International evaluation of an AI system for breast cancer screening

Scott Mayer McKinney[1†*], Marcin T. Sieniek[1†], Varun Godbole[1†], Jonathan Godwin[2†], Natasha Antropova[2], Hutan Ashrafian[4], Trevor Back[2], Mary Chesus[2], Greg C. Corrado[1], Ara Darzi[5], Mozziyar Etemadi[6], Florencia Garcia-Vicente[6], Fiona J Gilbert[7], Mark Halling-Brown[8], Demis Hassabis[2], Sunny Jansen[9], Alan Karthikesalingam[3], Christopher J Kelly[3], Dominic King[3], Joseph R. Ledsam[2], David Melnick[6], Hormuz Mostofi[1], Bernardino Romera-Paredes[2], Lily Peng[1], Joshua Jay Reicher[10], Richard Sidebottom[11,12], Mustafa Suleyman[2], Daniel Tse[1], Kenneth C. Young[8], Jeffrey De Fauw[2‡], and Shravya Shetty[1‡*]

[1]Google Health, Palo Alto, CA, USA
[2]DeepMind, London, UK
[3]Google Health, London, UK
[4]Imperial College London, London, UK
[5]Institute of Global Health Innovation, Imperial College London, London, UK
[6]Northwestern Medicine, Chicago, IL, USA
[7]University of Cambridge School of Clinical Medicine, Cambridge, UK
[8]Royal Surrey County Hospital, Guildford, UK
[9]Verily Life Sciences, South San Francisco, CA, USA
[10]Stanford Health Care and Palo Alto Veterans Affairs, Palo Alto, CA, USA
[11]The Royal Marsden Hospital, London, UK
[12]Thirlestaine Breast Centre, Cheltenham, UK

[†]These authors contributed equally.
[‡]These authors supervised jointly.
*e-mail: sshetty@google.com; scottmayer@google.com

Screening mammography aims to identify breast cancer before symptoms appear, enabling earlier therapy for more treatable disease[1]. Despite the existence of screening programmes worldwide, interpretation of mammograms suffers from suboptimal rates of false positives and false negatives[2]. Here we present an AI system capable of surpassing expert readers in breast cancer prediction performance. To assess its performance in the clinical setting, we curated a large representative dataset from the United Kingdom (UK) and a large enriched dataset from the United States (US). We show an absolute reduction of 5.7%/1.2% (US/UK) in false positives and 9.4%/2.7% (US/UK) in false negatives. We show evidence of the system's ability to generalise from the UK sites to the US site. In an independently-conducted reader study, the AI system out-performed all six radiologists with an area under the receiver operating characteristic curve greater than the average radiologist by an absolute margin of 11.5%. By simulating the AI system's role in the double-reading process, we maintain noninferior performance while reducing the second reader's workload by 88%. This robust assessment of the AI system paves the way for prospective clinical trials to improve the accuracy and efficiency of breast cancer screening.

Breast cancer is the second leading cause of cancer death in women[3], but early detection and treatment can dramatically improve outcomes[1,4,5]. As a consequence, many developed nations have implemented large-scale mammography screening programmes. Major medical and governmental organisations recommend screening for all women starting between the ages of 40 and 50[6–8]. In the US and UK combined, over 42 million exams are performed each year[9,10].

Despite mammography's widespread adoption, interpretation of these images remains challenging. There is high variability in experts' cancer detection accuracy, and the performance of even the best clinicians leaves room for improvement[11,12]. False positives can lead to patient anxiety[13], unnecessary follow up, and invasive diagnostic procedures. Cancers missed at screening may not be identified until they are more advanced and less amenable to treatment[14].

Artificial intelligence (AI) may be uniquely poised to help. Recent studies have demonstrated AI's ability to meet or exceed the performance of human experts on several medical image analysis tasks[15–19]. As a shortage of mammography professionals threatens availability and adequacy of breast screening services around the world[20–23], the scalability of AI could improve access to high quality care for all.

Computer-aided detection (CAD) software for mammography was introduced in the 1990s, and multiple assistive tools have been approved for medical use[24]. Despite early promise[25,26], this generation of software failed to improve reader performance in real-world settings[12,27,28]. More recently, the field has seen a renaissance owing to the success of deep learning. A few studies have shown breast cancer prediction systems with standalone performance approaching that of human experts[29,30]. Still, existing work has several limitations. Most studies evaluate on small, enriched datasets with limited follow-up, and few have compared performance to readers in actual clinical practice, instead relying on lab-based simulations of the reading environment. To date, there has been little evidence of the ability of AI systems to translate between different

screening populations and settings without additional training data[31]. Critically, the pervasive use of follow-up intervals no longer than 12 months[29,30,32,33] means that more subtle cancers, not identified until the next screen, may be ignored.

In this study, we evaluate the performance of a new AI system for breast cancer prediction using two large, clinically-representative datasets from the UK and US. We compare the system's predictions to those made by readers in routine clinical practice and show performance better than individual radiologists. These observations are confirmed with an independently-conducted reader study. We further show how this system might be integrated into screening workflows, and provide evidence that the system can generalise across continents. Figure 1 depicts a high-level overview.

Screening programme datasets

A deep learning model for identifying breast cancer in screening mammograms was developed and evaluated using two large datasets from the UK and the US. We report results on test sets withheld from AI development.

The UK test set consisted of screening mammograms from 25,856 women collected between 2012 and 2015 at two screening centers in England, where women are screened every three years. It included 785 women who had a biopsy, and 414 women with cancer diagnosed within 39 months (3 years and 3 months) of imaging. This was a random sample of 10% of all women with screening mammograms at these sites during this time period. The UK cohort resembled the broader screening population in age and disease characteristics (Extended Data Table 1a).

The test set from the US, where women are screened every 1 or 2 years, consisted of screening mammograms from 3,097 women collected between 2001 and 2018 at one academic medical center. We included images from all 1,511 women biopsied during this time period and a random subset of women who never underwent biopsy (Methods). Among the women who received a biopsy, 686 were diagnosed with cancer within 27 months (2 years and 3 months) of imaging.

Breast cancer outcome was determined on the basis of multiple years of follow up (Figure 1). We chose the follow-up duration based on the screening interval in each dataset's country of origin. Following previous work[34], we augment each interval with a 3-month buffer to account for variability in scheduling and latency of follow up. Cases designated as cancer positive were accompanied by a biopsy-confirmed diagnosis within the follow-up period. Cases labeled as cancer negative had at least one follow-up non-cancer screen; cases without this follow up were excluded from the test set.

Retrospective clinical comparison

We used biopsy-confirmed breast cancer to evaluate predictions of the AI system as well as the original decisions made by radiologists in the course of clinical practice. Human performance was computed based on the clinician's decision to recall the patient for further diagnostic investigation. The receiver operating characteristic (ROC) curve of the AI system's cancer prediction is shown in Figure 2.

In the UK, each mammogram is interpreted by two readers. In cases of disagreement, an arbitration process is used, invoking a third opinion. These interpretations occur serially such that each reader has access to prior readers' opinions. The records of these decisions yield three human performance benchmarks for cancer prediction.

Compared to the first reader, the AI system demonstrated a statistically significant absolute specificity improvement of 1.2% (95% CI [0.29%, 2.1%]; $p = 0.0096$ for superiority) and an absolute sensitivity improvement of 2.7% (95% CI [-3%, 8.5%]; $p = 0.004$ for noninferiority at a prespecified 5% margin; Extended Data Table 2a).

Compared to the second reader, the AI system showed non-inferiority (at a 5% margin) for both specificity ($p < 0.001$) and sensitivity ($p = 0.02$). The AI system likewise showed non-inferiority (at a 5% margin) to the consensus judgment for specificity ($p < 0.001$) and sensitivity ($p = 0.0039$).

