Using Time as a Measure of Impact for AI Systems: Implications in Breast Screening

William Hsu, PhD • *Anne C. Hoyt, MD*

William Hsu, PhD, is an associate professor of radiological sciences at the David Geffen School of Medicine, University of California, Los Angeles. His research interests are in data integration, machine learning, and imaging informatics with applications in cancer screening.

Anne C. Hoyt, MD, is a clinical professor of radiological sciences at the David Geffen School of Medicine, University of California, Los Angeles. She serves as section chief and medical director of breast imaging for UCLA Health. Dr Hoyt's research and clinical interests include digital breast tomosynthesis, breast ultrasonography, percutaneous biopsy challenges, and AI applications.

Time is a scarce resource in the modern fast-paced, high-pressure clinical breast imaging environment. Radiologist fatigue has been a long-standing concern but is being exacerbated by expanding practices, growing examination volume, and increasing complexity of imaging data that need to be interpreted (1). As radiology services expand, the number of examinations performed increases, and patient imaging data grow in complexity, and an immediate need exists for artificial intelligence (AI) and machine learning (ML) tools to facilitate the triaging, analysis, and interpretation of this data to streamline workflow, limit fatigue, and ultimately improve patient outcomes.

Breast screening is one area where reducing read time and improving reader performance would have a significant impact. While computer-aided detection (CAD) systems for mammography have had regulatory clearance and have been reimbursable by Medicare since 2002, their use has had mixed success. CAD systems applied to two-dimensional mammograms benefit less-experienced radiologists compared with expert breast imagers but also introduce automation bias, which may lead to missed cancers due to an overreliance on CAD (2). With the adoption of digital breast tomosynthesis (DBT), radiologists view a series of multiple thin images through the breast rather than a single two-dimensional image, improving cancer detection and reducing recall rates (3). The growing proportion of women undergoing screening using DBT and the exponentially higher number of images that need to be viewed to render a final assessment have increased interpretation time and compounded fatigue. AI and ML can aid the screening mammography workflow by reviewing and identifying clinically significant findings within DBT studies. To date, studies that examine the impact of an AI or ML system on reading time and reader performance of DBT examinations have been few and small.

In one of the first and largest studies of its kind for DBT, Conant et al (4) examined the concurrent use of AI on the performance of human readers in terms of reading time, sensitivity, specificity, and recall rate in identifying in situ and invasive cancers. The retrospective, nonclinical study focused on a single system called PowerLook Tomo Detection developed by a commercial vendor, iCAD, that processed the images and then displayed outlines of the detected lesions on each DBT image. Of the 130 studies that had suspicious or recalled findings, the suspicious findings comprised either soft-tissue densities and/or calcifications. Identified suspicious lesions had sizes ranging from 0.1 to 6 cm. The readers reviewed images from an enriched 260-case dataset that consisted of 65 (25%) malignancies. In addition to outlining lesions, the system generated a malignancy score between 0 and 100, where 100 means the system was highly confident the finding was malignant. The authors recruited 24 radiologists (11 of which were general radiologists who read mammograms), asking them to interpret all 260 cases with and without AI across two sessions.

Conant et al (4) provided compelling evidence of how AI and ML aid readers in interpreting DBT studies. Across all of the 24 readers, the use of an AI system was found to significantly reduce interpretation time by an average of 34.7 seconds (from 64.1 to 30.4 seconds), while improving case-level sensitivity by an average of 0.080, specificity by an average of 0.069, and reducing recall rates by an average of 0.072. Tables 3–5 in Conant et al (4) give a detailed picture of the impact that AI had on each reader. We performed a subset analysis of Table 3 in Conant et al (4) and examined five readers (20.8%) who had statistically significant differences in area under the receiver operating characteristic curve (AUC), following an approach described in Hanley and McNeil (5) based on a P value less than .05. Of these five readers (reader 1, 10, 13, 18, and 24), four (80%) were general radiologists who devoted less

From the Department of Radiological Sciences, David Geffen School of Medicine, University of California, 924 Westwood Blvd, Suite 420, Los Angeles, CA 90024.
Received June 21, 2019; accepted June 26. **Address correspondence to** W.H. (e-mail: *whsu@mednet.ucla.edu*).

