The Breast 52 (2020) 146–149

Contents lists available at ScienceDirect

The Breast

journal homepage: www.elsevier.com/brst

Original article

Pathways to breast cancer screening artificial intelligence algorithm validation

Christoph I. Lee [a, *], Nehmat Houssami [b], Joann G. Elmore [c], Diana S.M. Buist [d]

[a] Department of Radiology, University of Washington School of Medicine, Department of Health Services, University of Washington School of Public Health, Hutchinson Institute for Cancer Outcomes Research, Seattle, WA, USA
[b] The University of Sydney, Faculty of Medicine and Health, Sydney School of Public Health, Australia
[c] Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA
[d] Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA

ARTICLE INFO

Article history:
Received 30 July 2019
Accepted 8 September 2019
Available online 9 September 2019

Keywords:
Artificial intelligence
Breast cancer
Screening
Mammography
Population health
Validation
Transparency
Reproducibility

ABSTRACT

As more artificial intelligence (AI)-enhanced mammography screening tools enter the clinical market, greater focus will be placed on external validation in diverse patient populations. In this viewpoint, we outline lessons learned from prior efforts in this field, the need to validate algorithms on newer screening technologies and diverse patient populations, and conclude by discussing the need for a framework for continuous monitoring and recalibration of these AI tools. Sufficient validation and continuous monitoring of emerging AI tools for breast cancer screening will require greater stakeholder engagement and the creation of shared policies and guidelines.

1. Introduction

Promising reports of artificial intelligence (AI) algorithms from reader studies involving limited imaging case sets indicate that they may improve mammography screening accuracy beyond radiologist interpretation alone [1–3]. Several of these AI tools have garnered medical device regulatory approval within multiple countries, including in the U.S. from the Food and Drug Administration (FDA) [4]. With regulatory approval, these commercial products can now be marketed for clinical use directly to stakeholders including radiologists and physician groups.

However, rigorous validation in large, diverse patient populations that were not involved in the original AI algorithm development is required before clinical translation. Moreover, key stakeholders, including major payers, providers, and women undergoing routine screening, need convincing evidence that these new tools can reliably improve screening performance beyond current practice standards. Given the "black box" nature of AI algorithms, there are a number of unique challenges in the process of algorithm validation and stakeholder acceptance. There is a myriad of technical, social, political, and ethical issues regarding AI algorithm validation, as well as multiple stakeholder viewpoints, beyond the scope of a single article. Thus, in this viewpoint, we focus on some of the major pressing issues in validating algorithms from the perspective of AI developers, organizations with imaging data, and regulatory agencies.

1.1. Learning from the past

In the U.S., the bar for FDA medical device approval remains low with small reader studies showing non-inferiority to existing performance being sufficient for regulatory clearance [5]. Case sets used in FDA approval reader studies usually number in the hundreds of exams and are enriched with positive cases, making performance measures unreliably applicable to routine screening at the population level where positive cases are more seldom encountered [6]. Once this initial regulatory bar is met, however, an AI software device can be marketed directly to radiology groups and health systems without the need for reproducibility in multiple

* Corresponding author. Department of Radiology University of Washington School of Medicine, 1144 Eastlake Avenue East, LG-212, Seattle, WA, 98109, USA.
E-mail address: stophlee@uw.edu (C.I. Lee).

https://doi.org/10.1016/j.breast.2019.09.005

populations and settings.

There are serious consequences of adopting new technologies without supporting validation and reproducibility in medicine [7]. In mammography screening, we have encountered these consequences with traditional computer aided detection (CAD) software. CAD was rapidly adopted around the turn of the century in combination with digital mammography (which was concurrently replacing screen-film mammography as a primary screening modality) without robust observational studies or randomized trials to suggest improved screening performance at the population level [8]. Instead, based on small reader studies used to gain FDA approval and heavy lobbying from vendors to obtain reimbursement, CAD was widely adopted into clinical practice [8]. Unfortunately, through observational studies performed over the next decade, CAD was eventually found to increase false-positives and benign biopsies without increasing cancer detection rate [9–11]. The result was a substantial increased cost to healthcare systems and women undergoing screening, without realization of the promised benefit [10,11].

