Editorial Commentary

External validation of AI algorithms in breast radiology: the last healthcare security checkpoint?

Teodoro Martin-Noguerol, Antonio Luna

Radiology Department, HTmédica, Clinica Las Nieves, Jaén, Spain

Correspondence to: Antonio Luna, MD, PhD. Radiology Department, HTmédica, Clinica Las Nieves, Carmelo Torres 2, 23007 Jaén, Spain.
Email: aluna70@htmedica.com.

Comment on: Salim M, Wåhlin E, Dembrower K, Azavedo E, Foukakis T, Liu Y, Smith K, Eklund M, Strand F. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA Oncol 2020;6:1581-8.

Submitted Dec 31, 2020. Accepted for publication Jan 26, 2021.
doi: 10.21037/qims-20-1409
View this article at: http://dx.doi.org/10.21037/qims-20-1409

Breast cancer is the most common cancer and the second most frequent cause of cancer death in women in the United States (1). In order to alleviate this major healthcare problem, breast cancer screening programs were launched in the eighties in developed countries. Although not free from controversy, population-based mammography screening has been proven to reduce breast cancer mortality between 20% to 40% in randomized controlled trials (1,2). However, there are still substantial differences between countries and regions in different aspects of these screening programs, such as starting age, mammography interval or number of readers.

In addition, the interpretation of mammographies by radiologists is qualitative and therefore, susceptible to reading errors. In general, the rate of false-negative results remains low (between 10–30%), but the rate of false positive results is not negligible, and requires subsequent additional imaging and a significant number of benign biopsies (3). False positive results have been linked to a number of factors related to patient characteristics, technical issues and radiologists' skills. Higher recall rates have been recorded in patients younger than 49-year-old, with higher density category, risk factors or previous mammographic findings (4). In addition, digital breast tomosynthesis, also known as 3D mammography, has widely demonstrated its superiority to conventional 2D mammography for breast lesion detection and characterization (5). In this direction, a recent meta-analysis suggests that breast tomosynthesis can improve cancer detection rates and decrease recall rates compared to traditional 2D mammography (6). Finally, double versus single reading, geographic location of the readers and particularly, radiologists' reading volume and experience have also been linked to differences in results and recall rates of screening programs (4,7,8). Clearly, there is room for improvement in several aspects of this screening approach.

Radiology is one of the most obvious gateways for the use of artificial intelligence (AI) in healthcare. The large amount of data present in medical imaging studies offers the ideal scenario for developing AI-based solutions. In this direction, deep learning (DL), a specific subtype of AI-based on artificial neural networks with representation learning, is expected to help radiologists to handle the growing number of radiological images and the resulting quantitative information (9). In this sense, breast tomosynthesis provides a greater number of images than 2D mammography, which involves an increase in radiologists' workload and decision times (10). Also, AI is expected to alleviate the general deficit of radiologist numbers, as in a growing number of countries the number of radiologists per 100,000 population is commonly below the desired standard, making this growing workload overwhelming and unsustainable. In the specific case of breast imaging, the shortage of subspecialized radiologists may result in a significant problem, bearing in mind the increasing demand for mammographies. (11).

The idea of using computer-aided detection (CAD) systems to improve both the radiologist's performance and productivity in the reading of screening mammography is not new (12). However, the results to date have been contradictory and without clear benefits (13). Over recent years, AI has been applied to CAD tools, first based on machine learning (ML) and later on DL algorithms, to assist radiologists' decisions in their routine evaluation of mammograms. With these tools, an increase of overall sensitivity is achieved; however, these computer solutions are not completely independent or automatic. These AI systems require the continuous supervision of radiologists to reach internal consensus or confirm suspicious findings, which is also a time-consuming task. Thus, there is a real need to increase the level of human independence from these AI-based tools, or at least, to improve their overall accuracy when ruling out malignancy (14).

More recently, convolutional neural networks (CNNs) have been applied to mammography CAD software with the aim of increasing the sensitivity and negative predictive values of this technique (15). Furthermore, these systems have been proposed as an option for independent reading. Along these lines, a recent report by Dembrower *et al.* concluded that the use of AI CAD systems for triage mammograms without radiologist supervision could potentially reduce radiologists' workload by more than half and also pre-emptively detect a substantial proportion of cancers, which would otherwise be diagnosed later (16). In the short-term, the most realistic scenario regarding the use of AI CAD tools for breast screening diagnosis will probably be focused on their use as an independent second reader in order to reduce the radiologist's workload (17).

