Radiology

REVIEWS AND COMMENTARY • EDITORIAL

Assessing Cancer Risk from Mammograms: Deep Learning Is Superior to Conventional Risk Models

Arkadiusz Sitek, PhD • *Jeremy M. Wolfe, PhD*

From IBM Watson Health, 75 Binney St, Cambridge, MA 02142 (A.S.); and Brigham and Women's Hospital, Harvard Medical School, Boston, Mass (J.M.W.). Received April 5, 2019; revision requested April 12; revision received April 15; accepted April 15. **Address correspondence to** A.S. (e-mail: *arek@ibm.com*).

J.M.W. supported by the National Institutes of Health National Cancer Institute (grant CA207490).

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

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

Radiology 2019; 00:1–2 • https://doi.org/10.1148/radiol.2019190791 • Content code: BR

The last few years have produced enormous growth in the radiologic application of computer vision deep learning (DL) algorithms and machine learning, often referred to as artificial intelligence (1). In this issue of *Radiology*, Yala et al (2) demonstrated the effectiveness of DL methods in assessing breast cancer risk by using clinical data, breast density scores, and mammograms. On March 28, 2019, the U.S. Food and Drug Administration announced a proposed rule (3) to update the landmark policy passed by Congress in 1992 to ensure quality of mammography for early breast cancer detection (known as the Mammography Quality Standards Act). One of the major amendments proposed by the Food and Drug Administration is a requirement to report breast density and communicate this finding in lay language to patients. This announcement makes the Yala et al article timely because it illustrates the limits of a simple breast density score and points to better estimates of the risk of developing breast cancer that can be derived from information obtained at mammography.

Yala et al (2) retrospectively examined nearly 90 000 consecutive screening mammographic examinations from almost 40 000 women obtained over 4 years (2009–2012) at Massachusetts General Hospital (Boston, Mass). The authors defined four breast cancer risk models intended to quantify the probability of discovering breast cancer within 5 years after a mammographic screening examination that was negative for cancer. The first model used clinical data including breast density described by using four categories (almost entirely fatty, scattered areas of fibroglandular tissue, heterogeneously dense, and extremely dense). Risk was calculated by using a model developed by the authors. The second model calculated risk by using the clinical standard Tirer-Cuzick model that also included breast density (4,5). The third model calculated risk on the basis of DL analysis of full-resolution mammography only. The fourth model was a hybrid that combined DL analysis of full-resolution mammography with clinical data, including breast density.

Yala et al (2) found that DL risk models (ie, the third and fourth models) consistently performed better than the first and second models for prediction of breast cancer within 5 years after mammography. The hybrid DL model was the best and the clinical data model was the worst. DL risk models performed similarly across demographic subgroups, and the fourth model produced an area under the receiver operating characteristic curve of 0.71 for white and African American women. This lack of difference between groups is clinically important because traditional models, such as the second model, do not perform equally well for these two groups. For example, Yala et al showed that the diagnostic performance of the Tirer-Cuzick model was different for white versus African American women (areas under the receiver operating characteristic curve, 0.62 and 0.45, respectively).

The third model, which used only DL analysis of full-resolution mammography, outperformed the first and second models. This suggested that there was more information about breast cancer risk on the mammograms than in the clinical data used by standard risk models. Because breast density scores are included in the standard models, it follows that those density scores do not reflect all the relevant information on the mammogram. This is, perhaps, unsurprising when one considers that a four-value density score (2 bits of information) is a result of compression of the hundreds of millions of bits of information contained on a mammogram.

The study by Yala et al (2) offered some interesting and significant evidence that DL can contribute to risk assessment. Of course, as the authors noted, there are limitations to this work. Whereas this is a relatively large study, it is retrospective and carries all dangers of biases associated with this approach to study design. Moreover, the study was performed by using data from a single institution and mammograms acquired by using the scanners from a single vendor.

Like computer algorithms, radiologists can also find unanticipated information at mammography. Experts can distinguish normal from abnormal mammograms after just a 500-msec exposure (6). Importantly, for these purposes, they can perform this task with the images of the breast contralateral to the lesion (7) and, like the DL described here, human experts can classify images as normal or abnormal at above-chance levels even when those images were acquired 3 years before the diagnosis of breast cancer (8). These human observer studies have been modeled on studies of so-called gist perception in the vision science literature. Those studies show that an exposure time of a fraction of a second is all that is needed for humans to extract basic semantic information about a scene (eg, is this natural or man-made? Is

an animal present? Is this scene navigable? [9]). Now that it has been shown that some so-called gist of cancer risk can be detected years before the cancer is diagnosed, new experiments can try to optimize this ability, perhaps by encouraging the radiologist to formally assess the gist or texture of the breast as part of the effort to assess the patient's risk.

