Published in final edited form as:
J Natl Cancer Inst. 2003 September 17; 95(18): 1384–1393.

International Variation in Screening Mammography Interpretations in Community-Based Programs

Joann G. Elmore, **Connie Y. Nakano**, **Thomas D. Koepsell**, **Laurel M. Desnick**, **Carl J. D'Orsi**, and **David F. Ransohoff**
J. G. Elmore, C. Y. Nakano, L. M. Desnick (Department of Medicine), T. D. Koepsell (Department of Health Services and Epidemiology), University of Washington, Seattle; C. J. D'Orsi, Department of Radiology, Emory University, Atlanta, GA; D. F. Ransohoff, Department of Medicine, University of North Carolina at Chapel Hill

Abstract

Background—Variations in mammography interpretations may have important clinical and economic implications. To evaluate international variability in mammography interpretation, we analyzed published reports from community-based screening programs from around the world.

Methods—A total of 32 publications were identified in MEDLINE that fit the study inclusion criteria. Data abstracted from the publications included features of the population screened, examination technique, and clinical outcomes, including the percentage of mammograms judged to be abnormal, positive predictive value of an abnormal mammogram (PPV_A), positive predictive value of a biopsy performed (PPV_B), and percentages of breast cancer patients with ductal carcinoma *in situ* (DCIS) and minimal disease (DCIS and/or tumor size ≤10 mm). North American screening programs were compared with those from other countries using meta-regression analysis. All statistical tests were two-sided.

Results—Wide ranges were noted for the percentage of mammograms judged to be abnormal (1.2%–15.0%), for PPV_A (3.4%–48.7%), for PPV_B (5.0%–85.2%), for percentage diagnosed with DCIS (4.3%–68.1%), and for percentage diagnosed with minimal disease (14.0%–80.6%). The percentage of mammograms judged to be abnormal were 2–4 percentage points higher in North American screening programs than they were in programs from other countries, after adjusting for covariates such as percentage of women who were less than 50 years of age and calendar year in which the mammogram was performed. The percentage of mammograms judged to be abnormal had a negative association with PPV_A and PPV_B (both $P < .001$) and a positive association with the frequency of DCIS cases diagnosed ($P = .008$) and the number of DCIS cases diagnosed per 1000 screens ($P = .024$); no consistent relationship was observed with the proportion of breast cancer diagnoses reported as having minimal disease or the number of minimal disease cases diagnosed per 1000 screens.

Conclusion—North American screening programs appear to interpret a higher percentage of mammograms as abnormal than programs from other countries without evident benefit in the yield of cancers detected per 1000 screens, although an increase in DCIS detection was noted.

Substantial intra- and interobserver variability has been noted among radiologists interpreting screening mammograms in research situations (1–3). This variability is similar to that seen in other areas of medicine where observation and interpretation are subjective (4,5). Several studies in the United States (6–8) have suggested that variation in

Correspondence to: J. G. Elmore, MD, MPH, Department of Medicine, Division of General Internal Medicine, Harborview Medical Center, 325 Ninth Ave., Box 359780, Seattle, WA 98104-2499 (jelmore@u.washington.edu).

mammography interpretation also exists among radiologists in community-based facilities. One review (6) of U.S. screening programs found that the percentage of screening mammograms for which additional work-up is recommended (i.e., percentage of mammograms judged to be abnormal, often referred to as the recall rate) ranged from approximately 2% to more than 50%, with an average of 11%. Another study (7) found that the positive predictive value of a biopsy performed (PPV_B) (i.e., the percentage of women who were actually found to have breast cancer among those referred for biopsy following screening mammography) ranged from 17% for radiologists practicing in the community to 26% for radiologists practicing at an academic center. A recent study by Elmore et al. (8) reported that radiologists varied widely in their false-positive rates for interpretation of screening mammograms, even after controlling for patient, radiologist, and testing characteristics.

Variability in screening mammography interpretations may have important clinical and economic implications. Although clinicians do not wish to miss breast cancers, it is important to minimize unnecessary follow-up diagnostic procedures, costs, and patient anxiety associated with false-positive screening mammograms.

In this article, we compare published data from community-based mammography screening programs in North America with similar screening programs in other countries to address two important questions: 1) To what extent is variability in mammographic interpretation in community-based screening mammography programs observed between programs in North America and other countries? and 2) Is variability in mammographic interpretation associated with different intermediate measures of breast cancer outcome (i.e., percentage of breast cancer cases with ductal carcinoma *in situ* [DCIS] and/or minimal disease)? Based on our findings, we discuss possible explanations for variability in mammography interpretations and of the implications that this variability might have on future research, health policy, and patient care.

Methods

Search Methods and Selection Criteria

A MEDLINE search identified 227 English and non–English language candidate publications between January 1, 1985, and June 15, 2002. Search terms included 1) mammography, 2) mass screening, and 3) biopsy. In addition, the references cited within these publications were searched for any publications that may fit the study inclusion criteria. To be included in this study, a publication had to report results of screening mammography performed in or after 1985. Screening mammograms were defined as mammograms obtained on women with no known breast-related symptoms or abnormalities and no known breast cancer at the time of the screen. In addition, at least two of the three measures of physicians' interpretation (i.e., percentage of mammograms judged to be abnormal, positive predictive value of an abnormal mammogram [PPV_A], and PPV_B) and at least one of the two intermediate breast cancer outcomes (i.e., percentage of breast cancer cases with DCIS or percentage with minimal disease) had to be available. For breast cancer screening programs that had published more than one article, we attempted to select the most recent data covering the largest geographic area and the largest number of women. Results from randomized clinical trials and case–control studies were excluded from this study because data from research settings may have less generalizability to current community-based practice.

A total of 32 published articles met the study inclusion criteria. Each of these final publications was reviewed independently by at least two authors (J. G. Elmore, L. Desnick, and C. Y. Nakano) to determine if it met the study inclusion criteria. When information was

not clear from the publication, the authors of the article were contacted to obtain that information (*see* supplemental information available at http://jncicancerspectrum.oupjournals.org/jnci/content/vol95issue18/). When disagreements between the text of an article and the data presented in the tables were noted, we chose to use the data from the tables. Information on the characteristics of the population screened, features of the mammogram examination, and outcome of readings were abstracted from each publication that met the study inclusion criteria.

Characteristics of the Population Screened

Age of the women screened was abstracted from each article because mammography sensitivity, specificity, and positive predictive value have been found to increase in older women (9). Each article was also classified as including initial screens only, subsequent screens only, or mixed screens (i.e., both initial and subsequent). Results from first-time (e.g., initial) mammographic screening include prevalent cancers, so the percentage of mammograms judged to be abnormal and cancer rates may be expected to be higher than in subsequent screens of the same woman (9).