1.2. Validation on newer screening technologies

One of the first steps towards clinical translation for promising AI algorithms trained and tested on existing mammography datasets will be adaptation to and validation using frequently evolving screening technologies. To date, nearly all of the published reader studies demonstrating improved screening accuracy with AI have used 2D digital mammography or screen-film mammography [6]. The largest mammography AI study to date, the Digital Mammography DREAM Challenge, provided 2D mammography images and associated clinical data representing >640,000 images from >86,000 women to DREAM Challenge participants for training and validation of their deep learning algorithms for automated mammography interpretation [12].

While digital mammography is currently the most widely used imaging modality for breast cancer screening and further AI algorithm development is needed prior to dissemination and clinical implementation, screening imaging technologies themselves are rapidly changing. Digital breast tomosynthesis (DBT, 3D mammography) is quickly usurping the role of digital mammography as the first-line screening modality of choice in many settings. This is in part due to population-based studies suggesting higher cancer detection rate and possibly lower interval cancer rate with DBT compared to digital mammography [13,14]. With the majority of U.S. facilities and many European population-based programs evaluating or transitioning to DBT screening, a strong argument can be made that current AI algorithms need to be effectively scaled from 2D to 3D volumetric data. While the assumption is that 2D to 3D algorithm scaling will be straightforward, this is not guaranteed. Thus, in order to remain relevant for clinical application, emerging AI algorithms will need to be trained and validated on large DBT imaging datasets and be able to adapt to further advancements in primary screening imaging modalities in the future.

1.3. Defining sufficient algorithm validation

After a promising AI algorithm has been trained and tested on a large modern imaging dataset and has gained regulatory approval, external validation in diverse patient populations is needed to demonstrate generalizability and clinical effectiveness. There is increasing concern that AI models have structural biases based on the imaging exams and populations included in initial training and testing with calls for more distributive justice in the initial model design, evaluation, and deployment [15]. AI developers will need to demonstrate improved screening performance in large diverse populations, including women of minority race/ethnicity and differing breast cancer risk factors.

External validation should be performed in population-based screening programs and also in many different clinical settings (i.e. double reader and single reader environments). In countries without centralized screening programs, such as the U.S., algorithms need to be validated in large health systems and in different geographic regions in order to ensure that there is no unintended bias against specific subpopulations, especially traditionally vulnerable populations. It is also uncertain whether retrospective validation (the predominant approach used in studies of AI for breast cancer detection thus far [6]) is sufficient or if prospective randomized and/or pragmatic trials are needed to convince key stakeholders that AI-driven mammography screening (with or without radiologist involvement) is more accurate than traditional radiologist interpretation alone. In other words, the actual threshold required to validly claim external validation of a promising AI algorithm remains up for debate. The guideposts for the adequacy in size of validation population datasets, diversity of the validation populations, and improved accuracy measures of AI-based screening over human interpretation alone are currently unknown.

2. Access to population-based data

Gaining access to validation datasets, even retrospectively, is currently fraught with differing priorities among imaging stakeholders and unequal access among AI developers. In order to be useful, mammography images representing populations served by regional screening programs and high quality registries need to be linked to eventual cancer outcomes in order to determine the ground truth. Thus, access to useful imaging datasets requires access to not only the images themselves but also complete cancer history (e.g., prior lumpectomy for breast cancer) and follow-up data on all women to define eventual cancer outcome status (e.g., clinical records, biopsy results). The result is the need for complex data use agreements, especially with intellectual property of eventual AI algorithms at stake.

From the perspective of imaging data owners (health systems, radiology groups, and women), privacy concerns for biomedical data remain a major concern. Institutions may be reluctant to release millions of imaging exams for private developer use [16]. One potential solution is a shift from a traditional model of transferring data directly to data modelers to an alternative “model to data” paradigm where the flow of data is reversed [17]. This paradigm was used successfully in the Digital Mammography DREAM Challenge where participants submitted containerized AI models to the Challenge organizer to train and validate the submitted models on untouched imaging data behind a firewall. While this has greater protection for health information, an important disadvantage is that AI data scientists have limited access to images, which could impede their ability to optimize their algorithms.