See also the article by Conant et al in this issue. Conflicts of interest are listed at the end of this article.

than 75% of their time to breast imaging. They all had AUCs (0.631–0.740) that were below the average AUC (0.795) of all readers when interpreting studies without AI. With the concurrent use of AI in this subset of five readers, the improvement in AUC ranged from 0.099 to 0.150. A driving factor in the observed improvement in AUC is better case-level sensitivity across these five readers: increases in sensitivity with AI ranged from 0.138 to 0.323. However, differences in specificity among these five readers were mixed, with two readers having lower specificity (differences ranging from −0.026 to −0.085) and the other readers having higher specificity (differences ranging from 0.010 to 0.149). In a separate subanalysis, we examined the top five performing readers without AI (reader 7, 8, 9, 11, 22). Two (40%) of these readers had minimally reduced AUCs when using AI, while the others had slightly improved AUCs; none of these differences were statistically significant. All of the top readers saw a reduction in reading time with AI. Notably, two of the top readers had lower sensitivities with AI and one reader's sensitivity was unchanged. All five readers had improved specificity with AI.

The results of Conant et al (4) reaffirm prior studies such as Balleyguier et al (6) that have demonstrated a greater impact of AI in the performance of less experienced readers than in subspecialists. Twenty-one (87.5%) of the readers experienced an improvement in sensitivity, though notably, all three readers who had a decrease in sensitivity were considered subspecialists. On the other hand, all subspecialist readers had an improvement in specificity. Conant et al (4) acknowledge that readers in nonclinical studies may behave differently than they do in clinical practice. Factors that may influence the actual impact of AI in practice include: *(a)* the trust and confidence a reader has on the performance of the AI system in identifying all clinically significant lesions while minimizing false-positive marks; *(b)* whether interaction with the AI system is intuitive and efficient (eg, the number of clicks required to toggle the display of AI results); *(c)* a radiologist's confidence in his or her own interpretation; *(d)* how transparent the model is in explaining the rationale behind its predictions to the user (7); and *(e)* the type of training and experience the reader has in interpreting DBT (eg, trained to interpret DBT always with the aid of AI or ML or prior to the AI and ML era).

Finally, traditional metrics such as AUC, sensitivity, and specificity are informative in conveying the technical accuracy and reliability of an AI or ML algorithm. Conant et al (4) demonstrated the value of evaluating the impact of AI and ML algorithms based on time saved during interpretation. Time savings is of clear value. Nevertheless, the concept of time saved could and should be applied broadly across the entire radiology value chain (8). Maximizing interpretative accuracy and time savings leads to earlier patient diagnoses, one of many factors tied to improved patient outcome. Understanding the impact of an AI or ML algorithm on the entire radiology value chain provides clarity into which algorithms are most impactful and helps elucidate areas for additional AI intervention, such as patient experience or check-in procedures. Another link in this chain may rely on AI and ML algorithms to facilitate appropriate follow-up of patients, particularly individuals who are at higher risk of not keeping screening appointments. The former could be addressed by automatically contacting patients who have not scheduled or misscheduled a follow-up examination. These types of algorithms could be assessed based on their ability to improve patient follow-up compliance, minimize nonattendance rates, and maximize facility utilization; all of which provide time and cost savings. While time is of the essence overall, it should be noted that time saved is not necessarily the only goal for individual tasks. Some AI and ML algorithms may add time to complete specific tasks but ultimately improve downstream system performance. In summary, time is a valuable measure of the impact of an AI or ML algorithm that should be evaluated not only for individual tasks but also for the entire radiology value chain. In doing so, we can improve patient care and shed light onto as of yet undiscovered territories that are ripe for the application of AI.

Acknowledgments: The authors would like to thank Dieter Enzmann, MD, for constructive discussions on this topic.

Disclosures of Conflicts of Interest: W.H. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: institution receives grant from Siemens Medical Solutions; paid deputy editor for *Radiology: Artificial Intelligence*. Other relationships: disclosed no relevant relationships. **A.C.H.** disclosed no relevant relationships.

References

1. Stec N, Arje D, Moody AR, Krupinski EA, Tyrrell PN. A systematic review of fatigue in radiology: is it a problem? AJR Am J Roentgenol 2018;210(4):799–806.
2. Gao Y, Geras KJ, Lewin AA, Moy L. New frontiers: an update on computer-aided diagnosis for breast imaging in the age of artificial intelligence. AJR Am J Roentgenol 2019;212(2):300–307.
3. 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(9):942–949.
4. Conant E, Toledano A, Periaswamy S, et al. Improving accuracy and efficiency with concurrent use of intelligence for digital breast tomosynthesis. Radiol Artif Intell 2019;1(4):e180096.
5. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148(3):839–843.
6. Balleyguier C, Kinkel K, Fermanian J, et al. Computer-aided detection (CAD) in mammography: does it help the junior or the senior radiologist? Eur J Radiol 2005;54(1):90–96.
7. Hsu W, Elmore JG. Shining light into the black box of machine learning. J Natl Cancer Inst 2019 Jan 10 [Epub ahead of print] https://doi.org/10.1093/jnci/djy226.
8. Enzmann DR. Radiology's value chain. Radiology 2012;263(1):243–252.