Title
Artificial Intelligence for Breast Cancer Imaging: The New Frontier?

Journal
Journal of the National Cancer Institute, 111(9)

ISSN
0027-8874

Authors
Lee, Christoph I
Elmore, Joann G

Publication Date
2019-09-01

DOI
10.1093/jnci/djy223

Peer reviewed

JNCI J Natl Cancer Inst (2019) 111(9): djy223

doi: 10.1093/jnci/djy223
First published online February 5, 2019
Editorial

EDITORIAL

Artificial Intelligence for Breast Cancer Imaging: The New Frontier?

Christoph I. Lee, Joann G. Elmore

See the Notes section for the full list of authors' affiliations.
Correspondence to: Christoph I. Lee, MD, MS, Department of Radiology, University of Washington School of Medicine, 1144 Eastlake Avenue East, LG-212, Seattle, WA 98109 (e-mail: stophlee@uw.edu).

There is great excitement around artificial intelligence (AI), the use of computers to mimic human cognitive functions, and its promise of automating time-consuming and repetitive tasks in medicine. One prime target for AI is the analyses of visual input data. For example, high-powered computers using AI algorithms already rival ophthalmologists in identifying diabetic retinopathy on fundus screening images (1). Now, screening mammography is taking center stage as an obvious target where 40 million women in the United States alone undergo the exam annually, and radiologists have traditionally been less than perfect in their interpretive performance (2, 3). Implied in these efforts is the provocative notion that certain tasks that are currently performed by highly specialized physicians can be completely replaced by super computers that run AI systems.

In this issue of the *Journal of the National Cancer Institute*, Rodriguez-Ruiz and colleagues (4) efficiently used existing imaging datasets from nine prior studies to compare the interpretations of their commercial AI system with interpretations of radiologists. In all, the study team compared the performance of their AI system to radiologists' performance on 2652 digital mammography screening exams. Their AI system generated a probability of malignancy score between 1 and 10, whereas radiologists provided an assessment based on the Breast Imaging and Reporting Data System (BI-RADS) assessment scale that ranges from 0 to 6. The area under the receiver operating characteristic curve (AUC) for the AI system was 0.840 compared with an AUC of 0.814 for the radiologists. Based on these findings, the authors conclude that the AI system was as accurate in screening mammography interpretation as the radiologist.

This study is a promising example of the emerging era of AI-based interpretation of medical image data. Rodriguez-Ruiz and colleagues (4) reported that their AI algorithm was previously trained and tested using 9000 true positive and 180 000 true negative mammograms not used in the current study for validation. In this analysis, the authors use a diverse set of screening digital mammograms obtained from seven countries and multiple imaging equipment vendors, with ground truth for the presence of cancer based on histopathologic data and clinical follow-up. While these results are compelling, multiple additional steps need to be taken prior to widespread adoption of commercial AI systems for automated mammography screening in routine practice.

First, Rodriguez-Ruiz and colleagues (4) used existing imaging sets from prior studies that compared mammography to another imaging technology (eg, digital breast tomosynthesis). To demonstrate differences in sensitivity and specificity between two screening technologies, these cohorts were enriched with a greater proportion of cancer-positive mammograms than seen in routine screening practice. The enhanced frequency of abnormal findings among these images could trigger greater radiologist scrutiny (context bias), resulting in higher recall rates of the radiologists with resultant lower observed accuracy. Indeed, the AUC of 0.814 among radiologists is considerably lower than the AUC of more than 0.90 observed in actual U.S. community practice (5). Thus, to truly demonstrate efficacy of an AI algorithm, larger validation datasets are needed that are more representative of a screening population.