In the standard US screening protocol, each mammogram is interpreted by a single radiologist. We used the BI-RADS[35] score assigned to each case in the original screening context as a proxy for the human cancer prediction (Methods, Interpreting clinical reads). Compared to the typical reader, the AI system demonstrated statistically significant improvements in absolute specificity of 5.7% (95% CI [2.6%, 8.6%]; $p < 0.001$) and sensitivity of 9.4% (95% CI [4.5%, 13.9%]; $p < 0.001$; Extended Data Table 2a).

Generalisation across populations

To evaluate the AI system's ability to generalise across populations and screening settings, we trained the same architecture using only the UK dataset and applied it to the US test set (Figure 2b). Even without exposure to US training data, the AI system's ROC curve envelops the point indicating the average performance of US radiologists. Once again, the AI system showed superior specificity (+3.5%, $p = 0.0212$) and superior sensitivity (+8.1%, $p = 0.0006$; Extended Data Table 2b).

Reader study comparison

In a reader study conducted by an external clinical research organisation, six US board-certified radiologists compliant with Mammography Quality Standards Act (MQSA) requirements

interpreted 500 mammograms randomly sampled from the US test set. Where data were available, readers were equipped with contextual information typically available in the clinical setting, including patient age, breast cancer history, and prior screening mammograms.

Among the 500 cases selected for this study, 125 had biopsy-proven cancer within 27 months, 125 had a negative biopsy within 27 months, and 250 were not biopsied (Extended Data Table 3). These proportions were chosen to increase the difficulty of the screening task and increase statistical power; such enrichment is typical in observer studies[36].

Readers rated each case using the forced BI-RADS[35] scale. BI-RADS scores were compared to ground truth outcomes to fit an ROC curve for each reader. The scores of the AI system were treated in the same manner (Extended Data Table 2).

The AI system exceeded average radiologist performance by a significant margin (ΔAUC = +0.115, 95% CI: [0.055, 0.175], $p = 0.0002$). Similar results were observed when 1 year follow-up was used instead of 27 months (Figure 3c, Extended Data Figure 2).

In addition to producing case-level classification, the AI system was designed to highlight areas of suspicion for malignancy. Likewise, the readers in our study supplied rectangular region-of-interest (ROI) annotations surrounding concerning findings.

We used multi-localisation receiver operating characteristic (mLROC) analysis[37] to compare the ability of the readers and the AI system to identify malignant lesions within each case (Methods, Localisation analysis).

We summarised each mLROC plot by computing the partial area under the curve (pAUC) in the false positive fraction interval from 0 to 0.1[38] (Extended Data Figure 3). The AI system exceeded human performance by a significant margin (ΔpAUC = +0.0192, 95% CI: [0.0086, 0.0298], $p = 0.0004$).

Potential clinical applications

The AI system's classifications could be used to reduce the workload involved in the UK's double reading process while preserving the standard of care. We explored this scenario through simulation by omitting the second reader and any ensuing arbitration when the AI's decision agreed with the first reader. In these cases, the first reader's opinion was treated as final. In cases of disagreement, the second and consensus opinions were invoked as usual. This combination of human and machine displays performance equivalent to that of the traditional double reading process, while saving 88% of the second reader's effort (Extended Data Table 4a).

The AI system could also be used to provide automated, immediate feedback in the screening setting.

In order to identify normal cases with high confidence, we used a very low decision threshold. On the UK data, we achieved a negative predictive value (NPV) of 99.99% while retaining a specificity of 41.15%. Similarly, on the US data, we achieved a NPV of 99.90% while retaining a specificity of 34.79%. These data suggest that it may be feasible to pick out 35–41% of normal cases if we allow for one cancer in every 1,000–10,000 negative predictions (NPV 99.90–99.99% in US–UK). For comparison, consensus double reading in our UK dataset included one cancer in every 182 cases deemed normal.

To identify cancer cases with high confidence, we used a very high decision threshold. On the UK data, we achieved a positive predictive value (PPV) of 85.6% while retaining a sensitivity of 41.2%. Likewise, on the US data, we achieved a PPV of 82.4% while retaining a sensitivity of 29.8%. These data suggest that it may be feasible to rapidly prioritise 30–40% of cancer cases with approximately 5 of 6 follow ups leading to cancer diagnosis. By comparison, in our study only 22.8% of UK cases recalled by consensus double reading and 4.9% of US cases recalled by single reading were ultimately diagnosed with cancer.

Performance breakdown

Comparing the errors of the AI system with errors from clinical reads revealed many cases in which the AI system correctly identified cancer while the reader did not and vice versa (Supplementary Table 1). Most of the cases in which only the AI system identified cancer were invasive (Extended Data Table 5). On the other hand, cases in which only the reader identified cancer were split more evenly between in situ and invasive. Further breakdowns by invasive cancer size, grade, and molecular markers show no clear biases (Supplementary Table 2).

We also considered the disagreement between the AI system and the six radiologists that participated in the US reader study. Figure 4a shows a sample cancer case missed by all six radiologists, but correctly identified by the AI system. Figure 4b shows a sample cancer case caught by all six radiologists but missed by the AI system. While we were unable to determine clear patterns among these instances, the presence of such edge cases suggests potentially complementary roles for the AI system and human readers in reaching accurate conclusions.

We compared the performance of the 20 individual readers best represented in the UK clinical dataset with that of the AI system (Extended Data Table 7). The results of this analysis suggest that the aggregate comparison presented above is not unduly influenced by any particular readers. Breakdowns by cancer type, grade, and lesion size suggest no apparent difference in the distribution of cancers detected by the AI system and human readers (Extended Data Table 6a).

On the US test set, a breakdown by cancer type (Extended Data Table 6b) shows that the AI system's sensitivity advantage is concentrated on the identification of invasive cancers (e.g. invasive lobular/ductal carcinoma) rather than in situ cancer (e.g. ductal carcinoma in situ). A

breakdown by BI-RADS[35] breast density category shows that performance gains apply equally across the spectrum of breast tissue types represented in this data set (Extended Data Table 6c).

Discussion

In this study we present an AI system that outperforms radiologists on a clinically relevant breast cancer identification task. These results held on two large datasets representative of different country-specific screening populations and practices.

In the UK, the AI system showed specificity superior to that of the first reader. Sensitivity at the same operating point was noninferior. Consensus double reading has been shown to improve performance compared to single reading[39], and represents the current standard of care in the UK and many European countries[40]. Our system did not outperform this benchmark, but was statistically noninferior to the second reader and consensus opinion.

In the US, the AI system displayed specificity and sensitivity superior to that of radiologists practicing in an academic medical center. This trend was confirmed in an externally conducted reader study, which showed that the scores of the AI system stratify cases better than each of the six readers' BI-RADS ratings, the standard scale for mammography assessment in the US.

Remarkably, the human readers (both in the clinic and our reader study) had access to patient history and prior mammograms when making screening decisions. The US clinical readers may have also had access to breast tomosynthesis images. In contrast, the AI system only processed the most recent mammogram.

These comparisons are not without limitations. While the UK dataset mirrored the nationwide screening population in age and cancer prevalence (Extended Data Table 1a), the same cannot be said of the US data, which was drawn from a single screening centre and was enriched for cancer cases.

By chance, the vast majority of images used in this study were acquired on devices made by Hologic, Inc. Future research should assess the AI system's performance across a variety of manufacturers in a more systematic way.

In our reader study, all the radiologists were eligible to interpret screening mammograms in the US, but did not uniformly receive fellowship training in breast imaging. It is possible that a higher performance benchmark could have been obtained with more specialised readers[41].

To obtain high-quality ground-truth labels, we employed extended follow-up intervals chosen to encompass a subsequent screening round in each country. Although there is some precedent in clinical trials[34] and targeted cohort studies[42], this step is not usually taken when undertaking systematic evaluation of AI systems for breast cancer detection.

In retrospective datasets with shorter follow-up intervals, outcome labels tend to be skewed in favour of readers. Since they are gatekeepers for biopsy, asymptomatic cases will only receive a cancer diagnosis if a mammogram raised reader suspicion. A longer follow-up interval decouples the ground truth labels from reader opinions (Extended Data Figure 4) and includes cancers that may have been initially missed by human eyes.