Most of the AI CAD tools for mammography assessment are designed to detect features that suggest malignancy. For this reason, most algorithms have a similar profile and are highly suitable to compare their performance in terms of sensitivity, specificity and accuracy for breast lesion detection and characterization. In this regard, Salim *et al.* performed a retrospective external evaluation of three different commercially available AI algorithms as independent readers of a large set of population-based screening mammograms from Sweden (8,805 women including 739 women with cancer). Furthermore, they analyzed the screening results of the AI CAD systems as a supplementary read to the first reading by a radiologist (18). One of the algorithms demonstrated a significantly superior area under the curve (AUC) compared to the other two for cancer detection. Interestingly, at the estimated specificity (96.6%) of the radiologists, the best AI CAD system showed a similar or even higher performance level than any of the readers (81.9% for the algorithm versus 77.4% and 80.1% for first and second reader, respectively). Of note, the AI tools were blinded to prior mammograms and clinical information, which constituted a disadvantage compared to readers, as the authors recognized. Also, the combination of the first reader with the best AI CAD system achieved better sensitivity and specificity in cancer detection than the combination of both readers. However, when using the best algorithm as first reader followed by a human reader, the detection rate increased by 8% but with a dramatic increase of the abnormal calls (true positives plus false positives) by 77%. In summary, Salim *et al.* have demonstrated that there is a commercially available AI algorithm able to perform independent reading of screening mammograms with enough diagnostic performance to be considered as an independent reader in prospective clinical studies.

This work also discusses several possible causes of the differences in the performance between the three evaluated AI algorithms and between AI algorithms and radiologists. These causes range from the "well-known" black-box effect of most of AI tools to the different databases used for training algorithms, types of annotation and use of different AI technologies to build the algorithms, such as artificial neural networks, CNNs, augmented learning or generative adversarial networks (19). In this report, the size of the training dataset seems to be an important factor in the performance of the evaluated AI solutions, as the best algorithm used a significantly larger number of normal mammograms and cancer images for training than the other two. Another element usually considered crucial in the development of AI CAD systems is the diversity of the training dataset, as most of the algorithms are usually trained with specific ethnic, age or breast density groups, which may limit their performance when they are applied to other different populations. Surprisingly, in the series by Salim *et al.*, the best computer algorithm was trained using a completely different patient population and with different vendor equipment compared with the dataset and mammography device of the study. In the same vein, McKinney *et al.* performed an interesting study comparing the robustness of the same AI tool for breast cancer screening in datasets from USA and UK, providing evidence of how the system could be generalized from the UK to the USA population (17). They also found a significant reduction of false positive and negative results and higher AUC for the AI algorithm than the average radiologists.

All these data support the assertion that one of the main strengths of an AI CAD tool must be its robustness against heterogeneous populations.

Moreover, some of these AI tools are only trained with radiological breast images, without considering demographic and/or clinical data (i.e., age, prior breast cancer, palpable lump or hormone supply treatment), which may potentially decrease their overall accuracy for breast lesion detection and characterization compared to radiologists. For these reasons, the consideration of data different to imaging in the training dataset of AI algorithms is important in order to enhance their diagnostic performance and in an attempt to imitate the radiologist's workflow (20). Similarly, Schaffter *et al.* evaluated a recent challenge concerning the performance of 126 different AI algorithms in comparison to radiologists in the detection of cancer within a 12-month time frame using a dataset of more than 300,000 screening mammography studies from two different countries (21). None of the algorithms was able to beat radiologists' specificity at the radiologists' operating sensitivity. The results were similar when the AI CAD systems used only mammograms for the assessment or when prior imaging, clinical and demographic information were available. Subsequently, an ensemble model was created by the top performing teams but, again, the radiologists' specificity was superior. A significant improvement in the specificity for breast cancer detection was obtained only when the ensemble AI algorithm was followed by the radiologist assessment, carried out in a single-reader approach, compared with radiologists' performance. The authors also compared double reading strategies against this ensemble AI model, finding that consensus radiologist interpretation outperformed it. These results indicate that double-reading models are still better than a single-radiologist model assisted by AI (22). In the study by Salim *et al.*, the three AI CAD systems were blinded to previous imaging and clinical and demographic data, which may be a potential limitation. However, the trial conducted by Schaffter and colleagues highlighted the scant influence of having access to additional clinical information or imaging in the performance of current breast AI algorithms.