Second, incremental improvement in the AUC is not directly translatable to improved patient outcomes in the clinical setting. In the case of this commercial AI system, the output is an estimated probability of malignancy on a scale of 1–10. However, it is unclear what threshold probability of malignancy value would trigger further diagnostic workup if the AI system was used in clinical practice. Currently, subjective interpretation by radiologists using the standardized BI-RADS is categorized as 0–6 in a noncontinuous fashion (with BI-RADS 3, 4, and 5 requiring further diagnostic workup). It is not clear how to compare a 10-point linear scale from an AI system with our current noncontinuous BI-RADS scale. For instance, a radiologist's suspicion of more than 2% but less than 95% malignancy is

Received: November 15, 2018; **Accepted:** November 29, 2018

given a BI-RADS 4 assessment, whereas a radiologist's suspicion of more than 95% malignancy is given a BI-RADS 5 assessment. It is uncertain what proportion of exams that a commercial AI system would flag as having more than 2% malignancy, requiring additional diagnostic workup under our current clinical practice thresholds. If the vast majority of screening exams are given a probability of a malignancy score of more than 1 on a 10-point scale (suggesting >2% change of malignancy), then the technology would lead to an excessively large proportion of women called back for additional imaging (compared to just 10% of women currently), rendering the AI system useless under accepted practice thresholds.

Third, it is uncertain if patients and physicians would accept medical imaging devoid of any human involvement and interpretation (6). In the absence of a radiologist reviewing these mammography images, who would take ultimate responsibility for breast cancers missed by an imperfectly performing AI algorithm? Radiologists will still likely be tasked with reviewing abnormalities noted by the commercial AI algorithm on the screening exams, ordering and interpreting the subsequent diagnostic imaging, and then performing any final image-guided tissue biopsies to confirm malignancy. Exactly where and how AI algorithms will support the practice of interpreting radiologists within their current workflow remains unknown. These nuances will be important because missed breast cancers on screening mammography remain the most litigious situation for medical malpractice lawsuits (7).

For now, AI holds incredible promise for rapidly and reproducibly interpreting vast amounts of medical image data. However, like other emerging technologies, AI systems for automated breast cancer detection require robust evaluation for clinical effectiveness before broad adoption. We learned from the experience of computer-aided detection in mammography that adopting promising new technologies too quickly could be a costly mistake; computer-aided detection was accepted and reimbursed as an adjunct to digital mammography in the early 2000s based on hype but little evidence and later found to lead to more false positives without improved cancer detection (8). As we enter this exciting and rapidly evolving new frontier, it will be important that AI systems for breast cancer screening be validated on multiple, diverse imaging datasets representative of screening populations. Moreover, the optimal interface between radiologists and AI systems must be determined, with ascertainment of broader stakeholder tolerance for inevitable missed cancers and false positive workups.

Notes

Affiliations of authors: Department of Radiology, University of Washington School of Medicine, Seattle, WA (CIL); Department of Health Services, University of Washington School of Public Health, Seattle, WA (CIL); Hutchinson Institute for Cancer Outcomes Research, Seattle, WA (CIL); Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA (JGE).

The authors report no conflicts of interest related to this article. Dr Lee reports grant funding from GE Healthcare unrelated to this work.

References

1. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402–2410.
2. Trister AD, Buist DSM, Lee CI. Will machine learning tip the balance in breast cancer screening? *JAMA Oncol*. 2017;3(11):1463–1464.
3. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *N Engl J Med*. 1994;331(22):1493–1499.
4. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst*. 2019;111(9):916–922.
5. Barlow WE, Chi C, Carney PA, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. *J Natl Cancer Inst*. 2004;96(24):1840–1850.
6. Houssami N, Lee CI, Buist DSM, Tao D. Artificial intelligence for breast cancer screening: opportunity or hype? *Breast*. 2017;36:31–33.
7. Arleo EK, Saleh M, Rosenblatt R. Lessons learned from reviewing breast imaging malpractice cases. *J Am Coll Radiol*. 2014;11(12 pt A):1186–1188.
8. Fenton JJ, Lee CI, Xing G, Baldwin L, Elmore JG. Computer-aided detection in mammography: downstream impact on diagnostic testing, ductal carcinoma in-situ treatment, and costs. *JAMA Intern Med*. 2014;174(12):2032–2034.