Features of the Mammogram Examination

The number of views taken of each breast was noted (mediolateral oblique and craniocaudal views versus a single-view mammogram). The use of two-view mammography improves sensitivity and specificity and reduces the percentage of mammograms judged to be abnormal (10,11). The year in which the mammograms were performed (not the year of publication) was also abstracted because possible improvements in the quality of mammograms over time and secular trends toward an increasing false-positive rate have been noted in the United States (12).

The mammographic interpretive process in each publication was categorized as involving a single radiologist versus separate readings by two radiologists, because double reading of mammograms has been associated with improved accuracy (13,14). If the interpretive process was not stated explicitly, we relied on the use of the term “radiologist” (i.e., singular) in the published literature to imply single-reader interpretation. Although single-reader interpretation is the standard-of-care in most North American screening programs, in other countries, interpretation is often performed by two radiologists (15). The actual method of double reading of mammograms was not stated in most of the publications we studied, but it presumably included two radiologists interpreting each film independently (e.g., blinded), with disagreements decided by consensus or by two radiologists reading each film together.

Outcome of Mammographic Readings

Most published studies during this time period (i.e., between January 1, 1985, and June 15, 2002) did not describe screening mammography program results using standardized methods (e.g., Breast Imaging Reporting and Data System [BI-RADS] classification) (16). Instead, most studies used descriptive prose to define the percentage of mammograms judged to be abnormal and for the measures of accuracy (e.g., PPVs). Because cancer mortality was not an available outcome in these studies, intermediate outcomes of several types were used, as defined below.

The percentage of mammograms judged to be abnormal was defined as the percentage of screening mammograms that the radiologist believed required further diagnostic evaluation, which is also sometimes referred to as the callback or recall rate. For the purposes of this study, our definition included short-interval follow-up diagnostic mammogram within 12 months, additional immediate mammographic views, follow-up with a clinician for physical

examination correlation (i.e., follow-up on an abnormality), ultrasound, and fine-needle aspiration or core biopsy. This broad definition corresponds to the use of BI-RADS classifications 0, 3, 4, and 5 indicating the positivity of an initial screening mammogram (16). Although a short-interval follow-up is not considered to be a recall according to the American College of Radiology recommendations (16), we included it in our definition because of the variability of the published reports—that is, some reports included short-interval follow-up in their definition of the percentage of mammograms judged to be abnormal, whereas other reports did not or were not clear in their definition.

PPV_A was defined as the percentage of women with abnormal screening mammograms (i.e., those with a recall) that ultimately received a diagnosis of breast cancer. PPV_A is roughly equivalent to the American College of Radiology PPV_1 classification (16). However, as noted above, we also included short-interval follow-up (i.e., BI-RADS classification 3).

PPV_B was defined as the percentage of women undergoing biopsy as a result of an abnormal screening mammogram who ultimately received a diagnosis of breast cancer. PPV_B is roughly equivalent to the American College of Radiology PPV_3 classification. However, it is important to note that data on biopsy outcomes in the published articles included fine-needle aspiration, open and core biopsies, and combinations thereof. Because many articles did not specify the type of biopsy performed, we could not stratify or interpret by biopsy type.

Few published studies from community-based settings reported breast cancer mortality results because of the difficulty in tracking patients and the requirements for long-term follow-up. For these reasons, intermediate endpoints were used to assess the clinical outcomes of a screening program. Although tumornode–metastasis (TNM) staging was not available for most articles, the percentage of women with minimal disease at the time of diagnosis and the percentage of women with DCIS at the time of diagnosis were commonly noted. Minimal disease was defined as invasive breast cancer that was less than or equal to 10 mm in diameter and/or DCIS at the time of the diagnosis (17).

Statistical Analysis

To model the proportion of mammograms judged to be abnormal in relation to other study characteristics, we used meta-regression analysis, a form of regression analysis that extends random effects meta-analysis (18–20). Meta-regression analysis assumes that, after accounting for within-study variance in the outcome of interest, the remaining between-study variance can be divided into systematic and random components. The systematic component is modeled using regression analysis on study-level covariates. Restricted maximum likelihood was used to estimate the residual random component of the between-study variance. The within-study variance was specified as $p(1 - p)/n$, where p = the proportion of subjects experiencing the outcome of interest in a given study and n = the number of subjects on which p was based. All meta-regressions were conducted using the metareg procedure in Stata 7.0 (Stata Corp., College Station, TX). *P* values were obtained by dividing the appropriate regression coefficient by its estimated standard error and then treating the quotient as a standard normal deviate. All statistical tests were two-sided.

Results

We identified 227 candidate published articles from breast cancer screening programs, of which 32 (13,21–51) met the study inclusion criteria. A description of the patient populations screened and the mammography techniques used is shown in Table 1. Of the 32 screening programs identified, eight were located in North America (21–28) and 24 were located in other countries (13,29,52). All of the studies reviewed were based on screening

mammograms completed between 1985 and 2000 for the screening programs from North American and between 1987 and 1999 for the screening programs from other countries. The number of screening mammograms reported among studies ranged from 1813 (41) to 1 495 744 (34). North American programs had a larger percentage of women under age 50 years compared with programs from other countries. Most North American programs also used single radiologist interpreters and two views per breast, whereas programs located in other countries tended to use double readings and one view per breast. Only one North American screening program (22) provided data on initial screening examination results, whereas 10 programs (31,34,36,37,39–41,46,47,49) from other countries provided such data (data not shown), perhaps because many programs in other parts of the world were newly established in their geographic areas.

The percentage of mammograms judged to be abnormal ranged from 5.5% to 15.0% in the North American programs compared with the range of 1.2%–12.6% in programs from other countries (Fig. 1 and Table 2). All of the North American programs reported that the percentage of mammograms judged to be abnormal exceeded 5%, whereas only 10 of 24 programs from other countries reported percentages above this threshold (Table 2). Overall, the weighted mean percentage of mammograms judged to be abnormal was statistically significantly higher in North American screening programs than in programs from other countries (8.4% versus 5.6%, respectively; difference = 2.8%, 95% confidence interval = 0.5% to 5.1%; $P = .018$).

