Does Practice Make Perfect When Interpreting Mammography?

Joann G. Elmore, Patricia A. Carney

There are important trade-offs in the practice of interpreting mammography. Radiologists do not want to miss identifying breast cancer, yet performing additional imaging to rule out cancer increases false-positive rates. False-positive mammograms generate anxiety, excess costs and, at times, morbidity from subsequent biopsies. The false-positive rate for screening mammography is higher in the United States than in European countries. Reducing the false-positive rates and maintaining high levels of sensitivity, as suggested by Esserman et al. (1) in this issue of the Journal, is appealing. They hypothesize that interpreting a high volume of mammograms, as is the norm for radiologists in the U.K., results in higher sensitivity than interpreting a low volume, as is often the norm for some U.S. radiologists. In their study (1), a standardized test set of 60 screening films from asymptomatic women, 13 of which included nonoccult breast cancers, was interpreted by 60 U.S. radiologists and also by 194 U.K. radiologists. The U.S. radiologists were assigned to one of three categories on the basis of their self-reported monthly volume of mammography interpretations: low (≤100/month), moderate (101–300/month), or high volume (≥301/month). Their results suggest that both sensitivity and specificity are better among higher volume readers, although the differences for specificity are not statistically significant (1).

The study by Esserman et al. (1) underscores an important difference in the approach to screening mammography between the United States and the U.K. and raises questions relevant to both research and clinical practice. U.S. mammographers are currently required to interpret a minimum of 480 mammograms per year, whereas U.K. mammographers are required to interpret more than 5000 per year. What can we learn from comparisons of international practice styles? Should U.S. mammographers be required to interpret higher volumes of mammograms each year? Esserman et al. (1) have made an excellent initial step in addressing those questions. The findings deserve follow-up.

Previous research has shown that radiologists’ accuracy improves after special training courses (2). Experienced radiologists have higher sensitivity than do less experienced radiologists in diagnosing breast cancer; however, experienced radiologists may also order more tests that do not result in a cancer diagnosis (3). Although Esserman et al. (1) have found an association between volume and accuracy, it remains to be seen if practice really does make perfect.

In most areas of radiology, images are interpreted in conjunction with additional clinical and diagnostic information, so that interpretive judgment is not the sole basis for patient management. However, the variability in radiologists’ interpretations is of particular importance in screening mammography because the premise is to detect signs of breast cancer before there is any evidence of disease (e.g., a palpable mass). A decision to perform additional diagnostic imaging or to perform a breast biopsy...
may depend heavily on the interpretation of just the two-view screening mammograms. Thus, interpretive variability can directly affect patient management.

Esserman et al. (1) assume that the radiologists’ accuracy when interpreting a standardized test set of films corresponds to their accuracy in the clinical setting. Surprisingly, there is no evidence supporting this assumption. Radiologists’ interpretations of a test set are likely influenced by the evaluation process itself [i.e., the Hawthorne effect, (4,5)]. Rutter and Taplin (6) found no correlation between the performance of individual radiologists when interpreting a test set of 113 mammographic examination films and their individual level of accuracy noted during actual clinical care. Test sets, such as those used in the study by Esserman et al. (1), often contain a higher prevalence of disease cases than is seen in usual practice, which may bias the assessment of accuracy. For example, Egglin and Feinstein (7) reported that radiologists were more likely to interpret arteriograms as positive for pulmonary emboli when using a test film set with a higher prevalence of diseased cases than what occurs in practice. The federally funded Breast Cancer Surveillance Consortium, established in 1994, will allow us to directly evaluate the performance of mammography in a large sample drawn from diverse geographical areas and practice settings in the United States (8).

Another limitation of a testing situation is that it fails to provide a broader context of what is occurring in day-to-day community practices. Legal, financial, and personal characteristics, which would be expected to affect individual radiologists, differ greatly between the United States and the U.K. Studies (9–18) suggest that malpractice concerns are often on physicians’ minds and that those concerns influence their clinical practice patterns, such as patient referral and the use of tests and procedures. Issues related to breast cancer, particularly any delay in diagnosis, have been one of the most common causes of malpractice claims in the United States (19). It is possible that U.S. radiologists practicing in a more litigious environment have heightened concerns about medical malpractice, which ultimately influences their recall rates.

Financial environments such as those that offer incentives should also be considered when attempting to understand differences between radiologists in the United States and the U.K. For example, in the mid-1980s, a chain of ambulatory care centers in the United States changed its method of compensating physicians from a flat hourly wage to a compensation plan, including bonuses for the gross income generated by each individual physician (20). Physicians’ salaries and the frequency of ordering follow-up tests were tracked before and after instituting the new compensation plan. Results showed a 23% increase in the number of laboratory tests performed per patient, a 16% increase in the number of x-ray films ordered per patient, and a 19% increase in the wages of physicians who received bonuses. Would U.S. radiologists benefit from higher recall rates if they practiced in a for-profit system?

Although Esserman et al. (1) describe differences between the mammography screening systems of the United States and the U.K., they do not comment on possible differences in the expectations of the consumer: women in the United States may want and demand convenient access to mammography screening facilities and may take a more active role in deciding whether follow-up imaging should be done. Some women in the United States may receive mammography screening services in a facility where further diagnostic testing, if warranted, also can be done on the same day. Women may find this less stressful and more convenient. Facility management personnel, on the other hand, may find this approach costly and impractical from a staffing standpoint; this form of service delivery may affect interpretive practices of radiologists.

Finally, the legal and financial variations between the two countries may differently influence personal variables, such as the individual radiologist’s level of comfort with the ambiguity inherent in clinical decision making. Perhaps U.S. radiologists who enjoy interpreting mammograms feel more confident about their skills and, as a result, interpret higher volumes. Many factors may affect radiologists’ practice patterns; interpretation volume is just one of these. How personal characteristics of the radiologists might influence their clinical behavior is an important, but understudied, area.

Clinically significant variation exists among radiologists interpreting mammograms and is notably different between the United States and other countries. The reasons for this variability are not well understood but may relate to the financial, legal, clinical, and personal characteristics of the radiologists and/or the characteristics of the mammography facility. The study by Esserman et al. (1) in this issue of the Journal is a vital first step in improving our understanding of this important clinical topic. Their discussion, however, may oversimplify the task of improving the interpretation accuracy of U.S. mammography programs. We cannot look at the volume of films interpreted by radiologists as if they practice in a vacuum and then generalize the findings to the real world. Increasing the volume of interpretations for each mammographer in the United States may not result in improved accuracy if other influences, such as the fear of medical malpractice, financial rewards, or differing levels of comfort with ambiguity in clinical decision making, remain unchanged.

Practice probably does improve accuracy, but it may not make us perfect.

REFERENCES


**Note**

We appreciate the helpful comments from Drs. Suzanne Fletcher (Department of Ambulatory Care & Prevention, Harvard Medical School, Boston, MA) and Richard Deyo (Department of Medicine, University of Washington School of Medicine, Seattle, WA) on an early draft.