The use of an extended interval makes cancer prediction a more challenging task. Cancers diagnosed years later may include new growths for which there could be no mammographic evidence in the original images. Consequently, the sensitivity values presented here are lower than what has been reported for 12 month intervals[2] (Extended Data Figure 5).

We present early evidence of the AI system's ability to generalise across populations and screening protocols. We retrained the system using exclusively UK data, and then measured performance on unseen US data. In this context, the system continued to outperform radiologists, albeit by a smaller margin. This suggests that in future clinical deployments, the system might offer strong baseline performance, but may benefit from fine-tuning with local data.

The utility of the AI system within clinical workflows remains to be determined. The specificity advantage exhibited by the AI system suggests it could help reduce recall rates and unnecessary biopsies. The improvement in sensitivity, exhibited in the US data, shows that the AI system may be capable of detecting cancers earlier than the standard of care. An analysis of the AI system's localisation performance suggests the early promise of using it to flag suspicious regions for review by experts. Notably, the additional cancers identified tended to be invasive rather than in situ disease.

Beyond augmenting reader performance, the technology described here may have a number of other clinical applications. Through simulation, we suggest how the system could obviate the need for double reading in 88% of UK screening cases, while maintaining similar accuracy to the standard protocol. We also explore how high-confidence operating points can be used to triage high-risk cases and dismiss low-risk cases. These analyses highlight the potential of this technology to deliver screening results in a sustainable manner despite workforce challenges in countries like the UK[43]. Prospective clinical studies will be required to understand the full extent to which this technology can benefit patient care.

References

1. Tabár, L. *et al.* Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. *Radiology* **260**, 658–663 (2011).

2. Lehman, C. D. *et al.* National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. *Radiology* **283**, 49–58 (2017).

3. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424 (2018).
4. Tonelli, M. *et al.* Recommendations on screening for breast cancer in average-risk women aged 40-74 years. *CMAJ* **183**, 1991–2001 (2011).
5. Marmot, M. G. *et al.* The benefits and harms of breast cancer screening: an independent review. *Br. J. Cancer* **108**, 2205–2240 (2013).
6. Lee, C. H. *et al.* Breast cancer screening with imaging: recommendations from the Society of Breast Imaging and the ACR on the use of mammography, breast MRI, breast ultrasound, and other technologies for the detection of clinically occult breast cancer. *J. Am. Coll. Radiol.* **7**, 18–27 (2010).
7. Oeffinger, K. C. *et al.* Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. *JAMA* **314**, 1599–1614 (2015).
8. Siu, A. L. & U.S. Preventive Services Task Force. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. *Ann. Intern. Med.* **164**, 279–296 (2016).
9. Center for Devices & Radiological Health. MQSA National Statistics. *U.S. Food and Drug Administration* (2019). Available at: http://www.fda.gov/radiation-emitting-products/mqsa-insights/mqsa-national-statistics. (Accessed: 16th July 2019)
10. Breast screening. *Cancer Research UK* (2017). Available at: https://www.cancerresearchuk.org/about-cancer/breast-cancer/screening/breast-screening. (Accessed: 26th July 2019)
11. Elmore, J. G. *et al.* Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology* **253**, 641–651 (2009).
12. Lehman, C. D. *et al.* Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. *JAMA Intern. Med.* **175**, 1828–1837 (2015).

13. Tosteson, A. N. A. *et al.* Consequences of False-Positive Screening Mammograms. *JAMA Internal Medicine* **174**, 954 (2014).
14. Houssami, N. & Hunter, K. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. *NPJ Breast Cancer* **3**, 12 (2017).
15. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402–2410 (2016).
16. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
17. De Fauw, J. *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
18. Ardila, D. *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).
19. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
20. Moran, S. & Warren-Forward, H. The Australian BreastScreen workforce: a snapshot. *Radiographer* **59**, 26–30 (2012).
21. Wing, P. & Langelier, M. H. Workforce shortages in breast imaging: impact on mammography utilization. *AJR Am. J. Roentgenol.* **192**, 370–378 (2009).
22. Rimmer, A. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ* **359**, j4683 (2017).
23. Nakajima, Y., Yamada, K., Imamura, K. & Kobayashi, K. Radiologist supply and workload: international comparison. *Radiation Medicine* **26**, 455–465 (2008).
24. Rao, V. M. *et al.* How widely is computer-aided detection used in screening and diagnostic mammography? *J. Am. Coll. Radiol.* **7**, 802–805 (2010).

25. Gilbert, F. J. *et al.* Single reading with computer-aided detection for screening mammography. *N. Engl. J. Med.* **359**, 1675–1684 (2008).
26. Giger, M. L., Chan, H.-P. & Boone, J. Anniversary Paper: History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM. *Medical Physics* **35**, 5799–5820 (2008).
27. Fenton, J. J. *et al.* Influence of computer-aided detection on performance of screening mammography. *N. Engl. J. Med.* **356**, 1399–1409 (2007).
28. Kohli, A. & Jha, S. Why CAD Failed in Mammography. *J. Am. Coll. Radiol.* **15**, 535–537 (2018).
29. Rodriguez-Ruiz, A. *et al.* Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *J. Natl. Cancer Inst.* (2019). doi:10.1093/jnci/djy222
30. Wu, N. *et al.* Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. (2019).
31. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.* **15**, e1002683 (2018).
32. Becker, A. S. *et al.* Deep Learning in Mammography: Diagnostic Accuracy of a Multipurpose Image Analysis Software in the Detection of Breast Cancer. *Invest. Radiol.* **52**, 434–440 (2017).
33. Ribli, D., Horváth, A., Unger, Z., Pollner, P. & Csabai, I. Detecting and classifying lesions in mammograms with Deep Learning. *Sci. Rep.* **8**, 4165 (2018).
34. Pisano, E. D. *et al.* Diagnostic performance of digital versus film mammography for breast-cancer screening. *N. Engl. J. Med.* **353**, 1773–1783 (2005).
35. American College of Radiology. *ACR BI-RADS Atlas: Breast Imaging Reporting and Data System ; Mammography, Ultrasound, Magnetic Resonance Imaging, Follow-up and*

Outcome Monitoring, Data Dictionary. (2013).

36. Gallas, B. D. *et al.* Evaluating imaging and computer-aided detection and diagnosis devices at the FDA. *Acad. Radiol.* **19**, 463–477 (2012).
37. Swensson, R. G. Unified measurement of observer performance in detecting and localizing target objects on images. *Med. Phys.* **23**, 1709–1725 (1996).
38. Samulski, M. *et al.* Using computer-aided detection in mammography as a decision support. *Eur. Radiol.* **20**, 2323–2330 (2010).
39. Brown, J., Bryan, S. & Warren, R. Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. *BMJ* **312**, 809–812 (1996).
40. Giordano, L. *et al.* Mammographic screening programmes in Europe: organization, coverage and participation. *J. Med. Screen.* **19 Suppl 1**, 72–82 (2012).
41. Sickles, E. A., Wolverton, D. E. & Dee, K. E. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. *Radiology* **224**, 861–869 (2002).
42. Ikeda, D. M., Birdwell, R. L., O'Shaughnessy, K. F., Sickles, E. A. & Brenner, R. J. Computer-aided detection output on 172 subtle findings on normal mammograms previously obtained in women with breast cancer detected at follow-up screening mammography. *Radiology* **230**, 811–819 (2004).
43. The breast imaging and diagnostic workforce in the United Kingdom. *The Royal College of Radiologists* (2016). Available at: https://www.rcr.ac.uk/publication/breast-imaging-and-diagnostic-workforce-united-kingdom. (Accessed: 22nd July 2019)

Figures

Figure 1. Development of an AI system to detect cancer in screening mammograms. Datasets representative of the UK and US breast cancer screening populations were curated from three screening centers in the UK and one center in the US. Outcomes were derived from the biopsy record and longitudinal follow up. An AI system was trained to identify the presence of breast cancer from a screening mammogram; it was evaluated in three primary ways. AI predictions were compared with the historical decisions made in clinical practice. To evaluate the generalisability across populations, a version of the AI system was developed using only the UK data and retested on the US data. Finally, the AI system was compared with six independent radiologists using a subset of the US test set.

