to two randomly picked radiologists, her odds of having a false-positive reading would be 2.05 times greater on average for the radiologist at higher risk compared with the radiologist at lower risk. Some of this variability is due to the small number of films read by certain radiologists (i.e., <100

mammograms). However, after accounting for the correlation within woman and radiologist using hierarchical logistic regression, which indirectly adjusts for the varying number of mammograms read by each radiologist and each woman's overall tendency for having a false-positive mammogram, the false-positive rates ranged from 3.5% to 11.9%, with a mean OR between radiologists of 1.68 (95% HPD = 1.33 to 2.42; Fig. 2, line B). Adjusting for patient and testing characteristics in addition to the correlation within woman and radiologist did not further reduce the variability in false-positive rates between radiologists (Fig. 2, line C; range = 3.3%–10.2%; mean OR = 1.65, 95% HPD = 1.33 to 2.44). However, after additionally adjusting for radiologists' characteristics, the range of false-positive rates was reduced to 3.5%–7.9%, and the mean OR between radiologists was 1.48 (95% HPD = 1.17 to 2.08; Fig. 2, line D).

Discussion

Community radiologists varied substantially in their interpretation of screening mammograms; the variability in false-positive mammography rates was reduced by half, but not eliminated, after adjustment for differences in the patient population, the testing situation, and radiologists' characteristics. Before adjustments, the 24 radiologists varied in their false-positive interpretation rates from 2.6% to 15.9%; after full adjustment for patient, testing, and radiologist characteristics that may influence false-positive readings, variability was reduced to a range of 3.5%–7.9%. While patient, testing, and radiologists' characteristics were all important predictors of false-positive rates, radiologist characteristics were more important in accounting for variability among radiologists in this study than we had anticipated. The unexpected importance of radiologist characteristics was probably due to the similarity of patient populations and testing characteristics across radiologists in this study. However, these characteristics may not be similar in other studies; therefore, it will typically be important to adjust for all of these variables when studying radiologists' variability.

The most important radiologist characteristic appeared to be age and time since graduation from medical school, with younger radiologists and those more recently in training having higher rates of false-positive mammograms. The fact that younger radiologists and those more recently trained had two to four times the false-positive mammographic examination rates of older radiologists (Table 2) is especially noteworthy, because it is reasonable to hypothesize that those most recently trained would be more accurate than older mammographers, i.e., those trained a long time ago. It is possible that the younger radiologists missed fewer cancers than did older mammographers who were more distant from their training, because their training emphasized sensitivity over specificity.

Variability has been noted in many areas of clinical medicine (28,29). Microscopic review of breast tissue slides has an element of subjectivity in interpretation similar to that of interpretation of mammograms. For example, in the diagnosis of ductal carcinoma *in situ*, agreement among five pathologists with a standard interpretation on a test set of 24 breast tissue slides ranged from 71% to 92%, with individual false-positive rates ranging from 0% to 20% (30). Obviously, the CIs around the individual rates would be wide, given the small sample size, but the similarities with our findings in mammography are striking.

Several studies (2–7) have indicated that significant variability exists in the interpretation of mammograms. This variability indicates the possibility of wide ranges in false-positive mammogram interpretations by individual radiologists, which can be both alarming and expensive for the patient (10). By better understanding sources of variability in mammography interpretation, we can identify potential areas of improvement. The ultimate

goal is to enhance mammography performance by reducing the rate of false-positive interpretations while maintaining high levels of sensitivity and accuracy.

It has long been known that certain clinical and demographic characteristics of women make accurate reading of mammograms more difficult (31–33). More recently, several studies (34,35) have shown that time between mammograms and the availability of previous studies for comparison also affect accuracy. However, less attention has been directed to secular trends in false-positive mammographic examination rates. We found that rates almost doubled in this community setting between 1985 and 1993. This increase in false-positive rates may be related to fear of medical malpractice litigation, given the prominence in North America of malpractice litigation for delayed detection of breast cancer.

Strengths of this study include the fact that it was done within a community setting and with radiologists who had a broad range of years of experience and who had worked in different types of clinical settings. Data were available on the radiologists, the patients, and the testing characteristics, all of which were controlled for in the analysis. Most of the prior studies of radiologists' variability in mammography have been done in a testing situation, which might not be representative of real-life clinical practice (8).