Finally, since the vast majority of players in this arena are looking to develop and commercialize their AI tool, there is fierce competition in gaining access to limited numbers of data partners willing to broker cooperative agreements with AI developers, especially for imaging data that includes exams from vulnerable populations. With intellectual property at stake, major industry players with larger resources and the ability to pay for use of imaging data have a distinct advantage over smaller start-ups, making current access to larger, diverse validation datasets inequitable. The end result is that not all AI developers will have access to validation datasets across diverse populations, potentially further exacerbating screening disparities by rendering eventual clinical

algorithms less effective among already vulnerable patient populations.

3. Continuous improvement and monitoring

New AI tools for mammography have the advantage of continuously learning compared to traditional CAD. Ideally, AI mammography algorithms would not go through validation just once, but would undergo continuous refinement and validation over time in order to not repeat the missteps experienced with traditional CAD. However, the fluid nature of AI algorithms is inherently non-transparent, with factors leading to algorithm performance changes difficult to decipher and monitor by its end users or those developing benchmarks without explicit information provided by AI developers. These developers will likely need direct access to an institution's radiology information system to make such a continuous feedback loop possible, leading to data security concerns with potential exposure of personal health information.

Thus, developers, medical organizations, industry partners, and government agencies will have to work together to create new processes and guidelines. In the U.S., the FDA is working with stakeholders to draft a new regulatory process for AI devices spanning from the pre-market approval to post-market surveillance [18]. Previously, FDA clearance required that CAD software be "locked" prior to marketing with any changes to the algorithm requiring another FDA premarket review. As newer AI models have the ability to continuously learn and adapt to more available data with each new exam, the FDA proposes an adaptive total product lifecycle regulatory approach where manufacturers would be expected to monitor the AI algorithm clinical performance and incorporate a risk management approach after an initial premarket review [19]. Algorithm change components to be monitored and reported include data management changes (new training and testing data), re-training of machine learning architecture and parameters, changes in pre-determined assessment metrics, and software update procedures.

This type of medical AI device oversight will require adoption of standardized application programming interfaces (APIs) across diverse medical organization and government data networks. The real-world data requested by regulatory bodies such as the FDA for post-marketing surveillance of AI will also require large population-based registries that can help with continuous validation in the post-marketing setting. Moreover, large academic and private health systems will have to become willing partners in a new era of continuous monitoring by truly adapting into learning health care environments where AI-based imaging interpretation can continuously evolve and change. This latter enterprise will be challenging given a current environment of vendor-specific and proprietary data management tools for medical imaging without the ability for cross-communication. Yet, in the post-marketing period, regulators and manufacturers will have to work collaboratively to demonstrate that improved overall screening accuracy is maintained across different populations over time.

4. Summary

With multiple AI algorithms for mammography screening entering the clinical market and frequently evolving imaging technologies, external validation will be needed before and after clinical adoption. Medical organizations, AI developers, researchers, and government agencies must work together to help make evolving population-based imaging datasets representing diverse populations available for external validation in order to ensure clinical effectiveness and generalizability. As AI algorithms and screening modalities continue to undergo modifications in the post-marketing period, better standards are needed for continuous monitoring with greater transparency from AI algorithm developers. Moreover, better integration of biomedical informatics and data systems are needed for incorporating improvements in real-time and to avoid the missteps experienced with static traditional CAD software. These major paradigm shifts in validation and monitoring will be necessary before trust in "black box" algorithms for breast cancer screening are embraced by payers, health providers, and women alike. Without investment in novel validation pathways, generalized adoption of AI-enhanced breast cancer screening is unlikely to be successful.

Conflicts of interest

The authors declare no conflicts of interest directly related to this work. CL receives grant funding from GE Healthcare unrelated to this work and personal fees from the American College of Radiology for his role as Deputy Editor of the *Journal of the American College of Radiology*. All other authors declare no other conflicts of interest.

