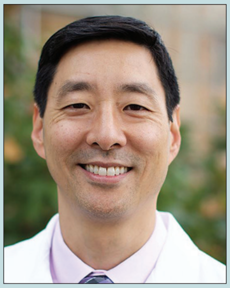
Radiology

A Warning about Warning Signals for Interpreting Mammograms

Solveig Hofvind, PhD • Christoph I. Lee, MD, MS

Dr Hofvind is head of BreastScreen Norway and professor at the University of Tromsø, Faculty of Health Sciences. Her main interests are the quality assurance, evaluation, and techniques and strategies used to improve mammographic screening.

Dr Lee is professor of radiology, adjunct professor of health systems and population health, and director of the Northwest Screening and Cancer Outcomes Research Enterprise at the University of Washington. He is a practicing breast imager with a research program focused on emerging breast cancer screening technology assessment.



Mammographic screening is aimed at detecting early-stage breast cancer, which leads to less aggressive treatment and reduced mortality. There is continuous effort to find the optimal balance between benefits and harms—that is, recall rate and cancer detection rate—to maximize the effectiveness of screening programs. In Europe, screening programs use double reading to ensure high accuracy for cancer detection while keeping recall rates low, usually below 4%, which is two to five times lower than in the United States where single reading is the norm (1). Use of warnings and markings given by computer-aided detection (CAD) has been used for more than 2 decades to optimize overall accuracy, and several new systems using artificial intelligence (AI) have been recently introduced with an ultimate goal of increasing the sensitivity and specificity of mammographic screening by supporting or replacing human radiologist interpretation (2,3).

The Dutch screening program has a unique warning system: Specialized mammography technologists preread screening examinations to identify and annotate possible abnormalities for their interpreting radiologists. This system was implemented as an effort to optimize the balance between the benefits and harms in the screening program while also making the technologists' role more attractive. In this issue of *Radiology,* Geertse et al (4) evaluated the effect of blinding or nonblinding the radiologists performing initial interpretations to technologists' warning signals. The authors found that blinding versus nonblinding readers to the technologists' prereading findings resulted in a lower recall rate (2.1% vs 2.4%, respectively), a higher positive predictive value of recall (30.6% vs 26.2%), and no difference in cancer detection rate (6.5 vs 6.4 per 1000 screening examinations). The authors concluded that prereading of screening mammograms by technologists and their warning signals to nonblinded radiologists at the start of their initial interpretations did not add any benefit to women. Of note, technologists' warning signals were considered for the blinded group at quality assurance sessions, where technologists and radiologists were both present after the initial independent radiologist interpretations. These consensus sessions led to additional recalls and cancer detection for the blinded group. However, the overall recall rate remained lower for the blinded versus nonblinded group, and the cancer detection rates were not significantly different.

Although the use of technologist prereading may be unique to the Dutch screening program, the findings of Geertse et al regarding timing and use of outside warning signals and annotations by interpreting radiologists is salient to current efforts in mammographic screening technologies. In both its objectives for improved accuracy and its process for identifying and alerting the radiologist to areas of potential suspicious imaging findings, the Dutch prereading strategy is analogous to CAD and newer AI-driven adjunct mammographic screening technologies. All of these warning systems are meant to help streamline the work of interpreting radiologists by helping them focus on the most suspicious area of a mammogram. However, as we have learned from the study by Geertse et al and seminal studies on the accuracy of traditional CAD for mammography, these systems may not be benefiting interpreting radiologists or the women undergoing routine screening (2,5).

One of the most likely reasons for this phenomenon is the bias introduced in the human interpretive process by early warning signals (6). Radiologists likely focus on areas marked as suspicious by an independent reviewer—whether human or computer—with a higher likelihood of

From the Section for Mammographic Screening, Cancer Registry of Norway, PO Box 5313, Majorstuen, Oslo 0304, Norway (S.H.); Department of Health and Care Sciences, UiT–The Arctic University of Norway, Tromsø, Norway (S.H.); Department of Radiology, University of Washington School of Medicine, Seattle, Wash (C.I.L.); and Department of Health Systems and Population Health, University of Washington School of Public Health, Seattle, Wash (C.I.L.). Received August 16, 2021; revision requested September 7; revision received and accepted September 8. **Address correspondence to** S.H. (e-mail: *sshh@kreftregisteret.no*).

C.I.L. supported in part by the National Cancer Institute (grant R37 CA240403).