To determine whether population characteristics and features of the mammogram examination confounded the difference in the percentage of mammograms judged to be abnormal between North American screening programs and programs from other countries, the meta-regression models summarized in Fig. 2 were fitted and adjusted for population characteristics and features of the mammogram examination. The adjusted difference in the percentage of mammograms judged to be abnormal by program location ranged between 2.0% and 3.9% across all models. In other words, the percentage of mammograms judged to be abnormal in North American programs was between 2 and 4 percentage points higher than it was in programs from other countries, even after adjusting for important study covariates. Confidence intervals excluded 0 in all models, except in the model that adjusted for the percentage of women who had a mammogram who were less than 50 years old, for which the confidence intervals were relatively wide. Because the number of screening programs was relatively small, no attempt was made to control for more than one other covariate at a time. Among the program characteristics assessed (Fig. 2), program location was the most statistically significant predictor of the percentage of mammograms judged to be abnormal. Once program location was in the meta-regression model, no other program characteristics added statistically significantly to the model.

The PPV_A in North American screening programs ranged from a low of 4.4% in New Mexico (24) to a high of 12.2% in Idaho (23) (Table 2). In the programs from other countries, the PPV_A ranged from a low of 3.4% in Portugal (46) to a high of 48.7% in The Netherlands (45). Similarly, the PPV_B was lower in North American screening programs, ranging from 16.9% to 51.8%, than it was in programs from other countries, which ranged from 5.0% to 85.2%. Every North American screening program except that in North Carolina (25) had a PPV_B less than 40%. In contrast, all but four programs from other countries (35,36,38,46) for which PPV_B could be obtained had a PPV_B greater than 40%.

The percentage of cancer cases with DCIS at the time of diagnosis ranged from 4.3% to 68.1% (Table 2). Four of eight of the North American screening programs, but only four of the 24 programs from other countries, reported more than 20% of cases with DCIS. The percentage of cases with minimal disease at the time of diagnosis ranged from 30.6% to

80.6% in North American programs and from 14.0% to 55.7% in programs from other countries (Table 2).

To evaluate the strength of the association between the percentage of mammograms judged to be abnormal and the six screening outcomes, the scatter plots shown in Fig. 3 were constructed. The plotting symbol area for each study is proportional to the number of screening mammograms in the audit. The percentage of mammograms judged to be abnormal had a strong negative association with PPV_A and PPV_B, both of which were highly statistically significant ($P<.001$). In contrast, the proportion of mammograms judged to be abnormal had a strong positive association with the proportion of DCIS cases among women diagnosed with breast cancer ($P = .008$) and the number of DCIS cases diagnosed per 1000 screens ($P = .024$). However, no consistent association was observed between the percentage of mammograms judged to be abnormal and either the proportion of breast cancer diagnoses reported as having minimal disease ($P = .21$) or the number of breast cancer cases diagnosed per 1000 screens ($P = .48$).

Discussion

This study assessed international variability in screening mammography interpretations in community-based programs. The published data show wide ranges in the percentage of mammograms judged to be abnormal, with the percentage for North American programs being 2–4 percentage points higher than that for programs from other countries, even after adjusting for important covariates, including percentage of women less than 50 years of age who had a mammogram, calendar year in which mammogram was performed, number of readers for each examination (i.e., single versus double reading), and number of mammographic views per breast. The percentage of screening mammograms judged to be abnormal was negatively associated with the positive predictive value (PPV_A, PPV_B), without evident benefit in the yield of cancers detected per 1000 screens; however, an increase was noted in DCIS detection. This finding suggests that many, if not most, of the presumptively abnormal mammographic interpretations in screening programs with high percentages of abnormal mammograms turn out to be false positives.

The impact of wide variability on the percentage of mammograms judged to be abnormal involves trade-offs between benefits and risks of mammography screening. On one hand, a low percentage of mammograms judged to be abnormal (i.e., <2%) may reflect decreased sensitivity and missed breast cancers or delay in diagnosis, potentially resulting in more advanced disease. On the other hand, a high percentage of mammograms judged to be abnormal (i.e., >10%) might reflect an increased false-positive rate or ‘over-reading’ of mammographic films. Indeed, the percentage of mammograms judged to be abnormal can be shown algebraically (see below) to be a weighted average of the true-positive rate (i.e., sensitivity) and the false-positive rate (1 – specificity)—that is, the percentage of mammograms judged to be abnormal = sensitivity × prevalence + (1 – specificity) × (1 – prevalence). Because the vast majority of women screened do not have breast cancer, (1 – prevalance) will normally be much greater than the prevalence. Hence, the percentage of mammograms judged to be abnormal will depend more on the false-positive rate than on the true-positive rate.

False-positive mammograms can lead to unnecessary diagnostic evaluations and high medical costs, anxiety for women receiving follow-up evaluations and even, in rare cases, morbidity (e.g., infections from biopsies, scars) (12,53). False-positive results do not have a negative impact on subsequent breast cancer screening behavior (54,55). It is estimated that 50% of women will have at least one false-positive mammogram after 10 screening examinations (12). In the United States, this false-positive rate translates into a substantial

annual financial burden (12). For example, an additional $750 000 000 might be spent on diagnostic evaluations each year in the United States if the percentage of mammograms judged to be abnormal were 10% instead of 5% [an estimated 30 million women are screened annually at an average cost of $500 per false-positive episode (56)]. Conversely, it is possible that diagnosing women with earlier stage breast cancer can obtain financial savings (57). It is important, therefore, to learn how clinical benefit, as well as financial cost and possible harm, is associated with variability in mammography interpretation.

In this descriptive study, marked heterogeneity between the screening studies was noted. This finding is to be expected in view of differences in characteristics of the population screened, features of the mammogram examination, and definitions used (e.g., what constitutes an abnormal mammogram). Although we attempted to deal with this heterogeneity in the selection of the studies and we applied statistical tests designed for this type of meta-analysis, interpretation of a possible difference in mammography interpretation between North American screening programs and those from other countries must be made with caution. In addition, only aggregate-level measures of demographic and mammography characteristics were available, rather than data for individual women. Hence, the circumstances of data collection (e.g., relying on what is available in the published literature) weaken any statistical conclusions.

Possible Explanations for the Variability Noted

The question that is implicitly raised by our results is, Why are North American radiologists calling for further evaluations of so many more women, most of whom turn out to have false-positive mammograms, than radiologists from other countries? Given the extent of variability in mammographic interpretation noted in the studies we examined and its potential clinical importance, it is worthwhile discussing the possible sources of this variability and the implications that these findings might have on future research and health policy. The variation in mammographic interpretation noted in the published articles that we studied is likely due to multiple factors as reviewed below and summarized in Table 3.

Characteristics of the population screened—Screening programs differed substantially in the basic features of the population screened, such as the number of women less than 50 years old who had a mammogram. Women in this younger age group have denser breast tissue and a lower incidence of breast cancer, so the sensitivity of mammographic screening may be lower, although this conclusion is controversial. Screening programs with a high percentage of younger women might therefore be expected to have a higher percentage of mammograms judged to be abnormal and to have more false positives than programs that limit screening examinations to older women. Because breast density and breast cancer incidence change gradually by decade, the binary age classification we used (i.e., <50 versus ≥ 50 years) is imperfect; however, data were not always reported by decade categories.