Funding Acknowledgement

CL and DB are funded by a grant from the National Cancer Institute (P01 CA154292). CL is also funded by grants from the American Cancer Society (MRSG-14-160-01-CPHPS), National Cancer Institute for CL (R37 CA240403), Safeway Foundation, and Earlier.org. NH is funded by Australia's National Breast Cancer Foundation (Cancer Research Leadership Fellowship). JE is also funded by grants from the National Cancer Institute (R01 CA225585; U01 CA231782; R01 CA200690; R01 CA201376). DB receives funding from the Patient Centered Outcomes Research Institute (PCS-1504-30370), the National Cancer Institute (R01CA207375, R01CA222090, UM1 CA221940, R01 CA240375) and the Agency for Health Care and Quality (K12 HS026369). The funding agencies had no role in the manuscript design or content selection; in the writing of the manuscript; and in the decision to submit the manuscript for publication.

References

[1] Aboutalib SS, Mohamed AA, Berg WA, Zuley ML, Sumkin JH, Wu S. Deep learning to distinguish recalled but benign mammography images in breast cancer screening. Clin Cancer Res 2018;24(23):5902–9.
[2] Rodriguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology 2019;290(2):305–14.
[3] Rodriguez-Ruiz A, Lang 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. https://doi.org/10.1093/jnci/djy222. Mar 5. pii: djy222.
[4] FDA News. FDA clears ScreenPoint Medical's AI system for reading mammograms. Available at: https://www.fdanews.com/articles/189314-fda-clears-screenpoint-medicals-ai-system-for-reading-mammograms. [Accessed 3 July 2019].
[5] Redberg RF, Dhruva SS. Moving from substantial equivalence to substantial improvement for 510(k) devices. J Am Med Assoc 2019 July 29. https://doi.org/10.1001/jama.2019.10191 [Epub ahead of print).
[6] Houssami N, Kirkpatrick-Jones G, Noguchi N, Lee CI. Artificial intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. Expert Rev Med Devices 2019;16(5):351–62.
[7] Iqbal SA, Wallace JD, Khoury MJ, Schully SD, Ioannidis JPA. Reproducible research practices and transparency across the biomedical literature. PLoS Biol 2018 Nov 20;16(11):e2006930. https://doi.org/10.1371/journal.pbio.2006930.
[8] Lee CI, Lehman CD. Digital breast tomosynthesis and the challenges of implementing an emerging breast cancer screening technology into clinical practice. J Am Coll Radiol 2013;10(12):913–7.
[9] Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. N Engl J Med 2007;356(14):1399–409.
[10] Fenton JJ, Lee CI, Xing G, Baldwin LM, Elmore JG. Computer-aided detection in

mammography: downstream effect on diagnostic testing, ductal carcinoma in situ treatment, and costs. JAMA Intern Med 2014;174(12):2032–4.
[11] Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL, Breast Cancer Surveillance Consortium. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med 2015;175(11):1828–37.
[12] Trister AD, Buist DSM, Lee CI. Will machine learning tip the balance in breast cancer screening? JAMA Oncol 2017;3(11):1463–4.
[13] Pattacini P, Nitrosi A, Giorgi Rossi P, et al. Digital mammography versus digital mammography plus tomosynthesis for breast cancer screening: the Reggio Emilia tomosynthesis randomized trial. Radiology 2018;288(2):375–85.
[14] Houssami N, Bernardi D, Caumo F, et al. Interval breast cancers in the 'screening with tomosynthesis or standard mammography' (STORM) population-based trial. Breast 2018;38:150–3.
[15] Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018;169(12):866–72.
[16] Wakabayashi D. Google and University of Chicago are sued over data sharing. *The New York Times*; June 27, 2019. Available at: https://www.nytimes.com/2019/06/26/technology/google-university-chicago-data-sharing-lawsuit.html.
[17] Guinney J, Saez-Rodriguez J. Alternative models for sharing confidential biomedical data. Nat Biotechnol 2018;36(5):391–2.
[18] Allen Jr B, Seltzer SE, Langlotz CP, et al. A road map for translational research on artificial intelligence in medical imaging: from the 2018 National Institutes of Health/RSNA/ACR/The Academy Workshop. J Am Coll Radiol 2019 Sep;16(9 Pt A):1179–89.
[19] U.S. FDA. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). Available at: https://www.fda.gov/media/122535/download. [Accessed 3 July 2019].