Radiology

REVIEWS AND COMMENTARY • EDITORIAL

Optimizing Breast Cancer Screening Programs: Experience and Structures

Robert D. Rosenberg, MD • *David Seidenwurm, MD*

From the Radiology Associates of Albuquerque, 4411 The 25 Way NE, Suite 150, Albuquerque, NM 87109 (R.D.R.); and Department of Diagnostic Imaging, Sutter Health, Sacramento, Calif (D.S.). Received April 22, 2019; revision requested April 29; final revision received and accepted May 1. **Address correspondence to** R.D.R. (e-mail: *rrosenb@unm.edu*).

Conflicts of interest are listed at the end of this article.

See also the article by Hoff et al in this issue.

Radiology 2019; 00:1–2 • **https://doi.org/10.1148/radiol.2019190924** • **Content code: BR**

The standards for accuracy and efficiency of a screening test are higher than those of most medical tests because they are applied to asymptomatic healthy people, with the expectation of long-term benefit. Poor quality is harmful and costly, while overregulation or unreasonable requirements are burdensome, limit access, and, therefore, need rigorous justification. Demanding image quality requirements for mammography are similar across Europe and the United States. However, the annual volume requirements and recommendations vary widely, from 960 mammograms within the past 2 years under the Mammography Quality Standards Act (1) to 3000–5000 mammograms per year required or recommended in European screening programs (2). In addition, the required initial experience under the Mammography Quality Standards Act is just 240 mammograms within a 6-month period. Precisely how annual interpretation, initial and overall cumulative experience, and program structure affect breast cancer screening program quality remain open questions despite decades of effort.

The recent study by Hoff and colleagues in this issue of *Radiology* (3) has three important findings concerning optimal interpretation volumes. First, radiologists' false-positive rates, and therefore specificity, are greatly impacted by experience and annual volumes. False-positive rates are high for the first 20 000 career mammograms and with a low annual volume of less than 2000 mammograms. Second, a radiologist's cancer detection rate is less associated with experience or annual volume and is optimal with 4000–10 000 annual mammograms. The few radiologists with annual volumes of more than 10 000 mammograms had the lowest cancer detection rates. Finally, double reading all mammograms with consensus review of discordant interpretations reduces false-positive rates and increases cancer detection.

Thus, optimal interpretation accuracy is improved with adequate initial experience, ongoing practice, and internal review. The screening environment in Norway with biennial mammography and double reading with consensus is different from that in single-reader systems, so the incremental benefits of suggested program changes must be validated cautiously as they are applied to existing practices.

Research in the United States by the Breast Cancer Surveillance Consortium (4–6) and in Canada (7) in single-reading screen-film environments showed results similar to those shown in Norway. There is improvement in performance associated with both higher initial experience and higher annual interpretation volumes. Specifically, performing work-up studies of recall cases (6), greater than 3 years of experience, and higher annual interpretation volumes are associated with improvement in the false-positive rate. Associations with sensitivity or cancer detection improvement were more difficult to identify (4,6,7).

What remains to be determined is which specific aspect of experience leads to improvement and how this can be targeted to reduce the time needed to acquire the needed skills. There may be two processes involved—one for cancer detection sensitivity and one for specificity improvement.

It seems that the consensus process, or review of the work-up process (7), provides useful feedback on false-positive cases because the overwhelming majority of diagnostic mammograms are negative. Cancer detection, however, requires adequate initial training and adequate interpretation time (3). Any research on sensitivity is limited by the relative scarcity of three to five cancers per 1000 in the screening environment and resultant poor statistical power of most research studies. This particularly limits the assessment of low-volume radiologists, as they will never achieve adequate cancer case volumes.

One important issue is the diversity in false-positive rates in different systems. Typical false-positive rates or recall rates in the United States and Canada are 9%–10% (4,6,7), compared with 4% in the study by Hoff et al. This is undoubtedly multifactorial. A mix of the medical-legal environments, habits, peer modeling, training customs, and societal differences contribute to this variation. There are no concrete incentives in the United States to reduce recall. Breast Cancer Surveillance Consortium work on optimizing recall in the screen-film environment suggests that a recall rate of 6%–8% is achievable at similar sensitivity for many radiologists. Recall may also decrease with use of multiple prior mammograms and avoidance of findings unlikely to be clinically important (8).

Several specific approaches for the improvement of overall accuracy are implied by Hoff and colleagues and the body of literature they build upon. However, these solutions may add new barriers to access or have other unintended consequences.

The first approach is to increase the minimal reader volume in screening while monitoring the screening detection rate in very high-volume readers.