Differences in the number of women undergoing their first or initial screening mammogram within each screening program could also explain some of the variability in the interpretation of mammograms. The percentage of mammograms judged to be abnormal generally decreases with subsequent screening examinations because comparison films establish stability of any abnormalities and because the number of incident cancers in a screened population should be lower than in a non-screened population (9). It is interesting that many of the screening programs that had the lowest percentage of mammograms judged to be abnormal and the highest positive predictive value were pilot projects in European countries where screening mammograms had previously been limited, so a high rate of abnormalities in those locations might be expected. In addition, if cancers were being

routinely missed in these programs because of the low percentage of mammograms judged to be abnormal, one might also expect more late-stage cancers to become apparent; however, no association was observed between the percentage of mammograms judged to be abnormal and the proportion of cancers reported as having minimal disease.

It is possible that the populations screened differ in regard to the presence of breast cancer risk factors, ethnic groups, presence of known breast cancer symptoms at the time of the mammogram, and the referral status of the woman; however, data for these variables were not available.

Features of the mammography examination—The screening studies we examined covered approximately the same time periods, so mammographic equipment and other technical issues were probably similar. However, all of the North American screening programs used two mammographic views per breast, whereas one or two mammographic views were used for programs from other countries. In addition, double reading of films was more common in screening programs from other countries than it was in programs from North America. Unfortunately, the published articles did not generally describe the specific details regarding how double reading was performed (i.e., two independent blind interpretations versus consensus interpretation). The differences noted between North American programs and programs from other countries persisted after adjustment for calendar year, number of views per breast, and use of single versus double readings.

Features of physicians interpreting the mammogram—Previous research has suggested a positive correlation between a physician's experience and accuracy of mammography interpretations (8,58–60), although definitions of experience vary and not all research has shown a statistically significant association (61,62). In addition, whether a physician's personal comfort with ambiguity influences clinical decision-making is an important but understudied area of research (63). It is possible that physicians who are uncomfortable with ambiguity interpret a higher percentage of mammograms as abnormal. Individual physicians have different thresholds for labeling mammograms as abnormal (59); decision-making aids may help in adjusting the thresholds of selected physicians (64–67).

Features of the health care system—Malpractice concerns may also be contributing to the high percentage of mammograms judged to be abnormal in some screening programs. For example, physicians in North America might be reluctant to label a mammogram as normal, thus leading to higher percentages of mammograms being judged as abnormal. This higher percentage of mammograms judged to be abnormal would probably lead to more false-positive mammograms and perhaps more true-positive mammograms with better sensitivity. It is interesting that malpractice concerns have been shown to affect physicians' ordering of tests and procedures (68–75). Failure or delay in cancer diagnosis is the most frequent allegation in medical malpractice claims in the United States (76), with issues related to breast cancer, particularly delay in diagnosis, being the most common cause of malpractice claims for all specialties (77). Malpractice concerns are on physicians' minds, and some physicians believe that these concerns influence their clinical practice in a variety of ways, including referring more patients to other physicians and increasing the use of tests and procedures (68–70,73,77–80). In academic and public screening programs, there is a possible buffer against malpractice concerns in that physicians are rarely responsible for their own malpractice insurance or defense in litigation.

Financial incentives inherent to the health care system can also affect physicians' use of health care resources (81–83). Physicians in nonprofit mammography programs might differ in their practice patterns compared with physicians who interpret both screening and diagnostic mammograms with an incentive plan for high volume. In addition, academic and

public screening programs might be isolated from the confounding issue of benefiting financially from additional diagnostic procedures that may follow an abnormal screening mammogram.

Some screening programs use specialized radiologists and have stated goals to develop high-quality screening programs that minimize false-positive results (11,22,41). These specialized programs monitor outcomes through quality control and self-auditing procedures. Programs that combine clinical breast examination with screening mammography may have higher biopsy rates because of abnormalities noted on the clinical examination and not necessarily on the mammography examination. In contrast, smaller, community-based screening programs may not have access to specialized radiologists.

Finally, national policy regarding desired goals for the percentage of screening mammograms judged to be abnormal varies (84,85) (Table 4). It is possible that physicians interpreting mammograms may, to some degree, simply be responding to those goals. Screening programs may also be using different methods for defining the percentage of mammograms judged to be abnormal and for calculating cancer outcomes. For example, some screening programs categorize short-term follow-up as a positive mammogram, which will elevate the false-positive rate. In addition, some European countries used fine-needle biopsies earlier than they were routinely used in the United States. The screening programs from these European countries do not consider fine-needle aspiration a biopsy; however, they use the aspiration information to decide if an excisional biopsy is warranted, which could well elevate the reported PPV_B for these programs.

In conclusion, substantial variation exists among published reports of community-based screening mammography programs. The percentage of mammograms judged to be abnormal in North American programs was 2–4 percentage points higher than it was in programs from other countries without evident benefit in the yield of cancers detected per 1000 women screened, although an increase was noted in DCIS detection. Across all screening programs, the percentage of mammograms judged to be abnormal was inversely associated with positive predictive value (both PPV_A and PPV_B). Although some variability might be explained by differences in patient populations and technique of mammography examinations, variability may also be due to differences in physicians' interpretations and in features of health care systems.

The mammography screening programs described in this article are heterogeneous both in the methods of auditing their practice and in factors that may affect the interpretation of results. Currently, it is not possible to determine which features may be most responsible for the variability. To address this problem, we recommend standardized reporting of mammography data in the literature, with results stratified for patient characteristics and features of the examination and data on both process and outcome. We also endorse the new international standards for reporting diagnostic tests evaluation (86,87), which will improve the quality of reporting methods and results. In addition, linkage of screening programs to tumor registries, as is currently being performed in the United States by the Breast Cancer Surveillance Consortium (88), is critical to obtaining outcome data that go beyond intermediary endpoints. Outcome data on breast cancer–specific mortality would be the best comparison but were not available in the published literature. Finally, a better understanding of the sources of variability in mammography may lead to more effective screening programs that have a lower percentage of mammograms judged as abnormal without substantially lowering the cancer detection rate.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Supported by Public Health Service grant HS-10591 (to J. G. Elmore) from the Agency for Healthcare Research and Quality and by the National Cancer Institute, National Institutes of Health, Department of Health and Human Services.

We appreciate suggestions on early drafts of this work by Drs. Suzanne Fletcher and Patricia A. Carney.