Due to all these potential sources of bias, there is a real need to perform robust external validation studies, like the one performed by Salim and colleagues, in order to compare the accuracy of different types of AI solutions from different vendors with the same cohort of patients. An essential step in the final development process of AI algorithms is to successfully overcome internal and external validation tests. Comparison between models from different vendors is the next and necessary step for ensuring an additional security checkpoint prior to their implementation in the clinical radiology arena.

In general, the practical application of AI-based tools in radiology, and in particular, their use in the assessment of mammography, raises several questions (23). The use of AI CAD systems combined with radiologists' readings increases the overall accuracy of mammograms for breast lesion detection and characterization. However, should these algorithms be launched prior to the radiologist's evaluation of images, as an assistant tool? Or, is it better to employ them after the radiologist's assessment, using them as a safety net? Should radiologists only have to check positive cases detected by AI algorithms? In this sense, economic and cost benefit issues also arise due to the possibility of saving money on a second radiologist for consensus reading or spending money purchasing a sophisticated AI CAD tool to be used as an independent reader. Finally, computational resources to deploy some of these algorithms may be difficult to find in a majority of breast imaging clinics or radiology departments (22).

For all the above reasons, we consider that a thoughtful analysis including external validation and comparison of different AI CAD systems in mammography assessment is mandatory prior to their introduction in radiological practice. Ethics and legal issues related with AI and radiology practice must also be carefully addressed before clinical trials begin, in order to preserve patient safety (24). All of these steps should constitute the last secure checkpoint for AI applications before being introduced into radiological clinical practice.

Acknowledgments

Funding: None.

Footnote

Provenance and Peer Review: This article was commissioned by the editorial office, *Quantitative Imaging in Medicine and Surgery*. The article did not undergo external peer review.

Conflicts of Interest: Both authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/qims-20-1409). Dr. AL has been lecturer for Siemens Healthineers, Philips, Canon and Bracco. The

other authors have no conflicts of interest to declare.

References

1. Helvie MA, Bevers TB. Screening mammography for average-risk women: the controversy and NCCN's position. J Natl Compr Canc Netw 2018;16:1398-404.
2. Biesheuvel C, Weigel S, Heindel W. Mammography screening: evidence, history and current practice in Germany and other European countries. Breast Care (Basel) 2011;6:104-9.
3. Nelson HD, O'Meara ES, Kerlikowske K, Balch S, Miglioretti D. Factors associated with rates of false-positive and false-negative results from digital mammography screening: an analysis of registry data. Ann Intern Med 2016;164:226-35.
4. Giess CS, Wang A, Ip IK, Lacson R, Pourjabbar S, Khorasani R. Patient, radiologist, and examination characteristics affecting screening mammography recall rates in a large academic practice. J Am Coll Radiol 2019;16:411-8.
5. Bernardi D, Macaskill P, Pellegrini M, Valentini M, Fantò C, Ostillio L, Tuttobene P, Luparia A, Houssami N. Breast cancer screening with tomosynthesis (3D mammography) with acquired or synthetic 2D mammography compared with 2D mammography alone (STORM-2): a population-based prospective study. Lancet Oncol 2016;17:1105-13.
6. Marinovich ML, Hunter KE, Macaskill P, Houssami N. Breast cancer screening using tomosynthesis or mammography: a meta-analysis of cancer detection and recall. J Natl Cancer Inst 2018;110:942-9.
7. Domingo L, Hofvind S, Hubbard RA, Román M, Benkeser D, Sala M, Castells X. Cross-national comparison of screening mammography accuracy measures in U.S., Norway, and Spain. Eur Radiol 2016;26:2520-8.
8. Euler-Chelpin MV, Lillholm M, Napolitano G, Vejborg I, Nielsen M, Lynge E. Screening mammography: benefit of double reading by breast density. Breast Cancer Res Treat 2018;171:767-76.
9. Martín Noguerol T, Paulano-Godino F, Martín-Valdivia MT, Menias CO, Luna A. Strengths, weaknesses, opportunities, and threats analysis of artificial intelligence and machine learning applications in radiology. J Am Coll Radiol 2019;16:1239-47.
10. Sechopoulos I, Teuwen J, Mann R. Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art. Semin Cancer Biol 2020. [Epub ahead of print]. doi: 10.1016/j.semcancer.2020.06.002.
11. Wing P, Langelier MH. Workforce shortages in breast imaging: impact on mammography utilization. AJR Am J Roentgenol 2009;192:370-8.
12. Vyborny CJ, Giger ML. Computer vision and artificial intelligence in mammography. AJR Am J Roentgenol 1994;162:699-708.
13. 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;175:1828-37.
14. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, Helbich TH, Chevalier M, Tan T, Mertelmeier T, Wallis MG, Andersson I, Zackrisson S, Mann RM, Sechopoulos I. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J Natl Cancer Inst 2019;111:916-22.
15. West E, Mutasa S, Zhu Z, Ha R. Global trend in artificial intelligence-based publications in radiology from 2000 to 2018. AJR Am J Roentgenol 2019;213:1204-6.
16. Dembrower K, Wåhlin E, Liu Y, Salim M, Smith K, Lindholm P, Eklund M, Strand F. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. Lancet Digit Health 2020;2:e468-74.
17. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GS, Darzi A, Etemadi M, Garcia-Vicente F, Gilbert FJ, Halling-Brown M, Hassabis D, Jansen S, Karthikesalingam A, Kelly CJ, King D, Ledsam JR, Melnick D, Mostofi H, Peng L, Reicher JJ, Romera-Paredes B, Sidebottom R, Suleyman M, Tse D, Young KC, De Fauw J, Shetty S. International evaluation of an AI system for breast cancer screening. Nature 2020;577:89-94.
18. Salim M, Wåhlin E, Dembrower K, Azavedo E, Foukakis