The limitations of our study include the fact that the radiologists in this study did not read the same films, and so direct comparisons are not possible (although we did adjust for patient characteristics in the models). Only 45 women were diagnosed with breast cancer; thus, we did not analyze sensitivity. In addition, some of the radiologists read fewer than 100 mammograms in the 8.5-year study period, which makes comparisons difficult because the CIs were wide. It should be noted, however, that these radiologists read additional films outside this study cohort; thus, the numbers do not represent the total number of mammograms they read during the study period. In addition, the American College of Radiology breast imaging reporting and data system (BI-RADS^{TM}) classification system was not in use at the time of the study (36). Although use of BI-RADS^{TM} may ultimately lead to less variability among radiologists, this has not yet been shown to be the case (5). The false-positive rates for our participating radiologists were lower than the national average; thus, our results possibly underestimate the variability among radiologists elsewhere. Finally, the data in this study are for 1985 through 1993, and reading patterns among radiologists may have changed since then.

Given the retrospective nature of this study, data on some variables were not available, which may have resulted in misclassification errors. For example, several factors related to radiologists that might be important and should be included in future research include fiscal incentives, medical malpractice concerns, and comfort with ambiguity in clinical decision making. Adjustments for these and other variables may further decrease the variability in false-positive rates.

In summary, community radiologists varied widely in their false-positive rates for screening mammograms. This variability was affected not only by the kind of patients seen but also by radiologists' age and experience. Younger radiologists and those more recently in training had higher rates of false-positive mammogram interpretations. This study was different from research designs that used test sets of films, because we looked at radiologists' decisions as they naturally occur in actual clinical practice. That the variability among radiologists in false-positive mammographic examination readings was reduced by half underscores the importance of adjusting for patient and radiologist characteristics when attempting to understand variability in clinical medicine.

Acknowledgments

Supported by grants from the American Cancer Society (to J. Elmore); by Public Health Service grant HS-10591 (to J. Elmore) from the Agency for Healthcare Research and Quality and the National Cancer Institute, National Institutes of Health, Department of Health and Human Services; by a Robert Wood Johnson Generalist Faculty Scholar Award (to J. Elmore); and by the Harvard Pilgrim Health Care Foundation (S. Fletcher and M. Barton).

References

1. Houn F, Elliott ML, McCrohan JL. The Mammography Quality Standards Act of 1992. History and philosophy. Radiol Clin North Am. 1995; 33:1059–65. [PubMed: 7480655]
2. Ciccone G, Vineis P, Frigerio A, Segnan N. Inter-observer and intraobserver variability of mammographic examination interpretation: a field study. Eur J Cancer. 1992; 28A:1054–8. [PubMed: 1627374]
3. Vineis P, Sinistrero G, Temporelli A, Azzoni L, Bigo A, Burke P, et al. Inter-observer variability in the interpretation of mammograms. Tumori. 1988; 74:275–9. [PubMed: 3400118]
4. Elmore J, Wells C, Lee C, Howard D, Feinstein A. Variability in radiologists' interpretations of mammograms. N Engl J Med. 1994; 331:1493–9. [PubMed: 7969300]
5. Kerlikowske K, Grady D, Barclay J, Frankel SD, Ominsky SH, Sickles EA, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. J Natl Cancer Inst. 1998; 90:1801–9. [PubMed: 9839520]
6. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Arch Intern Med. 1996; 156:209–13. [PubMed: 8546556]
7. Baines C, McFarlane D, Miller A. The role of the reference radiologist: estimates of interobserver agreement and potential delay in cancer detection in the national breast screening study. Invest Radiol. 1990; 25:971–6. [PubMed: 2211054]
8. Rutter CM, Taplin SH. Assessing mammographers' accuracy. A comparison of clinical and test performance. J Clin Epidemiol. 2000; 53:443–50. [PubMed: 10812315]
9. Meyer JE, Eberlein TJ, Stomper PC, Sonnenfeld MR. Biopsy of occult breast lesions. JAMA. 1990; 263:2341–3. [PubMed: 2157903]
10. 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. [PubMed: 9545356]
11. Christiansen CL, Wang F, Barton MB, Kreuter W, Elmore JG, Gelfand AE, et al. Predicting the cumulative risk of false-positive mammograms. J Natl Cancer Inst. 2000; 92:1657–66. [PubMed: 11036111]
12. Barnett G. The application of computer-based medical-record systems in ambulatory practice. N Engl J Med. 1984; 310:1643–50. [PubMed: 6427610]
13. Barnett GO, Justice NS, Somand ME, Adams JB, Waxman BD, Beaman PD, et al. COSTAR: a computer based medical information system for ambulatory care. Proc IEEE. 1979; 67:1226–37.
14. Bassett, LW.; Hendrick, RE.; Bassford, TL.; Butler, PF.; Carter, DC.; DeBor, JD., et al. Clinical practice guideline No 13. Rockville (MD): Agency for Health Care Policy and Research (AHCPR); 1994. Quality determinants of mammography. AHCPR Publ No. 95–0632Available at: http://hstat.nlm.nih.gov/hq/Hquest/db/local.arahcpr.arclin.mamc
15. Linver MN, Osuch JR, Brenner RJ, Smith RA. The mammography audit: a primer for the mammography quality standards act (MQSA). AJR Am J Roentgenol. 1995; 165:19–25. [PubMed: 7785586]
16. Sickles EA. Quality assurance: how to audit your own mammography practice. Radiol Clin N Am. 1992; 30:265–75. [PubMed: 1732933]
17. Bird RE. Low-cost screening mammography: report on finances and review of 21,716 consecutive cases. Radiology. 1989; 171:87–90. [PubMed: 2494683]
18. Brown M, Goun F, Sickles E, Kessler L. Screening mammography in community practice: positive predictive value of abnormal findings and yield of follow-up procedures. AJR Am J Roentgenol. 1995; 165:1373–7. [PubMed: 7484568]