Figure 2. Breast cancer prediction performance.
a. Receiver operating characteristic (ROC) curve of the AI system on the UK screening data. The area under the curve (AUC) is 0.889 (95% CI [0.871, 0.907]; n = 25,856 patients). Also shown are the (sensitivity, specificity) pairs of the human decisions made in clinical practice. Cases were considered positive if they received a biopsy-confirmed cancer diagnosis within 39 months (3 years and 3 months) from the time of screening. The consensus decision represents the standard of care in the UK, and will involve input from between 2 and 3 expert readers. The inset shows an enhancement of the gray shaded region. AI system operating points were selected on a separate validation dataset: point (i) was intended to match the sensitivity and exceed the specificity of the first reader; points (ii) and (iii) were selected to attain non-inferiority for both the sensitivity and specificity of the second reader and consensus opinion, respectively. **b.** ROC curve of the AI system on the US screening data. When trained on both datasets, the AUC is 0.8107 (95% CI [0.791, 0.831]; n = 3,097 patients). When trained only on the UK dataset (dotted curve), the AUC is 0.757 (95% CI [0.732, 780]). Also shown are the sensitivity and specificity achieved by radiologists in clinical practice using BI-RADS[35]. Cases were considered positive if they received a biopsy-confirmed cancer diagnosis within 27 months (2 years and 3 months) from the time of screening. AI system operating points were chosen to exceed the average reader's sensitivity and specificity. Negative cases were upweighted to account for the sampling protocol (Methods, Inverse probability weighting). Extended Data Figure 1 shows an unweighted analysis. See Extended Data Table 2a for statistical comparisons of sensitivity and specificity.

Figure 3. Breast cancer prediction performance compared to six independent readers.
a. Six readers rated each case (n = 465) using the 6-point BI-RADS scale. A fitted ROC curve for each of the readers is compared to the ROC curve of the AI system (Methods, Statistical analysis). For reference, a nonparametric ROC curve is presented in tandem. Cases were considered positive (n = 113) if they received a pathology-confirmed cancer diagnosis within 27 months (2 years and 3 months) from the time of screening. Note that this sample of cases was enriched for patients that had received a negative biopsy result (n = 119), making this a more challenging population for screening. The mean reader AUC was 0.625 (s.d. 0.032), while the AI system's AUC was 0.740 (95% CI: [0.696, 0.794]). The AI system exceeded human

performance by a significant margin (Δ = +0.115, 95% CI: [0.055, 0.175], p = 0.0002, two-sided ORH method). For results using a 12-month interval, see Extended Data Figure 2.
b. Pooled results from all six readers from (**a**).
c. Pooled results (n = 408) from all six readers using a 12-month interval for cancer definition. Cases were considered positive (n = 56) if they received a pathology-confirmed cancer diagnosis within 1 year (Extended Data Table 3).

Figure 4. Discrepancies between the AI system and human readers.
a. A sample cancer case missed by all six readers in the US reader study, but correctly identified by the AI system. The images show two views of a small, irregular mass with associated microcalcifications in the lower inner right breast.
b. A sample cancer case caught by all six readers in the US reader study, but missed by the AI system. The images show two views of a dense mass in the lower inner right breast. (left, mediolateral oblique; right, craniocaudal)

Methods

Ethical approval. Use of the UK dataset for research collaborations by both commercial and non-commercial organisations received ethical approval (REC reference 14/SC/0258). The US data was fully de-identified and released only after an Institutional Review Board approval (STU00206925).

UK dataset. The UK dataset was collected from three breast screening sites in the United Kingdom National Health Service Breast Screening Programme (NHSBSP). The NHSBSP invites women aged between 50 and 70 who are registered with a general practitioner (GP) for mammographic screening every 3 years. Women who are not registered with a GP, or who are older than 70, can self-refer to the screening programme. In the UK, the screening programme uses double reading: each mammogram is read by two radiologists, who are asked to decide whether to recall the woman for additional followup. When there is disagreement, an arbitration process takes place.

The data was initially compiled by OPTIMAM, a Cancer Research UK effort, between the years of 2010 and 2018 from St. George's Hospital (London, UK), Jarvis Breast Centre (Guildford, UK) and Addenbrooke's Hospital (Cambridge, UK). The collected data included screening and follow-up mammograms (comprising mediolateral oblique "MLO" and craniocaudal "CC" views of the left and right breast), all radiologist opinions (including the arbitration result, if applicable) and metadata associated with follow-up treatment.

The mammograms and associated metadata of 137,291 women were considered for inclusion in the study. Of these, 123,964 had both screening images and uncorrupted metadata. Exams that were recalled for reasons other than radiographic evidence of malignancy, or episodes that were not part of routine screening were excluded. In total, 121,850 women had at least one eligible exam. Women who were aged below 47 at the time of the screen were excluded from validation and test sets, leaving 121,455 women. Finally, women for whom there was insufficient follow up for any scan were excluded from validation and test. This last step resulted in the exclusion of 5,990 of 31,766 test set cases (19%). See Supplementary Figure 1.

The test set is a random sample of 10% of all women screened at two sites, St. George's and Jarvis, between the years 2012 and 2015. Insufficient data was provided to apply the sampling procedure to the third site. In assembling the test set, we randomly selected a single eligible screening mammogram from each woman's record. For women with a positive biopsy, eligible mammograms were those conducted in the 39 months (3 years and 3 months) prior to the biopsy date. For women that never had a positive biopsy, eligible mammograms were those with a non-suspicious mammogram at least 21 months later.

The final test set consisted of 25,856 women (see Supplementary Figure 1). When compared to the UK national breast cancer screening service we see a very similar cancer prevalence, age and cancer type distribution (see Extended Data Table 1a). Digital mammograms were acquired

predominantly on devices manufactured by Hologic, Inc. (95%), followed by General Electric (4%) and Siemens (1%).

US dataset. The US dataset was collected from Northwestern Memorial Hospital (Chicago, IL) between the years of 2001 and 2018. In the US, each screening mammogram is typically read by a single radiologist, and screens are conducted annually or biannually. The breast radiologists at this hospital are fellowship-trained and only interpret breast imaging studies. Their experience levels ranged from 1-30 years. The American College of Radiology (ACR) recommends that women start routine screening at the age of 40, while other organizations including the US Preventive Services Task Force (USPSTF) recommend initiation at 50 for women with average breast cancer risk[6–8].

The US dataset included records from all women that underwent a breast biopsy between 2001 and 2018. It also included a random sample of approximately 5% of all women who participated in screening, but were never biopsied. This heuristic was employed in order to capture all cancer cases (to enhance statistical power) and to curate a rich set of benign findings on which to train and test the AI system.

Supplementary Figure 2 distills the data processing steps involved in constructing the dataset. Among women with a completed mammogram order, we collected the records from all women with a pathology report containing the term “breast”. Among those that lacked such a pathology report, women whose records bore an International Classification of Diseases (ICD) code indicative of breast cancer were excluded. Approximately 5% of this unbiopsied negative population was sampled. After deidentification and transfer, women were excluded if their metadata was either unavailable or corrupted. The women in the dataset were split randomly among train (55%), validation (15%) and test (30%). For testing, a single case was chosen for each woman following a similar procedure as in the UK dataset. In women who underwent biopsy, we randomly chose a case from the 27 months preceding the date of biopsy. For women who did not undergo biopsy, one screening mammogram was randomly chosen from among those with a follow up event at least 21 months later.

Cases were considered complete if they possessed the four standard screening views (mediolateral oblique “MLO” and craniocaudal “CC” views of the left and right breast) acquired for screening intent. Here too, the vast majority of the studies were acquired using Hologic (including Lorad-branded) devices (99%) while manufacturers Siemens and General Electric together constituted less than 1% of studies.