References

1. Elmore J, Wells C, Lee C, Howard D, Feinstein A. Variability in radiologists' interpretations of mammograms. N Engl J Med. 1994; 331:1493–1499. [PubMed: 7969300]
2. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Arch Intern Med. 1996; 156:209–213. [PubMed: 8546556]
3. 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; 90:1801–1809. [PubMed: 9839520]
4. Elmore J, Feinstein A. A bibliography of publications on observer variability (final installment). J Clin Epidemiol. 1992; 45:567–580. [PubMed: 1607896]
5. Feinstein A. A bibliography of publications on observer variability. J Chronic Dis. 1985; 38:619–632. [PubMed: 3894405]
6. 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; 165:1373–1377. [PubMed: 7484568]
7. Meyer JE, Eberlein TJ, Stomper PC, Sonnenfeld MR. Biopsy of occult breast lesions. JAMA. 1990; 263:2341–2343. [PubMed: 2157903]
8. Elmore JG, Miglioretti DL, Reisch LM, Barton MB, Kreuter W, Christiansen CL, et al. Screening mammograms by community radiologists: variability in false-positive rates. J Natl Cancer Inst. 2002; 94:1373–1380. [PubMed: 12237283]
9. 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]
10. Bryan S, Brown J, Warren R. Mammography screening: an incremental cost effectiveness analysis of two view versus one view procedures in London. J Epidemiol Community Health. 1995; 49:70–78. [PubMed: 7707010]
11. Warren R, Duffy S, Bashir S. The value of the second view in screening mammography. Br J Radiol. 1996; 69:105–108. [PubMed: 8785635]
12. 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; 338:1089–1096. [PubMed: 9545356]
13. Warren R, Duffy S. Comparison of single reading with double reading of mammograms, and change in effectiveness with experience. Br J Radiol. 1995; 68:958–962. [PubMed: 7496693]
14. Brown J, Bryan S, Warren R. Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. BMJ. 1996; 312:809–812. [PubMed: 8608287]
15. de Waard F, Kirkpatrick A, Perry N, Tornberg S, Tubiana M, de Wolf C. Breast cancer screening in the framework of the Europe against Cancer programme. Eur J Cancer Prev. 1994; 3:3–5. [PubMed: 8130724]
16. Illustrated breast imaging reporting and data system. 3 rd ed.. Reston (VA): American College of Radiology; American College of Radiology.
17. Linver M, Osuch J, Brenner R, Smith R. The mammography audit: a primer for the Mammography Quality Standards Act (MQSA). AJR Am J Roentgenol. 1995; 165:19–25. [PubMed: 7785586]

18. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986; 7:177–188. [PubMed: 3802833]
19. Mosteller F, Colditz G. Understanding research synthesis (meta-analysis). Annu Rev Public Health. 1996; 17:1–23. [PubMed: 8724213]
20. Thompson S, Sharp S. Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med. 1999; 18:2693–2708. [PubMed: 10521860]
21. Burhenne LJ, Hislop TG, Burhenne HJ. The British Columbia mammography screening program: evaluation of the first 15 months. AJR Am J Roentgenol. 1992; 158:45–49. [PubMed: 1307850]
22. Kerlikowske K, Grady D, Barclay J, Sickles E, Eaton A, Ernster V. Positive predictive value of screening mammography by age and family history of breast cancer. JAMA. 1993; 270:2444–2450. [PubMed: 8230621]
23. Robertson CL. A private breast imaging practice: medical audit of 25,788 screening and 1,077 diagnostic examinations. Radiology. 1993; 187:75–79. [PubMed: 8451440]
24. Rosenberg RD, Lando JF, Hunt WC, Darling RR, Williamson MR, Linver MN, et al. The New Mexico mammography project. Cancer. 1996; 78:1731–1739. [PubMed: 8859186]
25. Bird RE. Low-cost screening mammography: report on finances and review of 21,716 consecutive cases. Radiology. 1989; 171:87–90. [PubMed: 2494683]
26. Freer T, Ulissey M. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. Radiology. 2001; 220:781–786. [PubMed: 11526282]
27. Peters GN, Vogel VG, Evans WP, Bondy M, Halabi S, Lord J, et al. The Texas Breast Screening Project: Part I. Mammographic and clinical results. South Med J. 1993; 86:385–390. [PubMed: 8465213]
28. Lynde J. Low-cost screening mammography: results of 21,141 consecutive examinations in a community program. South Med J. 1993; 86:338–343. [PubMed: 8451676]
29. Robinson J, Crane C, King J, Scarce D, Hoffman C. The South Australian breast x-ray service: results from a statewide mammographic screening programme. Br J Cancer. 1996; 73:837–842. [PubMed: 8611391]
30. A mammographic screening pilot project in Victoria 1988–1990. The Essendon Breast X-ray Program Collaborative Group. Med J Aust. 1992; 157:670–673. [PubMed: 1435408]
31. Rickard M, Lee W, Read J, Scott A, Stephen D, Grace J. Breast cancer diagnosis by screening mammography: early results of the Central Sydney Area Health Service Breast X-ray Programme. Med J Aust. 1991; 154:126–131. [PubMed: 1986190]
32. Bleyen L, Landeghem PV, Pelfrene E, Vriendt MD, DeSmet A, Backer GD. Screening for breast cancer in Ghent, Belgium: first results of a programme involving the existing health services. Eur J Cancer. 1998; 34:1410–1414. [PubMed: 9849425]
33. Van Oyen H, Verellen W. Breast cancer screening in the Flemish Region, Belgium. Eur J Cancer Prev. 1994; 3:7–12. [PubMed: 8130728]
34. Dean P, Pamilo M. Screening mammography in Finland--1.5 million examinations with 97 percent specificity. Acta Oncol. 1999; 38 Suppl 13:47–54. [PubMed: 10612496]
35. Renaud R, Gairard B, Schaffer P, Guldenfels C, Haehnel P, Dale G, et al. Europe against Cancer breast cancer screening programme in France: the ADEMAS programme in Bas-Rhin. Eur J Cancer Prev. 1994; 3:13–19. [PubMed: 8130721]
36. Seguret F, Daures J, Guizard A, Mathieu-Daude H, Bonifacj J, Cherifcheik J, et al. Herault breast screening programme: results after 30 months of mobile French schedule. Eur J Cancer Prev. 1995; 4:299–305. [PubMed: 7549822]
37. Garas I, Pateras H, Triandafilou D, Georgountzos V, Mihas A, Abatzoglou M, et al. Breast cancer screening in southern Greece. Eur J Cancer Prev. 1994; 3:35–39. [PubMed: 8130725]
38. Lau Y, Lau P, Chan C, Yip A. The potential impact of breast cancer screening in Hong Kong. Aust N Z J Surg. 1998; 68:707–711. [PubMed: 9768606]
39. Sigfusson B, Hallgrimsson P. Breast cancer screening in Iceland: preliminary results. Recent Results Cancer Res. 1990; 119:94–99. [PubMed: 2236868]