19. Kerlikowske K, Grady D, Barclay J, Sickles E, Ernster V. Likelihood ratios for modern screening mammography. Risk of breast cancer based on age and mammographic interpretation. JAMA. 1996; 276:39–43. [PubMed: 8667537]
20. Robertson CL. A private breast imaging practice: medical audit of 25,788 screening and 1,077 diagnostic examinations. Radiology. 1993; 187:75–9. [PubMed: 8451440]
21. Diggle, PJ.; Liang, KY.; Zeger, SL. Analysis of longitudinal data. New York (NY): Oxford University Press; 1994.
22. Hauck WW, Neuhaus JM, Kalbfleisch JD, Anderson S. A consequence of omitted covariates when estimating odds ratios. J Clin Epidemiol. 1991; 44:77–81. [PubMed: 1986061]
23. Heagerty PJ, Zeger SL. Multivariate continuation ratio models: connections and caveats. Biometrics. 2000; 56:719–32. [PubMed: 10985208]
24. Larsen K, Petersen JH, Budtz-Jorgensen E, Endahl L. Interpreting parameters in the logistic regression model with random effects. Biometrics. 2000; 56:909–14. [PubMed: 10985236]
25. Spielgelhalter, DJ.; Thomas, A.; Best, NG.; Gilkes, WR. BUGS 0.5: Bayesian Inference Using Gibbs Sampling. MRC Biostatistics Unit, Cambridge University; Cambridge (U.K.): 1996.
26. Jones MP. Indicator and stratification methods for missing explanatory variables in multiple linear regression. J Am Stat Assoc. 1996; 91:222–30.
27. Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. Am J Epidemiol. 1991; 134:895–907. [PubMed: 1670320]
28. Feinstein AR. A bibliography of publications on observer variability. J Chronic Dis. 1985; 38:619–32. [PubMed: 3894405]
29. Elmore J, Feinstein A. A bibliography of publications on observer variability. J Clin Epidemiol. 1992; 45:567–80. [PubMed: 1607896]
30. Schnitt SJ, Connolly JL, Tavassoli FA, Fechner RE, Kempson RL, Gelman R, et al. Interobserver reproducibility in the diagnosis of ductal proliferative breast lesions using standardized criteria. Am J Surg Pathol. 1992; 16:1133–43. [PubMed: 1463092]
31. Laya MB, Larson EB, Taplin SH, White E. Effect of estrogen replacement therapy on the specificity and sensitivity of screening mammography. J Natl Cancer Inst. 1996; 88:643–9. [PubMed: 8627640]
32. Baines CJ. Menstrual cycle variation in mammographic breast density: so who cares? J Natl Cancer Inst. 1998; 90:875–6. [PubMed: 9637132]
33. Brenner RJ, Pfaff JM. Mammographic changes after excisional breast biopsy for benign disease. AJR Am J Roentgenol. 1996; 167:1047–52. [PubMed: 8819410]
34. Tabar L, Gad A, Holmberg L, Ljungquist U. Significant reduction in advanced breast cancer. Results of the first seven years of mammography screening in Kopparberg, Sweden. Diagn Imaging Clin Med. 1985; 54:158–64. [PubMed: 3896614]
35. 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; 164:1107–9. [PubMed: 7717214]
36. American College of Radiology (ACR). Breast imaging reporting and data system (BI-RADS™). 3rd. Reston (VA): American College of Radiology; 1998.