The radiology reports associated with cases in the test set were used to flag and exclude cases in the test set which depicted breast implants or were recalled for technical reasons. To compare the AI system against the clinical reads performed at this site, we employed clinicians to manually extract BI-RADS scores from the original radiology reports. There were some cases for which the original radiology report could not be located, even if a subsequent cancer diagnosis was biopsy-confirmed. This might have happened, for example, if the screening case

was imported from an outside institution. Such cases were excluded from the clinical reader comparison.

Inverse probability weighting. The US test set includes images from all biopsied women, but only a random subset of women who never underwent biopsy. This enrichment allowed us to accrue more positives in light of the low baseline prevalence of breast cancer, but led to underrepresentation of normal cases. We accounted for this sampling process by using inverse probability weighting to obtain unbiased estimates of human and AI system performance in the natural screening population[44,45].

We acquired images from 7,522 of the 143,238 women who underwent mammography screening but had no cancer diagnosis or biopsy record. Accordingly, we upweighted cases from women who never underwent biopsy by a factor of 19.04. Further sampling occurred when selecting one case per patient: to enrich for difficult cases, we preferentially chose cases from the timeframe preceding a biopsy, if one occurred. Although this sampling increases the diversity of benign findings, it again shifts the distribution from what would be observed in a typical screening interval. To better reflect the prevalence resulting when negative cases are randomly selected, we estimated additional factors by Monte Carlo simulation. When choosing one case per patient with our preferential sampling mechanism, we got 872 cases that were biopsied within 27 months, and 1,662 cases that were not (Supplementary Figure 2). However, 100 trials of pure random sampling yielded on average 557.54 and 2,056.46 cases, respectively. Accordingly, cases associated with negative biopsies were down-weighted by 557.54 / 872 = 0.64. Cases that were not biopsied were up-weighted by another 2,056.46 / 1,662 = 1.24, leading to a final weight of 19.04 x 1.24 = 23.61.Cancer positive cases carried a weight of 1.0. The final sample weights were used in sensitivity, specificity and ROC calculations.

Histopathological outcomes. In the UK dataset, benign and malignant classifications, given directly in the metadata, followed NHSBSP definitions[46]. To derive the outcomes labels for the US dataset, pathology reports were reviewed by US board-certified pathologists and categorized according to the findings they contained. An effort was made to make this categorization consistent with UK definitions. Malignant pathologies included ductal carcinoma in situ, microinvasive carcinoma, invasive ductal carcinoma, invasive lobular carcinoma, special type invasive carcinoma (including tubular, mucinous, and cribriform carcinomas), intraductal papillary carcinoma, non-primary breast cancers (including lymphoma and phyllodes), and inflammatory carcinoma. Any woman who received a biopsy resulting in any of these malignant pathologies was considered to have a diagnosis of cancer.

Benign pathologies included lobular carcinoma in situ, radial scar, columnar cell changes, atypical lobular hyperplasia, atypical ductal hyperplasia, cyst, sclerosing adenosis, fibroadenoma, papilloma, periductal mastitis, and usual ductal hyperplasia. None of these findings qualified a woman for a cancer diagnosis.

Interpreting clinical reads. In the UK screening setting, readers categorise mammograms from asymptomatic women as normal or abnormal, with a third option for technical recall due to inadequate image quality. An abnormal result at the conclusion of the double reading process results in further diagnostic workup. We treat mammograms deemed abnormal as a prediction of malignancy. Cases in which the consensus judgment recalled the patient for technical reasons were excluded from analysis, as the images were presumed incomplete or unreliable. Cases in which any single reader recommended technical recall were excluded from the corresponding reader comparison.

In the US screening setting, radiologists attach a BI-RADS[35] score to each mammogram. A score of 0 is deemed "incomplete", and will be later refined based on follow up imaging or repeat mammography to address technical issues. For computation of sensitivity and specificity, we dichotomized the BI-RADS assessments in line with previous work[34]. Scores of 0, 4 and 5 were treated as positive predictions if recall was not based on technical grounds and the recommendation was based on mammographic findings, not solely patient symptoms. Cases of technical recall were excluded from analysis, as the images were presumed incomplete or unreliable. BI-RADS scores were manually extracted from the free-text radiology reports. Cases for which the BI-RADS score was unavailable were excluded from the reader comparison.

In both datasets, the original readers had access to contextual information normally available in clinical practice. This includes the patient's family history of cancer, prior screening and diagnostic imaging, and radiology or pathology notes from past examinations. In contrast, only the patient's age was made available to the AI system.

Overview of the AI system. The AI system consisted of an ensemble of three deep learning models, each operating on a different level of analysis (individual lesions, individual breasts, and the full case). Each model produces a cancer risk score between 0 and 1 for the entire mammography case. The final prediction of the system was the mean of the predictions from the three independent models. A detailed description of the AI system is available in Supplementary Methods and Supplementary Figure 3.

Operating point selection. The AI system natively produces a continuous score representing the likelihood that cancer is present. To support comparisons with the predictions of human readers, we thresholded this score to produce analogous binary screening decisions. For each clinical benchmark, we used the validation set to choose a distinct operating point; this amounts to a score threshold separating positive and negative decisions. To better simulate prospective deployment, the test sets were never used in selecting operating points.

The UK dataset contains three clinical benchmarks--the first reader, second reader, and consensus. This last decision is the outcome of the double reading process and represents the standard of care in the UK. For the first reader, we chose an operating point aimed at demonstrating statistical superiority in specificity and non-inferiority for sensitivity. For the second reader and consensus reader, we chose an operating point aimed at demonstrating statistical non-inferiority to the human reader for both sensitivity and specificity.

The US dataset contains a single operating point for comparison, corresponding to the single radiologist using the BI-RADS rubric for evaluation. In this case, we used the validation set to choose an operating point aimed at achieving superiority on both sensitivity and specificity.

Reader study. For the reader study, 6 US board-certified radiologists interpreted a sample of 500 cases from 500 women in the test set. All radiologists were compliant with MQSA requirements for interpreting mammography and had an average of 10 years of clinical experience (Extended Data Table 7). Two of them were fellowship-trained in breast imaging. The sample of cases was stratified to contain 50% normal cases, 25% biopsy negative cases and 25% of biopsy positive cases. A detailed description of the reader study case composition can be found in Extended Data Table 3. Readers were not informed of the enrichment levels in the dataset.

Readers recorded their assessments on a 21CFR11-compliant electronic case report form within the Ambra Health (New York, NY) viewer v3.18.7.0R. They interpreted the images using 5MP MSQA-compliant displays. Each reader interpreted the cases in a unique randomized order.

For each study, readers were asked to first report a BI-RADS[35] 5th edition score among 0, 1, and 2, as if they were interpreting the screening mammogram in routine practice. They were then asked to render a forced diagnostic BI-RADS score using one of the following values: 1, 2, 3, 4A, 4B, 4C or 5. Readers also gave a finer-grained score between 0 and 100 indicating their suspicion that the case contains a malignancy.

In addition to the 4 standard mammographic screening mages, clinical context was provided to better simulate the screening setting. Readers were presented with the preamble of the deidentified radiology report produced by the radiologist originally interpreting the study. This contained information such as the patient's age and family history of cancer. The information was manually reviewed to ensure that no impression or findings were included.

Where possible (in 43% of cases), prior imaging was made available to the readers. Readers could review up to four sets of prior screening exams, acquired between 1 and 4 years earlier, accompanied by deidentified radiologist reports. If prior imaging was available, the study was read twice by each reader--first without the prior information and immediately after, with prior information present. The system ensured that readers could not update their initial assessment after the prior information was presented. For cases where prior exams were available, reader assessment after having reviewed priors was used for the analysis.

Cases for which at least half of the readers indicated image quality concerns were excluded from analysis. Cases in which breast implants were noted were excluded as well. The final analysis was performed on the remaining 465 cases.