40. Codd M, Laird O, Dowling M, Dervan P, Gorey T, Stack J, et al. Screening for breast cancer in Ireland: the Eccles breast screening programme. Eur J Cancer Prev. 1994; 3:21–27. [PubMed: 8130722]
41. de Placido S, Nuzzo F, Perrone F, Carlomagno C, Noviello A, Delrio P, et al. The first breast cancer screening program in southern Italy: preliminary results from three municipalities of the Naples Province. Tumori. 1995; 81:7–12. [PubMed: 7754547]
42. Giorgi D, Ambrogetti D, Bianchi S, Catarzi S, Ciatto S, Morrone D, et al. Design and preliminary results of the Florence Breast Cancer Screening Programme (Progetto Firenze Donna). Eur J Cancer Prev. 1994; 3:29–34. [PubMed: 8130723]
43. Filippini L, Braga M, Perna E, Bianchi A, Bettoni C, Lucini L, et al. Results of a mammographic and clinical screening in a health district (USSL) of Brescia, Italy. Tumori. 1996; 82:430–436. [PubMed: 9063517]
44. Autier P. A breast cancer screening programme operating in a liberal health care system: the Luxembourg Mammography Programme, 1992–1997. Int J Cancer. 2002; 97:828–832. [PubMed: 11857363]
45. De Koning HJ, Fracheboud J, Boer R, Verbeek AL, Hubertine JC, Hendriks JH, et al. Nation-wide breast cancer screening in The Netherlands: support for breast-cancer mortality reduction. Int J Cancer. 1995; 60:777–780. [PubMed: 7896444]
46. Alves JG, Cruz DB, Rodrigues VL, Goncalves ML, Fernandes E. Breast cancer screening in the central region of Portugal. Eur J Cancer Prev. 1994; 3 Suppl 1:49–53. [PubMed: 8130727]
47. Ascunce N, del Moral A, Murillo A, Alfaro C, Apesteguia L, Ros J, et al. Early detection programme for breast cancer in Navarra, Spain. Eur J Cancer Prev. 1994; 3:41–48. [PubMed: 8130726]
48. Vizcaino I, Salas D, Vilar J, Ruiz-Perales F, Herranz C, Ibanez J. Breast cancer screening: first round in the population-based program in Valencia, Spain. Collaborative group of readers of the Breast Cancer Screening Program of the Valencia Community. Radiology. 1998; 206:253–260. [PubMed: 9423680]
49. Thurfjell E. Population-based mammography screening in clinical practice. Acta Radiol. 1994; 35:487–491. [PubMed: 8086260]
50. Litherland J, Evans A, Wilson A. The effect of hormone replacement therapy on recall rate in the National Health Service Breast Screening Programme. Clin Radiol. 1997; 52:276–279. [PubMed: 9112944]
51. O'Driscoll D, Britton P, Bobrow L, Wishart G, Sinnatamby R, Warren R. Lobular carcinoma in situ on core biopsy-what is the clinical significance? Clin Radiol. 2001; 56:216–220. [PubMed: 11247699]
52. Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. Radiology. 1994; 191:241–244. [PubMed: 8134580]
53. Lerman C, Trock B, Rimer B, Boyce A. Psychological and behavioral implications of abnormal mammograms. Ann Intern Med. 1991; 114:657–661. [PubMed: 2003712]
54. Burman ML, Taplin SH, Herta DF, Elmore JG. Effect of false-positive mammograms on interval breast cancer screening in an HMO. Ann Intern Med. 1999; 131:1–6. [PubMed: 10391809]
55. Pisano E, Earp J, Schell M, Vokaty K, Denham A. Screening behavior of women after a false-positive mammogram. Radiology. 1998; 208:245–249. [PubMed: 9646820]
56. Cyrlak D. Induced costs of low-cost screening mammography. Radiology. 1988; 168:661–663. [PubMed: 3406395]
57. Taplin S, Barlow W, Urban N, Mandelson M, Timlin D, Ichikawa L, et al. Stage, age, comorbidity, and direct costs of colon, prostate, and breast cancer care. J Natl Cancer Inst. 1995; 87:417–426. [PubMed: 7861461]
58. Linver M, Paster S, Rosenberg R, Key C, Stidley C, King W. Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,663 cases. Radiology. 1992; 184:39–43. [PubMed: 1609100]
59. Elmore JG, Wells CK, Howard DH. Does diagnostic accuracy in mammography depend on radiologists' experience? J Womens Health. 1998; 7:443–449. [PubMed: 9611702]

