Digital medicine

Readiness for mammography and artificial intelligence

One area that has attracted great attention for the use of deep learning artificial intelligence (AI) in health care is medical imaging, especially mammography. Many initial AI studies proclaimed remarkable improvement in accuracy over the performance of radiologists, but a recent systematic review highlighted there is insufficient scientific evidence to support such findings. The UK National Screening Committee commissioned Freeman and colleagues to review the quality and results of studies that assessed the accuracy of AI algorithms, alone or in combination with radiologists, to detect cancer in digital mammograms; 34 of 36 AI systems evaluated were less accurate than a single radiologist, and all were less accurate than the consensus of two or more radiologists.

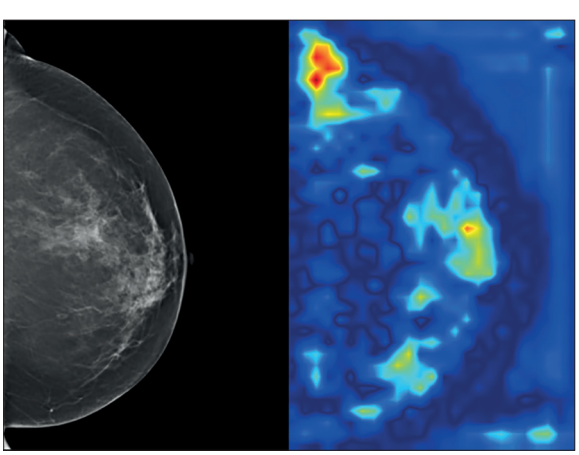
This situation is reminiscent of how computer-aided detection (CAD) tools were trumpeted as improving accuracy in screening mammography. However, subsequent large-scale studies showed no benefit of CAD tools in improving radiologist diagnostic performance. Deep learning models can extend human vision and knowledge in ways that CAD cannot and their strongest potential role could be in new applications beyond assisting the radiologist to detect cancer. Research has shown the potential of deep learning models to triage low-risk mammograms, maintaining sensitivity and improving specificity and efficiency. But these studies are based on retrospective, in-silico data resources that may not be representative of real-world clinical practice.

Screening mammography provides a rich domain for AI because of evidence supporting decreased morbidity and mortality of breast cancer through early detection, knowledge that mammography is the best tool available to detect preclinical breast cancer, and agreement that mammography has shortcomings. These flaws, which AI has the potential to address, include wide variation in human interpretation, false-positive and false-negative results, overall costs, and restricted global access due to shortages of specialised radiologists to interpret mammograms.

Screening mammography has key elements that make it suitable for deep learning models. The screening test itself has a binary outcome. From the screening mammogram, the patient is either cleared or recalled for additional imaging and possible biopsy. The diagnosis is also binary in that the patient is classified as disease positive or negative. Even with these narrow parameters, there has been substantial variation in the methods used to train, test, and validate deep learning models in the interpretation of mammograms. In the review by Freeman and colleagues, the AI studies were deemed to be of poor methodological quality because the study designs were retrospective and had high risk of bias based on cancer-enriched samples, reader study laboratory effect, differential verification of outcomes, and insufficient follow-up. These biases probably led to overestimation of sensitivity and underestimation of specificity. This cautionary note is precisely what occurred in CAD studies. Indeed, there is much overlap in the methods that were used historically with CAD and those now used with AI tools.

The development of deep learning models has not yet been followed by implementation and assessment in routine clinical practice. We can learn from the history of CAD applied to mammography and leverage that knowledge to more rapidly translate our discoveries in AI to improved patient care. Future studies should consider how to design studies best suited for the precise clinical application intended for the deep learning model. A model to assist radiologists to interpret mammograms with a higher degree of accuracy will require a different study design from one designed to fully replace human screening interpretation. The fusion of machine and expert human vision is the combination that needs to be emphasised. The promise of AI has not been diminished by awareness of the limitations of published studies. We are moving closer to the goal of leveraging the true power of AI to solve our greatest challenges in early and accurate breast cancer detection and diagnosis. Ultimately, this will require compelling evidence from optimally designed prospective studies in large cohorts.

Constance D Lehman, Eric J Topol
Mass General Brigham, Harvard Medical School, Radiology, Boston, MA 02114, USA (CDL); Scripps Research Translational Institute, Scripps Research La Jolla, CA, USA (EJT)
clehman@mgh.harvard.edu


Clairity, Inc

Further reading

Freeman K, Geppert J, Stinton C, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *BMJ* 2021; **374:** n1872

Lång K, Dustler M, Dahlblom V, Åkesson A, Andersson I, Zackrisson S. Identifying normal mammograms in a large screening population using artificial intelligence. *Eur Radiol* 2021; **31:** 1687–92

Yala A, Mikhael PG, Strand F, et al. Toward robust mammography-based models for breast cancer risk. *Sci Transl Med* 2021; **13:** eaba4373

CDL is supported by the Breast Cancer Research Foundation and receives institutional support from GE Healthcare and Hologic, Inc. She is co-founder with equity in Clairity, Inc, a company developing deep learning models to predict cancer risk. EJT is supported by the US National Institutes of Health/National Center for Advancing Translational Sciences grant UL1TR001114.