Localisation analysis. For this purpose, we considered all screening exams from the reader study for which cancer developed within 12 months. See Extended Data Table 3 for a detailed description of how the dataset was constructed. To collect ground truth localisations, two board-certified radiologists inspected each case, using follow-up data to identify the location of malignant lesions. Instances of disagreement were resolved by one radiologist with fellowship training in breast imaging. To identify the precise location of the cancerous tissue, radiologists consulted subsequent diagnostic mammograms, radiology reports, biopsy notes, pathology reports, and post-biopsy mammograms. Rectangular bounding boxes were drawn around the locations of subsequent positive biopsies in all views in which the finding was visible. In cases where no mammographic finding was visible, the location where the lesion later appeared was highlighted. Of the 56 cancers considered for analysis, location information could be obtained with confidence in 53 cases. Three cases were excluded due to ambiguity in the index examination and the absence of follow-up images. On average, there were 2.018 ground truth regions per cancer-positive case.

In the reader study, readers supplied rectangular region-of-interest (ROI) annotations surrounding suspicious findings in all cases they rated BI-RADS 3 or higher. A limit of 6 ROIs per case was enforced. On average, the readers supplied 2.04 annotations per suspicious case. In addition to an overall cancer likelihood score, the AI system emits a ranked list of rectangular bounding boxes for each case. To conduct a fair comparison, we allowed the AI system only its top two bounding boxes to match the number of ROIs produced by the readers.

To compare the localisation performance of the AI system with that of the readers, we used a method inspired by location receiver operating characteristic (LROC) analysis[37]. LROC analysis differs from traditional ROC analysis in that the ordinate is a sensitivity measure that factors in localisation accuracy. Although LROC analysis traditionally involves a single finding per case[37,47], we permitted multiple unranked findings to match the format of our data. We use the term multi-localization ROC analysis (mLROC) to describe our approach. For each threshold, a cancer case was considered a true positive if its casewide score exceeded this threshold and at least one culprit area was correctly localised in any of the four mammogram views. Correct localisation required an intersection-over-union (IoU) of 0.1 with the ground truth ROI. False positives were defined as usual.

CAD systems are often evaluated on the basis of whether the center of their marking falls within the boundary of a ground truth annotation[48]. This is potentially problematic since it doesn't properly penalize predicted bounding boxes that are so large as to be nonspecific, but whose center nevertheless happens to fall within the target region. Similarly, large ground truth annotations associated with diffuse findings might be overly generous to the CAD system. We prefer the IoU metric because it balances these considerations. We chose a threshold of 0.1 to account for the fact that indistinct margins on mammography findings lead to region-of-interest annotations of vastly different sizes depending on subjective factors of the annotator. See Supplementary Figure 4. Similar work in 3D chest computed tomography [18] used *any* pixel overlap to qualify for correct localisation. Likewise, an FDA-approved software device for wrist

fracture detection reports statistics in which true positives require at least one pixel of overlap [49]. An IoU value of 0.1 is strict by these standards.

Statistical analysis. To evaluate standalone AI system performance, the area under the ROC curve was estimated using the normalized Wilcoxon (Mann-Whitney) *U* statistic[50]. This is the standard nonparametric method employed by most modern software libraries. For the UK dataset, nonparametric confidence intervals on the AUC were computed with DeLong's method [51,52]. For the US dataset, in which each sample carried a scalar weight, the bootstrap was used with 1000 replications.

On both datasets, we compared the sensitivity and specificity of the readers with that of a thresholded score from the AI system. For the UK dataset, we knew the identity of each reader, so statistics were adjusted for the clustered nature of the data using Obuchowski's method for paired binomial proportions[53,54]. Confidence intervals on the difference are Wald intervals [55] and a Wald test was used for noninferiority [56]. Both used the Obuchowski variance estimate.

For the US dataset, in which each sample carried a scalar inverse probability weight[45], we used resampling methods [57] to compare the AI system's sensitivity and specificity with that of the pool of radiologists. Confidence intervals on the difference were generated with the bootstrap method with 1000 replications. A *p*-value on the difference was generated through the use of a permutation test [58]. In each of 10000 trials, the reader and AI system scores were randomly interchanged for each case, yielding a reader-AI system difference sampled from the null distribution. A two-sided *p*-value was computed by comparing the observed statistic to the empirical quantiles of the randomization distribution.

In the reader study, each reader graded each case using a forced BI-RADS protocol (a score of 0 was not permitted), and the resulting values were treated as a 6-point index of suspicion for malignancy. Scores of 1 and 2 were collapsed into the lowest category of suspicion; scores 3, 4a, 4b, 4c, and 5 were treated independently as increasing levels of suspicion. Because none of the BI-RADS operating points reached the high sensitivity regime (see Figure 3), to avoid bias from nonparametric analysis [59] we fit parametric ROC curves to the data using the proper binormal model [60]. This issue was not alleviated by using the readers' malignancy suspicion ratings, which showed very strong correspondence with the BI-RADS scores (Supplementary Figure 5). Since BI-RADS is used in actual screening practice, we elected to focus on these scores for their superior clinical relevance. In a similar fashion, we fit a parametric ROC curve to discretized AI system scores on the same data.

The performance of the AI system was compared to that of the panel of radiologists using methods for the analysis of multi-reader multi-case (MRMC) studies standard in the radiology community [61]. More specifically, we compared the AUC-ROC and pAUC-mLROC for the AI system to that of the average radiologist using the ORH procedure, which was proposed in [62] and updated in [63]. Originally formulated for the comparison of multiple imaging modalities, this analysis has been adapted to the setting in which the population of radiologists operate on a single modality and interest lies in comparing their performance to that of a standalone

algorithm [61]. The jackknife method was used to estimate the covariance terms in the model. The *p*-value and confidence interval computation was conducted in Python using the numpy and scipy packages and benchmarked against a reference implementation in the RJafroc library for the R computing language [64].

Our primary comparisons numbered seven in total: sensitivity and specificity for the UK first reader; sensitivity and specificity for the US clinical radiologist; sensitivity and specificity for the US clinical radiologist using a model trained using only UK data; and the AUC-ROC in the reader study. For comparisons with the clinical reads, the choice of superiority or non-inferiority was based on what seemed attainable from simulations conducted on the validation set. For non-inferiority comparisons, a 5% absolute margin was prespecified before inspecting the test set. We employed a statistical significance threshold of 0.05. All seven *p*-values survived correction for multiple comparisons using the Holm-Bonferroni method[65].

Code availability. The code used for training the models has a large number of dependencies on internal tooling, infrastructure and hardware, and its release is therefore not feasible. However, all experiments and implementation details are described in sufficient detail in the Supplementary Methods section to allow independent replication with non-proprietary libraries. Several major components of our work are available in open source repositories: Tensorflow: https://www.tensorflow.org; Tensorflow Object Detection API: https://github.com/tensorflow/models/tree/master/research/object_detection

Data availability. The dataset from Northwestern Medicine was used under license for the current study, and is not publicly available. Applications for access to the OPTIMAM database can be made at https://medphys.royalsurrey.nhs.uk/omidb/getting-access/.

Methods References

44. Pinsky, P. F. & Gallas, B. Enriched designs for assessing discriminatory performance--analysis of bias and variance. *Stat. Med.* **31**, 501–515 (2012).
45. Mansournia, M. A. & Altman, D. G. Inverse probability weighting. *BMJ* **352**, i189 (2016).
46. Pathology reporting of breast disease in surgical excision specimens incorporating the dataset for histological reporting of breast cancer. *Royal College of Pathologists* (2016). Available at: https://www.evidence.nhs.uk/document?id=1777849. (Accessed: 22nd July 2019)
47. Chakraborty, D. P. & Yoon, H.-J. Operating characteristics predicted by models for diagnostic tasks involving lesion localization. *Medical physics* **35**, 435–445 (2008).
48. Ellis, R. L., Meade, A. A., Mathiason, M. A., Willison, K. M. & Logan-Young, W. Evaluation of computer-aided detection systems in the detection of small invasive breast carcinoma. *Radiology* **245**, 88–94 (2007).
49. U.S. Food & Drug Administration. Evaluation of automatic class III designation for OsteoDetect. (2018). Available at: https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180005.pdf. (Accessed: 2nd October 2019)
50. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
51. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
52. Gengsheng Qin & Hotilovac, L. Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. *Stat. Methods Med. Res.* **17**, 207–221 (2008).