60. Esserman L, Cowley H, Eberle C, Kirkpatrick A, Chang S, Berbaum K, et al. Improving the accuracy of mammography: volume and outcome relationships. J Natl Cancer Inst. 2002; 94:321–323. [PubMed: 11880465]
61. Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. J Natl Cancer Inst. 2003; 95:282–290. [PubMed: 12591984]
62. Elmore JG, Miglioretti DL, Carney PA. Does practice make perfect when interpreting mammography? Part II. J Natl Cancer Inst. 2003; 95:250–252. [PubMed: 12591973]
63. Pearson S, Goldman L, Orav E. Triage decisions for emergency department patients with chest pain: do physicians' risk attitudes make a difference? J Gen Intern Med. 1995; 10:557–567. [PubMed: 8576772]
64. D'Orsi C, Getty D, Swets J, Pickett R, Seltzer S, McNeil B. Reading and decision aids for improved accuracy and standardization of mammographic diagnosis. Radiology. 1992; 184:619–622. [PubMed: 1509042]
65. Swets J, Getty D, Pickett R, D'Orsi C, Seltzer S, McNeil B. Enhancing and evaluating diagnostic accuracy. Med Decis Making. 1991; 11:9–18. [PubMed: 2034078]
66. Getty D, Pickett R, D'Orsi C, Swets J. Enhanced interpretation of diagnostic images. Invest Radiol. 1988; 23:240–252. [PubMed: 3372189]
67. D'Orsi CJ, Swets JA. Variability in the interpretation of mammograms. N Engl J Med. 1995; 332(17):1172.
68. Bovbjerg RR, Dubay LC, Kenney GM, Norton SA. Defensive medicine and tort reform: new evidence in an old bottle. J Health Polit Policy Law. 1996; 21:267–288. [PubMed: 8723178]
69. Opinion Research Corporation. Prepared for the American College of Obstetricians and Gynecologists. Washington (DC): Opinion Research Corporation; 1988 March. Professional liability and its effects: report of a 1987 survey of ACOG's membership.
70. Jacobson PD, Rosenquist CJ. The use of low-osmolar contrast agents: technological change and defensive medicine. J Health Polit Policy Law. 1996; 21:243–266. [PubMed: 8723177]
71. California Medical Association. Professional liability issues in obstetrical practice. 1987 Socioecon Rep 25; Nos. 6 and 7.
72. Klingman D, Localio AR, Sugarman J, Wagner JL, Polishuk PT, Wolfe L, et al. Measuring defensive medicine using clinical scenario surveys. J Health Polit Policy Law. 1996; 21:185–217. [PubMed: 8723175]
73. Hershey N. The defensive practice of medicine: myth or reality. Milbank Mem Fund Q. 1972 Jan. 50:69–98. [PubMed: 5060142]
74. Hirsh HL, Dickey TS. Defensive medicine as a basis for malpractice liability. Trans Stud Coll Physicians Phila. 1983; 5:99–107. [PubMed: 6879665]
75. Edwards KS. Defensive medicine Health care with a pricetag. Ohio State Med J. 1985; 81:38–42. [PubMed: 3969252]
76. Strunk AL, Kenyon S. Medicolegal considerations in the diagnosis of breast cancer. Obstet Gynecol Clin North Am. 2002; 29:43–49. [PubMed: 11892873]
77. Physician Insurers Association of America. Breast cancer study: June 1995. Washington, DC: Physician Insurers Association of America; 1995 June.
78. Zuckerman S. Medical malpractice: claims, legal costs, and the practice of defensive medicine. Health Aff (Millwood) Fall. 1984; 3:128–133.
79. Weisman CS, Morlock LL, Teitelbaum MA, Klassen AC, Celentano DD. Practice changes in response to the malpractice litigation climate. Results of a Maryland physician survey. Med Care. 1989; 27:16–24. [PubMed: 2911218]
80. Voss JD. Prostate cancer, screening, and prostate-specific antigen: promise or peril? J Gen Intern Med. 1994; 9:468–474. [PubMed: 7525903]
81. Hillman AL, Pauly MV, Kerstein JJ. How do financial incentives affect physicians' clinical decisions and the financial performance of health maintenance organizations? N Engl J Med. 1989; 321:86–92. [PubMed: 2733758]

82. Hemenway D, Killen A, Cashman SB, Parks CL, Bicknell WJ. Physicians' responses to financial incentives. Evidence from a for-profit ambulatory care center. N Engl J Med. 1990; 322:1059–1063. [PubMed: 2320066]

83. Fairbrother G, Hanson KL, Friedman S, Butts GC. The impact of physician bonuses, enhanced fees, and feedback on childhood immunization coverage rates. Am J Public Health. 1999; 89:171–175. [PubMed: 9949744]

84. de Wolf, CJ.; Perry, NM. European guidelines for quality assurance in mammography screening. 2nd ed.. Luxembourg: European Commission, Europe Against Cancer Programme; 1996. I-5, I-7, I-10

85. Quality determinants of mammography. Clinical practice guidelines. No. 13. Agency for Health Care Policy and Research Publ. 1994 October. 95–0632.

86. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem. 2003; 49:7–18. [PubMed: 12507954]

87. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. Clin Chem. 2003; 49:1–6. [PubMed: 12507953]

88. Ballard-Barbash R, Taplin S, Yankaskas B, Ernster V, Rosenberg R, Carney P, et al. Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database. AJR Am J Roentgenol Oct. 1997; 169:1001–1008.

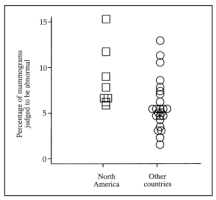

Fig. 1.

Percentage of mammograms judged to be abnormal from published studies by screening program location. North American screening programs (n = 8) are shown as **open squares**, and programs from other countries (n = 24) are shown as **open circles**. The weighted mean percentage of mammograms judged to be abnormal was statistically significantly higher in North American programs than it was in programs from other countries (8.4% versus 5.6%; difference in weighted mean percentage = 2.8%, 95% confidence interval = 0.5% to 5.1%; *P* = .018).

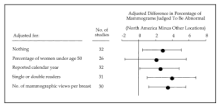

Fig. 2.

Differences in the percentage of mammograms judged to be abnormal between North American screening programs and programs from other countries, after adjustment for patient population and program characteristics. Differences in the percentages were calculated as the percentage in North American programs minus the percentage in programs from other countries. The difference in the percentage of mammograms judged to be abnormal between North American screening programs and programs from other counties remained between 2.0% and 3.9%, even after adjustment for the percentage of women less than 50 years old, reported calendar year in which the mammogram was performed, single or double mammography readers, and number of mammography views per breast.

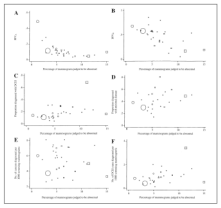

Fig. 3.

Study-level (i.e., at the individual publication level) association of the percentage of mammograms judged to be abnormal with the positive predictive value of an abnormal mammogram (PPV_A) (**A**), the positive predictive value of a biopsy performed (PPV_B) (**B**), the proportion of cancer cases diagnosed with ductal carcinoma *in situ* (DCIS) (**C**), the proportion of cancer cases diagnosed with minimal disease (DCIS and/or tumor size ≤ 10 mm) (**D**), the number of cancer cases diagnosed per 1000 screening mammograms (**E**), and the number of DCIS cases diagnosed per 1000 screening mammograms (**F**). North American screening studies are shown as **open squares**, and screening studies from other countries are shown as **open circles**. Symbol area for each study is proportional to the number of screening mammograms performed.