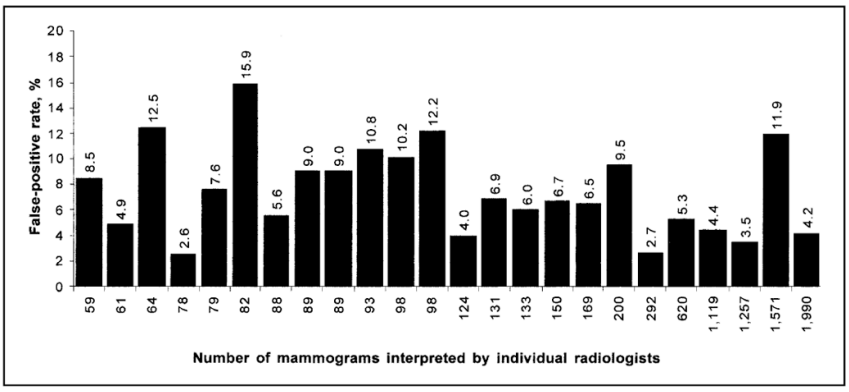

Fig. 1.

Observed false-positive rates for 24 radiologists reading 8734 mammograms. Each column represents a single radiologist. The number of mammograms interpreted by a given radiologist is given at the base of the column and the false-positive rate is plotted on the ***y*-axis**, with the exact number describing the false-positive rate for an individual radiologist given at the top of the respective column.