The second approach is to minimize the medical-legal fears associated with low false-positive rates. General double reading may be optimal but is impractical in many systems. A first step could be double reading and consensus for any initially recalled and discordant cases. This should produce a diminished liability risk because at least two radiologists agreed at the time of screening that further work-up was not indicated. Careful monitoring of recall and cancer detection rates would be required during a transition period. Although any double reading would increase the radiologist time in screening mammography, the burden would be at least partly recouped through savings in the time- and resource-intensive diagnostic imaging process.

The third approach is to address the problem of the low rate of cancers in the screening environment and decrease the training time needed for new readers. There are methods to either add known cancers to the screening environment (9,10) or provide access to cancer-enriched screening mammography case sets. Any current requirement of ongoing or initial minimal screening volumes should also include minimal numbers of cancers detected, including those in enriched environments, since cancer detection is the immediate screening goal. Availability of enriched case sets or enriched screening also allows for ongoing assessment of cancer detection ability.

Incentives in payment structures that reward high cancer detection and lower recall rates are needed. Replacement of the procedure-based, atomized fee-for-service payment structure with payment for the annual screening episode, appropriately priced to include all services related to screening mammography, would accomplish this goal and would be administratively feasible. A single payment for the annual screening episode would include relevant professional and technical services through the initial biopsy. This type of plan would incentivize cancer diagnosis and disincentivize false-positive recall while permitting local decision making by each practice. Certain very rare patient populations may not be appropriate for this approach, but the majority of screening systems service typical populations. Medicare would be an ideal testing ground for this model of care due to the national scope and consistent enrollment of beneficiaries.

Finally, technology has changed our work environment, so we should use its benefits. Digital imaging permits easy implementation and testing of cancer detection improvement strategies as original images may be shared. Tomosynthesis has reduced recall compared with digital mammography, and we anticipate further technology improvements (eg, computer-aided detection and/or artificial intelligence and screening MRI and breast US). However, for the foreseeable future the greatest gains are likely to be achieved by optimizing the most precious resources in breast cancer screening: the radiologists' time and our patients' well-being.

In conclusion, sufficient data exist to recommend structural changes in screening mammography programs. Treating the whole patient and the system of care will lead us toward optimally performing systems.

Disclosures of Conflicts of Interest: **R.D.R.** disclosed no relevant relationships. **D.S.** Activities related to the present article: receives fees as a professional liability witness and personal injury expert. Activities not related to the present article: receives fees for expert testimony as a professional liability witness and personal injury expert; is employed by Sutter Medical Group; has stock/stock options in Sutter Medical Group and Radiological Associated Medical Group. Other relationships: disclosed no relevant relationships.

References

1. Mammography Quality Standards Act Regulations. U.S. Food & Drug Administration. https://www.fda.gov/Radiation-EmittingProducts/MammographyQualityStandardsActandProgram/Regulations/ucm110906.htm. Updated November 29, 2017. Accessed April 21, 2019.
2. Perry N, Broeders M, de Wolf C, Törnberg S, Holland R, von Karsa L, eds. European guidelines for quality assurance in breast cancer screening and diagnosis. 4th ed. Luxembourg: Office for Official Publications of the European Communities, 2006.
3. Hoff SR, Myklebust TA, Lee CI, Hofvind S. Influence of mammography volume on radiologists' performance: results from BreastScreen Norway. Radiology 2019; https://doi.org/10.1148/radiol.20191826842. Published online May 28, 2019.
4. Elmore JG, Jackson SL, Abraham L, et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology 2009;253(3):641–651.
5. Miglioretti DL, Gard CC, Carney PA, et al. When radiologists perform best: the learning curve in screening mammogram interpretation. Radiology 2009;253(3):632–640.
6. Buist DS, Anderson ML, Haneuse SJ, et al. Influence of annual interpretive volume on screening mammography performance in the United States. Radiology 2011;259(1):72–84.
7. Théberge I, Chang SL, Vandal N, et al. Radiologist interpretive volume and breast cancer screening accuracy in a Canadian organized screening program. J Natl Cancer Inst 2014;106(3):djt461.
8. Sickles EA. Successful methods to reduce false-positive mammography interpretations. Radiol Clin North Am 2000;38(4):693–700.
9. Gordon PB, Borugian MJ, Warren Burhenne LJ. A true screening environment for review of interval breast cancers: pilot study to reduce bias. Radiology 2007;245(2):411–415.
10. Evans KK, Birdwell RL, Wolfe JM. If you don't find it often, you often don't find it: why some cancers are missed in breast cancer screening. PLoS One 2013;8(5):e64366.