Table 1

Features of the population screened and the mammography examination*

Study location (Ref.)	Population screened		Features of mammography examination		
	% of women less than 50 years of age	No. of screening mammograms†	No. of views	Year of mammogram‡	Single vs. double interpretation§
		North America			
British Columbia, Canada (21)	N/A	11824	2	1988–1989	S
California (22)	61.7	29838	2	1985–1992	S
Idaho (23)	N/A	25788	2	1989–1990	S
New Mexico (24)	43.7	126 466	2	1991–1993	S
North Carolina (25)‖	48	21716	2‖	1987	D
Texas (26)	N/A	12860	2	1999–2000	S
Texas (27)‖	N/A‖	64 459	1–2	1987	S‖
Virginia (28)‖	N/A‖	21 141	2	1988–1991	S
		Other countries			
Australia (30)	0	N/A	2	1988–1990	D
Australia (31)	19.5	7163	2	1988–1989	D
Australia (29)	23.3	76106	2	1989–1993	D
Belgium (32)	35.2	9215	2	1992–1994	D
Belgium (33)	0‖	16017	1	1989–1992	D
Finland (34)	N/A	1495 744	2	1987–1997	D
France (35)	0‖	40 535‖	1‖	1989–1992	D
France (36)	38.9	26026	2	1990–1992	D
Greece (37)‖	37	22258	2	1989–1990	D
Hong Kong (38)	74	8504	N/A	1993–1995	D
Ireland (40)‖	100	17903	2	1989–N/A	S‖
Iceland (39)‖	46	5416	2	1987–1989	D
Italy (41)‖	0	1813	2	1990–1992	S‖
Italy (43)	0	N/A	2	1990–1993	S
Italy (42)	0‖	20149	2	1990–1992	S

Study location (Ref.)	Population screened		Features of mammography examination		
	% of women less than 50 years of age	No. of screening mammograms†	No. of views	Year of mammogram‡	Single vs. double interpretation§
Luxembourg (44)	0	16249	2	1994–1997	D
Spain (47)	21.5	48691	1–2	1990–1992	S/D
Sweden (49,52)	40.3	41761	1–2	1988–1990	S/D
The Netherlands (45)//	0	416 020	2	1990–1992	D
Portugal (46)//	23.3	27416	1	1990–1992//	D
Spain (48)	27.6	78224	2	1992–N/A	D
United Kingdom (50)//	0//	5699	1–2	1995	S//
United Kingdom (51)	0	47975	N/A	1994–1999	N/A
United Kingdom (13)//	6.6//	26193	2	1987–1991	S/D

*Please see supplemental data online (available at http://jncicancerspectrum.oupjournals.org/jnci/content/vol95/issue18/) for additional information on the abstraction process for specific articles. N/A = not available.

†The total number of screening mammograms is equal to the total number of women in all studies but three: New Mexico (n = 87 443), Virginia (n = 17 547), and Spain (n = 49 631).

‡Reported year in which the mammograms were obtained.

§S = Single interpretation for each mammogram; D = double interpretation for each mammogram; S/D = a combination of single and double interpretation.

//Some data were obtained via personal communication with author.

J Natl Cancer Inst. Author manuscript; available in PMC 2011 July 29.

Table 2

Outcomes of screening mammography programs*

Study location (Ref.)	% of mammograms judged to be abnormal	PPV_A	PPV_B	% of cases with DCIS	% of cases with minimal cancer†
North America					
British Columbia, Canada (21)	8.7	4.6	22.3	27.7	59.6
California (22)	5.5	8.0	27.5	N/A	63.3
Idaho (23)	6.0	12.2	N/A	12.2	N/A
New Mexico (24)	11.4	4.4	16.9	68.1	49.1
North Carolina (25)	7.3	9.8	51.8	N/A	30.6
Texas (26)	6.5	4.9	38.3	29.3	N/A
Texas (27)	15.0	9.9	19.1	16.0	N/A
Virginia (28)	6.4	7.2	36.0	29.6	80.6
Other countries					
Australia (30)	7.5	7.5	55.9	18.3	55.7
Australia (31)	12.6	5.8	53.5	22.6	50.9
Australia (29)	4.9	14.2	57.8	17.6	52.3
Belgium (32)	2.9	28.5	58.1	12.0	42.7
Belgium (33)	4.1	7.1	51.7	30.4	23.9
Finland (34)	3.3	11.4	57.7	11.0	29.7
France (35)	10.4	5.0	5.0	18.9	41.0
France (36)	7.0	8.4	34.9	19.7	50.4
Greece (37)	5.3	5.9	43.7	4.3	14.0
Hong Kong (38)	11.0	4.5	37.2	23.8	N/A
Iceland (39‖	4.3	15.3	70.6	5.6	N/A
Ireland (40)	4.2	17.1	50.0	11.6	30.2
Italy (41)	5.0	12.1	57.9	9.1	27.3
Italy (43)	2.1	24.6	60.1	11.2	53.6
Italy (42)	4.3	17.4	85.2	9.3	51.3
Luxembourg (44)	5.2	11.2	56.0	10.6	40.4
The Netherlands (45)	1.2	48.7	67.7	14.0	38.0
Portugal (46)	8.3	3.4	12.3	16.7	38.5

Study location (Ref.)	% of mammograms judged to be abnormal	PPV_A	PPV_B	% of cases with DCIS	% of cases with minimal cancer[†]
Spain (47)	6.0	10.5	55.8	16.6	42.8
Spain (48)	5.0	8.6	56.9	12.6	35.1
Sweden (49,52)	4.8	12.0	57.0	11.2	19.9
United Kingdom (50)	2.9	9.0	N/A	13.3	N/A
United Kingdom (51)	4.9	13.3	41	30.9	N/A
United Kingdom (13)	6.6	12.3	79.5	17.7	31.8

[*] PPV_A = positive predictive value of an abnormal screen; PPV_B = positive predictive value of a biopsy done; DCIS = ductal carcinoma *in situ*; N/A = not available.

[†] Minimal cancer is defined as DCIS and/or invasive cancer with a tumor size of less than or equal to 10 mm.

Table 3

Possible explanations for the variability noted among published studies of screening mammography

Characteristics of the population screened
• Age (e.g., percentage of women <50 years of age)
• Initial versus subsequent screening examination
• Presence of risk factors for breast cancer
• Presence of breast symptoms
• Self-referral versus physician referral
Features of the mammography examination
• Equipment type and year
• One or two views of each breast
• Single or double readings
• Technician training
Features of physicians interpreting the mammogram
• Experience of the physician
• Level of personal comfort with ambiguity
• Individual thresholds to label film as abnormal
Features of the health care system
• Malpractice concerns
• Financial incentives
• Private versus academic/public programs
• Different stated goals for the percentage of mammograms judged abnormal and positive predictive value
• Quality control and auditing procedures
• Variability of definitions used to calculate outcomes

Table 4

Stated goals for the percentage of screening mammograms judged to be abnormal and the positive predictive value of a biopsy done (PPV_B)

Organization (Ref.)	Stated goals	% of mammograms judged to be abnormal	% PPV_B
Agency for Health Care Policy and Research (United States) (85)	General	≤10	25–40
Europe Against Cancer Programme* (84)	Acceptable levels	<7	>34
Europe Against Cancer Programme* (84)	Desirable levels	<5	>50

* For women aged 50–64 years at their initial screen.