Similar to the DL algorithm, this human gist perception is detecting something beyond a simple density measure (6–8). With both the human and DL capabilities, we are not sure what signal is detected. DL results are not easily explainable to humans. DL methods are so-called black boxes that provide little guidance about why conclusions were reached. Yala et al (2) do not speculate on what exactly makes one mammography image predictive of the future occurrence of breast cancer. Further research is needed to uncover the signals supporting the DL and human abilities. If radiologists and DLs are uncovering different signals at mammography, there is potential for further improvement in a hybrid human-DL system or a DL algorithm that learns the signal that human experts use. We suggest that more rapid progress will be made if the hybrid models combine not just DL and clinical data but also combine human and machine capabilities.

Effort to make deep learning (DL) methods fully explainable is an area of active research in academia and industry (10). At the moment, it is unclear whether DL methods will ever be fully explainable. Are we willing to follow computer recommendations in radiology without completely understanding them if we know that they provide sound advice? This is a complex, multilevel question that we in the community must be asking at the dawn of an era when artificial intelligence will be affecting decision making in radiology and health care. There can be little doubt that more deep learning studies will produce further advances of the sort described by Yala et al (2). The effect of DL on patient outcomes is yet to be demonstrated. As a community, we must learn how to embed and use these technologies in clinical care and how to synergistically use DL and human visual perception to build confidence and trust in DL systems in radiology, including in mammogram-based breast cancer risk assessment.

Disclosures of Conflicts of Interest: A.S. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed money paid to author for employment by IBM Watson Health. Other relationships: disclosed no relevant relationships. **J.M.W.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed money paid to author's institution for grants/grants pending from NIH/NCI, NIH/NEI, General Electric, and NSF; disclosed paid honoraria for talks from University of Minnesota, St Vincent's Hospital (Worcester, Mass), University of California, Berkely; disclosed royalties from Sinauer; disclosed stock/stock options in Synapse Technology Corporation. Author disclosed travel/accommodations/meeting expenses from Pinterest; Center for Brain, Minds, and Machines; Zurich University of Applied Sciences; AgeLab MIT; Philips Healthcare; Childrens Hospital (Boston, Mass); IBM Watson Health; and Woods Hole Brains, Minds, and Machines course. Other relationships: disclosed that National Cancer Institute funds a lab for the author at the RSNA Annual Meeting and Barco provides some in-kind support.

References

1. Nam JG, Park S, Hwang EJ, et al. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. Radiology 2019;290(1):218–228.
2. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology 2019. https://doi.org/10.1148/radiol.2019182716. Published online May 7, 2019.
3. Proposed rule to Mammography Quality Standards Act, FDA, March 28, 2019. https://www.federalregister.gov/d/2019-05803. Accessed April 22, 2019.
4. Brentnall AR, Harkness EF, Astley SM, et al. Mammographic density adds accuracy to both the Tyrer-Cuzick and Gail breast cancer risk models in a prospective UK screening cohort. Breast Cancer Res 2015;17(1):147.
5. IBIS Breast Cancer Risk Evaluation Tool. http://www.ems-trials.org/riskevaluator/. Accessed April 22, 2019.
6. 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.
7. Evans KK, Haygood TM, Cooper J, Culpan AM, Wolfe JM. A half-second glimpse often lets radiologists identify breast cancer cases even when viewing the mammogram of the opposite breast. Proc Natl Acad Sci U S A 2016;113(37):10292–10297.
8. Brennan PC, Gandomkar Z, Ekpo EU, et al. Radiologists can detect the 'gist' of breast cancer before any overt signs of cancer appear. Sci Rep 2018;8(1):8717.
9. Greene MR, Oliva A. The briefest of glances: the time course of natural scene understanding. Psychol Sci 2009;20(4):464–472.
10. Seah JCY, Tang JSN, Kitchen A, Gaillard F, Dixon AF. Chest radiographs in congestive heart failure: visualizing neural network learning. Radiology 2019;290(2):514–522.