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

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

Radiology 2022; 302:284–285 • https://doi.org/10.1148/radiol.2021212092 • Content codes: BR OI • ©RSNA, 2021

calling back women with mammograms marked as suspicious. Providing warning signals to radiologists may lead to their overreliance on the signals without maintaining their own vigilance in the independent interpretive process, or may reduce their attention to other suspicious areas on mammographic images. With traditional CAD, studies have demonstrated that there were 1.5 to four false-positive markings per screening examination, leading to more false-positive screening results (7).

The introduction of bias in what is meant to be an independent interpretive process for radiologists is an especially important concern as newer AI technologies for mammographic interpretation are gaining regulatory approval and becoming commercially available. AI technologies, compared with traditional CAD or technologist prereading, are being marketed as having higher interpretive accuracy than that of radiologist interpretation alone. If adopted and applied broadly based on solely these promises, then AI is likely to lead to higher introduction of bias or, more specifically, automation bias (6). In automation bias, the human interpreter finds it difficult to disagree with what is perceived as a smarter and more accurate supercomputer. How can a radiologist not call back women with a preannotated finding marked as suspicious on the mammogram by algorithms that are supposedly more accurate than they are?

The study by Geertse et al also points out the importance of when such additional warning signals should be presented to radiologists, if at all. In double-reading environments such as the Dutch screening program, consensus meetings are meant to keep the recall rate low while maintaining appropriate cancer detection rates (8). The objective of these meetings is to deselect cases and thereby increase the positive predictive value of recalls and the specificity of the screening program. Two human interpretations have already been performed and additional assessment by a third radiologist is often used as a tiebreaker for recall or no recall. In the study by Geertse et al, the technology prescreening results were applied to the blinded group only at the consensus stage (4). As we look ahead on how to implement AI, this study suggests that one route to diminish automation bias at the time of initial interpretation is to hold AI results until a later decision point. Perhaps AI should be used to replace the consensus process, rather than influencing initial independent human interpretation in double-reading settings.

Finally, the Dutch study points out an important aspect of warning signals like CAD and AI for mammographic screening: the ideal format for informing radiologists with warning signals or AI scores (4). The user interface for computer algorithms and human radiologists is vastly understudied. Although automation bias among radiologists is likely with newer AI technologies for mammography, this has not been shown in prospective studies. Future research efforts for the optimal user interface for these early warning signals are needed prior to their widespread adoption. The findings from the Dutch study suggest availability to warning signals should occur after independent interpretation to avoid the introduction of interpretive bias. This may be possible in double-reading environments with consensus but not for single-reader environments. Moreover, additional interpretation by expert technologists does not appear to add value to mammographic screening programs. Any additional warning signals will certainly have to have substantially higher cancer detection rates and lower recall rates than radiologists. The optimal timing and approach for incorporating early warning signals from emerging AI tools in both single- and double-reading environments remain to be determined and are open, critical areas for further research.

Disclosures of Conflicts of Interest: **S.H.** No relevant relationships. **C.I.L.** Textbook royalties from McGraw Hill, Oxford University Press, Wolters Kluwer; personal fees for journal editorial board work for American College of Radiology; personal fees for work on Data Safety Monitoring Board for GRAIL.

References

1. Smith-Bindman R, Ballard-Barbash R, Miglioretti DL, Patnick J, Kerlikowske K. Comparing the performance of mammography screening in the USA and the UK. J Med Screen 2005;12(1):50–54.
2. Chan HP, Samala RK, Hadjiiski LM. CAD and AI for breast cancer-recent development and challenges. Br J Radiol 2020;93(1108):20190580.
3. Le EPV, Wang Y, Huang Y, Hickman S, Gilbert FJ. Artificial intelligence in breast imaging. Clin Radiol 2019;74(5):357–366.
4. Geertse TD, Setz-Pels W, van der Waal D, et al. Added value of prereading screening mammograms for breast cancer by radiologic technologists on early screening outcomes. Radiology 2022;302(2):276–283.
5. Lehman CD, Wellman RD, Buist DS, et al. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med 2015;175(11):1828–1837.
6. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc 2017;24(2):423–431.
7. Houssami N, Given-Wilson R, Ciatto S. Early detection of breast cancer: overview of the evidence on computer-aided detection in mammography screening. J Med Imaging Radiat Oncol 2009;53(2):171–176.
8. Perry N, Broeders M, de Wolf C, Törnberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis. 4th edition. Luxembourg: Office for Official Publications of the European Communities. Published 2006. Accessed October 17, 2021. https://screening.iarc.fr/doc/ND7306954ENC_002.pdf.