53. Obuchowski, N. A. On the comparison of correlated proportions for clustered data. *Stat. Med.* **17**, 1495–1507 (1998).
54. Yang, Z., Sun, X. & Hardin, J. W. A note on the tests for clustered matched-pair binary data. *Biom. J.* **52**, 638–652 (2010).
55. Fagerland, M. W., Lydersen, S. & Laake, P. Recommended tests and confidence intervals for paired binomial proportions. *Stat. Med.* **33**, 2850–2875 (2014).
56. Liu, J.-P., Hsueh, H.-M., Hsieh, E. & Chen, J. J. Tests for equivalence or non-inferiority for paired binary data. *Stat. Med.* **21**, 231–245 (2002).
57. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap*. (Springer US, 1993).
58. Chihara, L. M., Hesterberg, T. C. & Dobrow, R. P. *Mathematical Statistics with Resampling and R & Probability with Applications and R Set*. (Wiley, 2014).
59. Gur, D., Bandos, A. I. & Rockette, H. E. Comparing areas under receiver operating characteristic curves: potential impact of the 'Last' experimentally measured operating point. *Radiology* **247**, 12–15 (2008).
60. Metz, C. E. & Pan, X. 'Proper' Binormal ROC Curves: Theory and Maximum-Likelihood Estimation. *Journal of Mathematical Psychology* **43**, 1–33 (1999).
61. Chakraborty, D. P. *Observer Performance Methods for Diagnostic Imaging: Foundations, Modeling, and Applications with R-Based Examples*. (CRC Press, 2017).
62. Obuchowski, N. A. & Rockette, H. E. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an anova approach with dependent observations. *Communications in Statistics - Simulation and Computation* **24**, 285–308 (1995).
63. Hillis, S. L. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. *Stat. Med.* **26**, 596–619 (2007).
64. CRAN - Package RJafroc. Available at: https://cran.r-project.org/web/packages/RJafroc/index.html. (Accessed: 29th January 2019)

65. Aickin, M. & Gensler, H. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *Am. J. Public Health* **86**, 726–728 (1996).

66. Breast Screening Programme - NHS Digital. *NHS Digital* Available at: https://digital.nhs.uk/data-and-information/publications/statistical/breast-screening-programme. (Accessed: 17th July 2019)

Acknowledgments

We would like to acknowledge multiple contributors to this international project: Cancer Research UK, the OPTIMAM project team, and staff at the Royal Surrey County Hospital who developed the UK mammography imaging database; Sandra Tymms and Suzanne Steer for providing patient perspectives and invaluable advice and input throughout the project; Robin Wilson for providing screening programme and expert clinical perspective; all members of the Etemadi Research Group for their efforts in data aggregation and deidentification; members of Northwestern Medicine leadership without whom this couldn't be possible: Mark Schumacher, Carl Christensen, Doug King, and Charles Hogue. We also thank the tireless efforts of everyone at NMIT, including Mark Lombardi, Dalila Fridi, Paul Lendman, Bob Slavicek, Stavroula Xinos, Bob Milfajt and others; Victoria Cornelius who provided expert advice on statistical planning; Ross West & T Saensuksopa for assistance with data visualisation; Ali Eslami & Olaf Ronneberger for machine learning expertise; Hannah Forbes and Claire Zaleski for assistance with project management; Jonny Wong and Fraser Tan for coordinating labeling resources; Razia Ahmed, Rory Pilgrim, Anthony Phalen and Michelle Bawn for meticulous work on partnership formation; Roger Eng, Vashita Dhir, Rajshri Shah for data annotation and interpretation; Cameron Chen for critical review of the manuscript; Diego Ardila for infrastructure development; Cían Hughes and Diogo Moitinho de Almeida for early engineering work; Jessica Yoshimi, Xiang Ji, Will Chen, Teagan Daly, Huy Doan, Eric Lindley, Quang Duong for development of the labeling infrastructure. Professor Fiona Gilbert receives funding from the National Institute for Health Research (Senior Investigator award). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Author contributions

A.K., A.D., D.H., D.K., H.M, G.C.C., J.D.F., J.R.L., K.C.Y., L.P., M.D.H.B., M.T.S., M.S., R.S., S.M.M., S.S, and T.B. contributed to conception; A.K., B.R.P., C.J.K., D.H., D.T., F.J.G., J.D.F., J.R.L., K.C.Y., L.P., M.D.H.B., M.C., M.E., M.T.S., M.S., N.A., R.S., S.J., S.M.M., S.S., T.B. and V.G. contributed to design; D.M. D.T., F.G.V., G.C.C., H.M., J.D.F., J.G., K.C.Y., L.P., M.D.H.B., M.C., M.E., M.T.S., S.M.M., S.S., and V.G. contributed to acquisition; A.K., A.D, B.R.P., C.J.K., F.J.G., H.A., J.D.F., J.G., J.J.R., M.S., N.A., R.S., S.J., S.M.M., S.S. and V.G. contributed to analysis and interpretation; A.K., C.J.K., D.T., F.J.G., J.D.F., J.G., J.J.R, M.T.S., N.A., R.S, S.J., S.M.M., S.S., and V.G. contributed to drafting and revising the manuscript.

Competing interests

This study was funded by Google LLC and/or a subsidiary thereof (“Google”). S.M.M., M.T.S., V.G., J.G., N.A., T.B., M.C., G.C.C., D.H., S.J., A.K., C.J.K, D.K., J.R.L., H.M., B.R.-P., L.P., M.Su., D.T., J.D.F., and S.S. are employees of Google and own stock as part of the standard compensation package. J.J.R., R.S., F.J.G., and A.D. are paid consultants of Google. M.E., F.G.-V., D.M., K.C.Y., and M.H-.B received funding from Google to support the research collaboration. The authors have no other competing interests to disclose.

Extended Data Tables

Extended Data Table 1. Characteristics of the UK and UK test sets. For each feature, we constructed a joint 95% confidence interval on the proportions in each category. **a**, The UK test set was drawn from two sites in the UK over a four-year period. For reference, we present the corresponding statistics from the broader UK Breast Screening Programme (BSP)[66]. For comparison with national numbers, only screen-detected cancers are reported here. **b**, The US test was drawn from one academic medical center over an eighteen-year period. For reference, we present the corresponding statistics from the broader US screening population, as reported by the Breast Cancer Surveillance Consortium (BCSC)[2]. Cancers reported here occurred within 12 months of screening.

Extended Data Table 2. Detailed comparison between human clinical decisions and AI predictions.

a. Comparison of sensitivity and specificity between human benchmarks, derived retrospectively from the clinical record, and the predictions of the AI system. Score thresholds were chosen, based on separate validation data, to match or exceed the performance of each human benchmark (Methods, Operating point selection). These points are depicted graphically in Figure 2a. Bolded quantities represent estimated differences which are statistically significant for superiority; all others are statistically noninferior at a prespecified 5% margin. Note that the number of cases (N) differs from Figure 2a because a radiologist opinion was not available for all images. We also note that sensitivity and specificity metrics are not easily comparable to most prior publications in breast imaging (eg. the DMIST Trial [34]) given differences in follow up interval. Negative cases in the US dataset were upweighted to account for the sampling protocol (Methods, Inverse probability weighting).

b. Same columns as A, but using a version of the AI system trained exclusively on the UK dataset. It was tested on the US dataset to show generalisability of the AI across different populations and healthcare systems. Superiority comparisons on the UK data were conducted using Obuchowski's extension of the two-sided McNemar test for clustered data. Noninferiority comparisons were Wald tests using the Obuchowski correction. Comparisons on the US data were performed with a two-sided permutation test. All *p*-values survived correction for multiple comparisons (Methods, Statistical analysis).

Extended Data Table 3. Detailed description of reader study case composition.

