Artificial intelligence to complement rather than replace radiologists in breast screening

See Articles page e507

In *The Lancet Digital Health*, Christian Leibig and colleagues[1] evaluated an artificial intelligence (AI) system designed to detect cancer on mammograms using a retrospective external validation test set of mammograms from 82 851 women attending breast cancer screening in Germany. Current practice in breast cancer screening is for one or two radiologists to examine each women's mammograms to determine whether there is sufficient suspicion of cancer to recall her for further tests. This study examined the potential for the same AI system to contribute in different ways to the testing pathway.

Retrospective test-set evaluation of new AI systems for the examination of breast screening mammograms is not new. Our previous systematic review identified seven studies reporting the test accuracy of AI as a standalone system in a retrospective test-accuracy study.[2] Three of these studies evaluated AI as a replacement for one or all radiologists.[3–5] Four of these studies evaluated AI as a triage tool to identify women whose mammograms show no indication of cancer so do not need radiologist review,[6–9] and one of the four studies evaluated AI after screening to detect missed cancers in women who were not recalled for further tests by the radiologist.[6] Overall, AI tests were not yet specific enough to replace a human radiologist, but separate studies indicate promise for using AI systems in the triage process and after radiologist review.

The Article by Leibig and colleagues[1] is new because it directly compares the performance of the same AI system in two different roles using the same dataset. The first role for the AI was directly replacing the radiologist, using an AI test threshold similar to that of the radiologist in the middle of the receiver operating characteristic (ROC) curve. In this role, the AI system was less accurate than the radiologist on the external validation set, with lower sensitivity and specificity. Specificity was 2·1 percentage points lower in the AI system, (93·4%, 95% CI 93·1–93·7, *vs* 91·3%, 91·1–91·5) a large difference in a population screening programme. The second role was called a decision-referral approach, which included the AI before and after radiologist review. Here, using the same retrospective-test set, the sensitivity and specificity of the decision-referral approach was projected to be higher than a single radiologist, by 2·6 percentage points for sensitivity and 1·0 percentage points for specificity. This result was driven by the accuracy of the AI system at the two ends of the ROC curve. Using AI as a triage test requires a low test threshold (the top-right part of the ROC curve). This low threshhold means the test has to be good at detecting cancer (so women with cancer are sent for radiological review), but does not have to be very good at identifying which women do not have cancer, because the radiologist will decide who is recalled for further tests. Using AI after radiological review and after screening to detect cancers missed by the radiologist needs to be very specific (the bottom-left part of the ROC curve). This specificity means that it must be very accurate in identifying those who do not have cancer, because at this stage the AI system is responsible for recalling women for further tests, therefore cannot cause many false-positive recalls to assessment. However, the system does not have to be sensitive to the point of detecting all cancers, because any cancers picked up at this stage represent extra detection in addition to those found through radiological review.

For each new AI test that is developed, the shape of the ROC curve can be informative for the potential role of the AI system in the testing pathway, with two caveats. First, we require sufficient data to make accurate estimates at each end of the ROC curve. Second, the method of fitting such a curve has to be an accurate representation of the raw data. The results from Leibig and colleagues[1] indicate that for this AI system, further evaluation for use in a decision-referral approach might be appropriate, but the accuracy of the underlying AI system would have to be improved before further consideration in a replacement role.

The evaluation by Leibig and colleagues[1] is reported more thoroughly than the existing literature, because they report accuracy by mammography system and by type of mammographic finding, which is important for understanding the generalisability of the results. Leibig and colleagues[1] also report results for the detection of ductal carcinoma in situ and invasive cancer seperately, which is informative in assessing the potential effect of

introducing AI on health outcomes. Future studies should follow this example of reporting accuracy of AI in these subgroups, and by other prognostic characteristics, such as grade and stage.

Retrospective test-set studies are insufficient for implementation of AI because they cannot accurately predict effect on test accuracy and outcomes (because of differential verification and inability to measure the effect on radiologist behaviour). However, these types of studies are informative in determining which AI systems are accurate enough for assessment in prospective studies, and in which role the AI system has most potential to contribute in the breast screening pathway. This study highlights that designing novel testing pathways to optimise the contribution of AI might have more potential than simply designing the AI system to directly replace humans.

ST-P and KF received funding from Public Health England to systematically review the literature on AI in breast screening. ST-P is supported by the National Institute for Health and Care Research (NIHR) through an NIHR Career Development fellowship (NIHR-CDF-2016-09-018) for the evaluation of screening tests. KS is supported by the NIHR through an NIHR Development and Skills Enhancement Fellowship (NIHR302371) for training in data science.

*Sian Taylor-Phillips, Karoline Freeman
s.taylor-phillips@warwick.ac.uk

Warwick Screening, The University of Warwick, CV4 7AL Coventry, UK

1 Leibig C, Brehmer M, Bunk S, Byng D, Pinker K, Umutlu L. Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. *Lancet Digit Health* 2022; **4:** e507–19.

2 Freeman K, Geppert J, Stinton C, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *BMJ* 2021; **374:** n1872.

3 McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020; **577:** 89–94.

4 Salim M, Wåhlin E, Dembrower K, et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. *JAMA Oncol* 2020; **6:** 1581–88.

5 Schaffter T, Buist DSM, Lee CI, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Net Open* 2020; **3:** e200265.

6 Dembrower K, Wåhlin E, Liu Y, et al. 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.

7 Raya-Povedano JL, Romero-Martín S, Elías-Cabot E, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. AI-based strategies to reduce workload in breast cancer screening with mammography and tomosynthesis: a retrospective evaluation. *Radiology* 2021; **300:** 57–65.

8 Balta C, Rodríguez-Ruiz A, Mieskes C, Karssemeijer N, Heywang-Köbrunner SH. Going from double to single reading for screening exams labeled as likely normal by AI: what is the impact? 15th International Workshop on Breast Imaging; May 22, 2020 (115130D).

9 Lång K, Dustler M, Dahlblom V, Åkesson A, Andersson I, Zackrisson S. Identifying normal mammograms in a large screening population using artificial intelligence. *Eur Radiol* 2021; **31:** 1687–92.