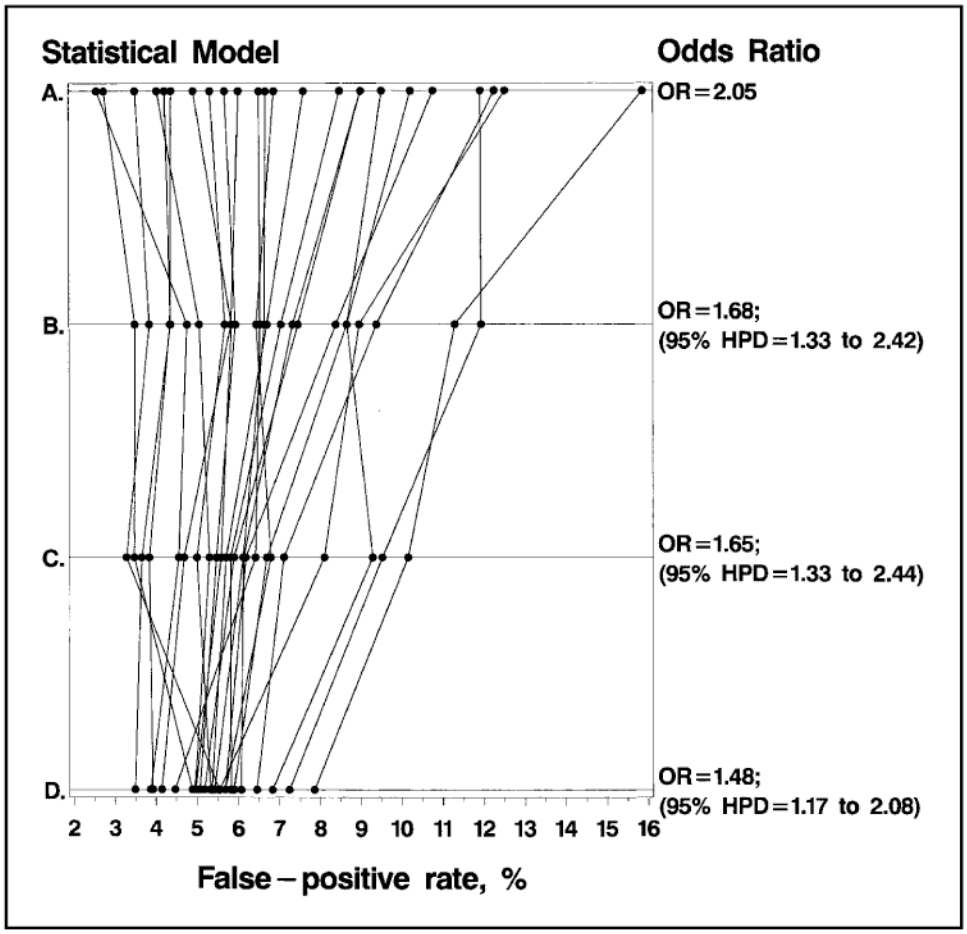

Fig. 2.

Results of statistical modeling for observed (unadjusted, **line A**) and adjusted (**lines B, C**, and **D**) false-positive rates for 24 radiologists. The details of the adjustments are given below. The **right side** shows the odds ratios (ORs) with 95% highest posterior density intervals (HPDs) (similar to classical 95% confidence intervals). **Line A** shows observed unadjusted false-positive rates and summary OR values. **Line B** shows false-positive rates and summary ORs after adjusting for correlation between multiple mammograms within the same woman and by the same radiologist. **Line C** shows ORs after adjustments given at **line B**, plus adjustment for patient characteristics and testing characteristics. **Line D** shows all adjustments at **line C** plus adjustment for radiologists' characteristics.

Table 1

Results of 24 radiologists' interpretations of screening mammograms in a community setting

	Individual radiologist's observations, interpretations, and recommendations given as % of the mammograms they interpreted	
Result	**Median**	**Range**
Observation noted		
Mass	2.3	0.0–7.9
Calcification	7.8	0.0–21.3
Fibrocystic changes	7.3	1.6–27.8
Other (e.g., skin change)	2.4	0.0–25.7
Diagnostic interpretation		
Normal	72.0	55.1–83.6
Abnormal, benign	22.0	6.0–39.3
Abnormal, indeterminate	3.3	1.0–8.5
Abnormal, suggestive of cancer	1.1	0.5–2.7
Recommendation		
Additional mammographic views	4.0	1.1–11.0
Ultrasound studies	2.2	0.0–4.7
Biopsy	0.8	0.0–3.4

Table 2

Patient, radiologist, and testing characteristics, false-positive rates, and odds ratios (ORs) (unadjusted and adjusted)* for 8734 screening mammograms on 2169 women read by 24 radiologists over an 8.5-year period