T, Liu Y, Smith K, Eklund M, Strand F. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA Oncol 2020;6:1581-8.
19. Fujita H. AI-based computer-aided diagnosis (AI-CAD): the latest review to read first. Radiol Phys Technol 2020;13:6-19.
20. Hamidinekoo A, Denton E, Rampun A, Honnor K, Zwiggelaar R. Deep learning in mammography and breast histology, an overview and future trends. Med Image Anal 2018;47:45-67.
21. Schaffter T, Buist DSM, Lee CI, Nikulin Y, Ribli D, Guan Y, Lotter W, Jie Z, Du H, Wang S, Feng J, Feng M, Kim HE, Albiol F, Albiol A, Morrell S, Wojna Z, Ahsen ME, Asif U, Jimeno Yepes A, Yohanandan S, Rabinovici-Cohen S, Yi D, Hoff B, Yu T, Chaibub Neto E, Rubin DL, Lindholm P, Margolies LR, McBride RB, Rothstein JH, Sieh W, Ben-Ari R, Harrer S, Trister A, Friend S, Norman T, Sahiner B, Strand F, Guinney J, Stolovitzky G; and the DM DREAM Consortium, Mackey L, Cahoon J, Shen L, Sohn JH, Trivedi H, Shen Y, Buturovic L, Pereira JC, Cardoso JS, Castro E, Kalleberg KT, Pelka O, Nedjar I, Geras KJ, Nensa F, Goan E, Koitka S, Caballero L, Cox DD, Krishnaswamy P, Pandey G, Friedrich CM, Perrin D, Fookes C, Shi B, Cardoso Negrie G, Kawczynski M, Cho K, Khoo CS, Lo JY, Sorensen AG, Jung H. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Netw Open 2020;3:e200265.
22. Mello-Thoms C. The path to implementation of artificial intelligence in screening mammography is not all that clear. JAMA Netw Open 2020;3:e200282.
23. Mezrich JL, Siegel EL. Legal ramifications of computer-aided detection in mammography. J Am Coll Radiol 2015;12:572-4.
24. Gastounioti A, Conant EF. Beyond the AJR: "External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms". AJR Am J Roentgenol 2020. [Epub ahead of print]. doi: 10.2214/AJR.20.25196.

Cite this article as: Martin-Noguerol T, Luna A. External validation of AI algorithms in breast radiology: the last healthcare security checkpoint? Quant Imaging Med Surg 2021;11(6):2888-2892. doi: 10.21037/qims-20-1409