Characteristics	No. of mammograms	Population average false-positive rate (95% HPD)	Population average unadjusted OR (95% HPD)	Population average adjusted OR† (95% HPD)
Patient characteristics				
Age of patient at time of mammogram				
40s	1757	9.1 (6.6 to 11.3)	**1.96 (1.36 to 2.73)**	**1.55 (1.01 to 2.46)**
50s	3269	8.4 (6.5 to 10.8)	**1.85 (1.34 to 2.58)**	**1.86 (1.25 to 2.56)**
60s	2486	5.9 (4.2 to 7.6)	1.22 (0.87 to 1.72)	1.32 (0.88 to 1.85)
70s	1222	4.4 (3.1 to 6.6)	referent	referent
Menopause and hormone replacement therapy (HRT) use				
Premenopausal	1371	8.9 (6.9 to 12.0)	**1.45 (1.17 to 1.86)**	1.23 (0.89 to 1.71)
Postmenopausal				
Current HRT use	1059	8.8 (6.5 to 12.5)	**1.45 (1.09 to 1.81)**	1.34 (0.99 to 1.73)
Previous HRT use	651	7.7 (5.3 to 11.1)	1.19 (0.85 to 1.74)	1.12 (0.79 to 1.65)
Never use of HRT	5653	6.1 (5.1 to 8.1)	referent	referent
Family history of breast cancer				
Yes	1722	9.4 (7.1 to 12.1)	**1.41 (1.16 to 1.78)**	**1.44 (1.13 to 1.81)**
No/missing	7012	6.6 (5.4 to 8.4)	referent	referent
History of biopsy				
Yes (1 or more)	603	11.7 (8.7 to 16.2)	**1.79 (1.34 to 2.44)**	**1.84 (1.29 to 2.54)**
No	8131	6.7 (5.4 to 8.5)	referent	referent
BMI at time of mammogram‡				
Overweight/obese (>25 kg/m^2)	4149	7.8 (5.9 to 9.7)	1.17 (0.95 to 1.37)	1.17 (0.97 to 1.45)
Missing	149	6.4 (2.6 to 12.4)	0.85 (0.37 to 1.85)	N/A
Normal/underweight (≤25 kg/m^2)	4436	7.2 (5.2 to 8.7)	referent	referent
Race§				
Black	874	7.1 (5.1 to 9.9)	0.95 (0.70 to 1.31)	
Other	221	4.4 (2.1 to 8.9)	0.57 (0.27 to 1.16)	
Unknown	747	6.3 (4.5 to 9.4)	0.85 (0.62 to 1.24)	
White	6892	7.2 (5.8 to 9.3)	referent	N/A
Radiologist characteristics				
Age, No. of years since graduation from medical school				
30s, 5–15	1647	7.0 (4.9 to 9.5)	1.25 (0.82 to 1.80)	**2.08 (1.01 to 4.10)**
40s				
5–15	702	11.7 (8.8 to 16.1)	**2.18 (1.46 to 3.42)**	**3.87 (1.68 to 7.58)**
16–20	977	6.2 (4.5 to 9.5)	1.20 (0.79 to 1.79)	1.89 (0.84 to 4.20)
>20	2238	5.1 (3.6 to 7.3)	0.94 (0.68 to 1.27)	1.57 (0.75 to 3.58)
50s, >20	1571	6.7 (5.2 to 9.1)	**1.30 (1.07 to 1.53)**	2.10 (0.90 to 3.93)

Characteristics	No. of mammograms	Population average false-positive rate (95% HPD)	Population average unadjusted OR (95% HPD)	Population average adjusted OR† (95% HPD)
60s or 70s, >20	1599	5.3 (4.1 to 7.2)	referent	referent
Gender				
Male	8115	7.6 (5.9 to 9.4)	1.27 (0.78 to 2.94)	1.10 (0.62 to 2.44)
Female	619	5.4 (3.2 to 9.4)	referent	referent
Testing characteristics				
Year of mammogram				
1985–1987	2156	6.9 (5.4 to 9.3)	0.88 (0.68 to 1.08)	**0.55 (0.37 to 0.81)**
1988–1990	3386	6.4 (5.1 to 8.3)	**0.77 (0.64 to 0.93)**	0.77 (0.58 to 1.00)
1991–1993	3192	8.2 (6.3 to 10.3)	referent	referent
Prior mammogram available for comparison				
No/unknown	4062	9.0 (7.1 to 11.4)	**1.71 (1.43 to 2.05)**	**1.76 (1.39 to 2.10)**
Yes	4672	5.4 (4.2 to 7.1)	referent	referent
Time since previous mammogram				
Never/unknown	1696	9.4 (7.1 to 12.1)	**1.57 (1.24 to 2.00)**	**1.99 (1.38 to 2.66)**
>18 months	2732	7.5 (5.8 to 9.8)	**1.25 (1.04 to 1.51)**	**1.37 (1.07 to 1.68)**
<18 months	4306	6.1 (4.6 to 7.9)	referent	referent
Facility type				
Health maintenance organization	7669	7.8 (6.0 to 10.1)	1.28 (0.72 to 2.11)	1.28 (0.75 to 2.12)
Community	1065	5.5 (3.7 to 8.2)	referent	referent

*Statistically significant ORs for which the highest posterior density region (HPD) did not include 1 are shown in **bold** type. Unadjusted and adjusted ORs were obtained from models including woman- and radiologist-specific effects. BMI = body mass index; HRT = hormone replacement therapy.

†ORs are adjusted for all other variables in the full model.

‡Women with missing values for BMI were excluded from the full model.

§Race was not included in the full model because of the large number of women of unknown race.