Row 1. 500 cases were selected for the reader study. The case mixture was enriched for positives as well as challenging negatives.

Row 2. Cases containing breast implants and those for which at least half of the readers indicated image quality concerns were excluded from analysis. The remaining 465 cases are represented in Figure 3a and b.

Row 3. We also restricted the cancers to those that developed within 12 months. Those that developed cancer later (but within 27 months) were excluded because they did not meet the follow-up criteria to be considered negative. The remaining 408 cases are represented in Extended Data Table 2c and Extended Data Figure 2.

Row 4. To perform localisation analysis, the areas of malignancy were determined using follow-up biopsy data. In three instances, ground truth could not reliably be determined. The remaining 405 cases are represented in Extended Data Figure 3.

Extended Data Table 4. Potential utility of the AI system in two clinical applications.

a. Simulation using the UK test set in which the AI system is used in place of the second reader when it concurs with the first reader. In cases of disagreement (12.02%) the consensus opinion was invoked. The high performance of this combination of human and machine suggests that approximately 88% of the second reader's effort can be eliminated while maintaining the standard of care produced by double reading. The AI system's decision was generated using operating point (i) in Figure 2a. Confidence intervals are Wald intervals computed with the Obuchowski correction for clustered data.

b. Evaluation of the AI system for low-latency triage. Operating points were set to perform with high NPV and PPV for detecting cancer in 12 months.

Extended Data Table 5. Discrepancies between the AI system and human readers.

For the UK comparison, we used the first reader operating point (i) shown in Figure 2a. For the US comparison, we used the operating point shown in Figure 2b.

Extended Data Table 6. Performance breakdowns. Analysis excludes technical recalls and US cases for which BI-RADS scores were unavailable.

a. Sensitivity across cancer subtypes in the UK data. We used the AI system operating point (i) in Figure 2a. Also shown is the first reader performance on the same subset.

b. Sensitivity across cancer subtypes in the US data. We used the AI system operating point (i) in Figure 2a. Reader performance was derived from the clinical BI-RADS scores on the same subset. ILC = Invasive lobular carcinoma, IDC = invasive ductal carcinoma, DCIS = ductal carcinoma in situ.

c. Performance across breast density categories. BI-RADS[35] breast density was extracted from the radiology report rendered at the time of screening, only available in the US dataset. We used the AI system operating point shown in Figure 2b. Adjusted specificities were computed using inverse probability weighting (Methods).

Extended Data Table 7. Reader experience from the UK clinical dataset (a) and the independent reader study (b).

Extended Data Figures

Extended Data Figure 1. Unweighted evaluation of breast cancer prediction on the US test set. Unlike in Figure 2b, the sensitivity and specificity were computed without the use of inverse probability weights to account for the spectrum-enrichment of the study population. Since hard negatives are overrepresented, the specificity of both the AI system and the human readers is reduced. The unweighted human sensitivity and specificity are 48.10% (n = 553) and 69.65% (n = 2,185), respectively.

Extended Data Figure 2. Breast cancer prediction performance compared to six independent readers with a 12-month follow up for cancer status. While the mean reader AUC was 0.750 (s.d. 0.049), the AI system achieved an AUC of 0.871 (95% CI: [0.785, 0.919]). The AI system exceeded human performance by a significant margin (Δ = +0.121, 95% CI: [0.070, 0.173], p = 0.0018, two-sided ORH method). In this analysis, there were 56 positives of 408 total cases; see Extended Data Table 3. Note that this sample of cases was enriched for patients that had received a negative biopsy result (n=119), making this a more challenging population for screening. Since these external readers were not gatekeepers for follow up and eventual cancer diagnosis, there was no bias in favour of reader performance at this shorter time horizon. See Figure 3A for a comparison with a time interval chosen to encompass a subsequent screening exam.

Extended Data Figure 3. Multi-location receiver operating characteristic (mLROC) analysis.
Similar to Extended Data Figure 2, but true positives require localisation of a malignancy in any of the four mammogram views (Methods, Localisation analysis). Here, the cancer interval was 12 months (n = 53 positives of 405 cases; see Extended Data Table 3). The dotted line indicates a false positive rate of 10%, which was used as the right-hand boundary for the partial AUC (pAUC) calculation. The mean reader pAUC was 0.029 (s.d. 0.005), while the AI system's pAUC was 0.048 (95% CI: [0.035, 0.061]). The AI system exceeded human performance by a significant margin (Δ = +0.0192, 95% CI: [0.0086, 0.0298], *p* = 0.0004, two-sided ORH method).

Extended Data Figure 4. Evidence for the gatekeeper effect in retrospective datasets. These figures show the change in observed reader sensitivity in the UK (**a**) and the US (**b**) as the cancer follow-up interval is extended. At short intervals, measured reader sensitivity is extremely high, owing to the fact that biopsies are only triggered based on radiological suspicion. As the time interval is extended, the task becomes more difficult and measured sensitivity declines. Part of this decline stems from the development of new cancers that were impossible to detect at initial screening. However, more precipitous drops occur when the follow-up window encompasses the screening interval (36 months in the UK, 12 and 24 months

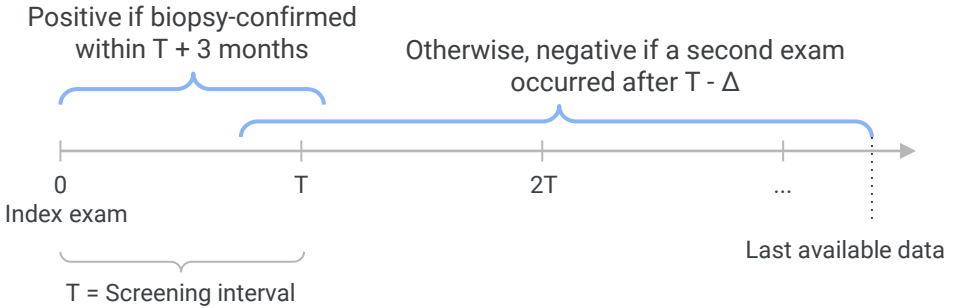
in the US). This is suggestive of what happens to reader metrics when gatekeeper bias is mitigated by another screening examination.

Extended Data Figure 5. Quantitative evaluation of reader and AI system performance with a 12-month follow-up interval for ground truth cancer positive status.
Because a 12-month follow-up interval is unlikely to encompass a subsequent screening exam in either country, reader-model comparisons on retrospective clinical data may be contaminated by the gatekeeper effect (Extended Data Figure 4). See Figure 2 for comparison with longer time intervals.
a. AI system performance on UK data. This plot was derived from a total of 25,717 eligible examples including 274 positives. The AI system achieved an AUC of 0.966 ([0.954, 0.977], 95% CI).
b. AI system performance on US data. This plot was derived from a total of 2,770 eligible examples including 359 positives. The AI system achieved an AUC of 0.883 ([0.859, 0.903], 95% CI).
c. Reader performance. In computing reader metrics on the UK data, we excluded cases for which the reader recommended repeat mammography to address technical issues. In the US data, radiologist performance could only be assessed on the subset of cases for which a BI-RADS grade was available.

Evaluation data sets

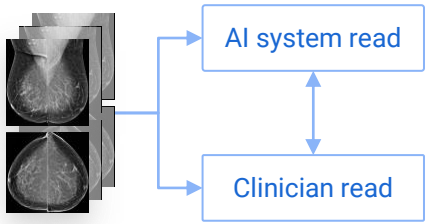
Number of women	25,856	3,097
Clinical evaluation	Double reader	Single reader
Screening interval	3 years	1 or 2 years
Cancer follow-up	39 months	27 months
Number of cancers	414 (1.6%)	686 (22.2%)

Ground truth determination



Evaluation

Comparison with retrospective clinical performance



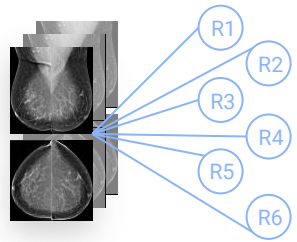
UK and US test sets

Generalization across data sets

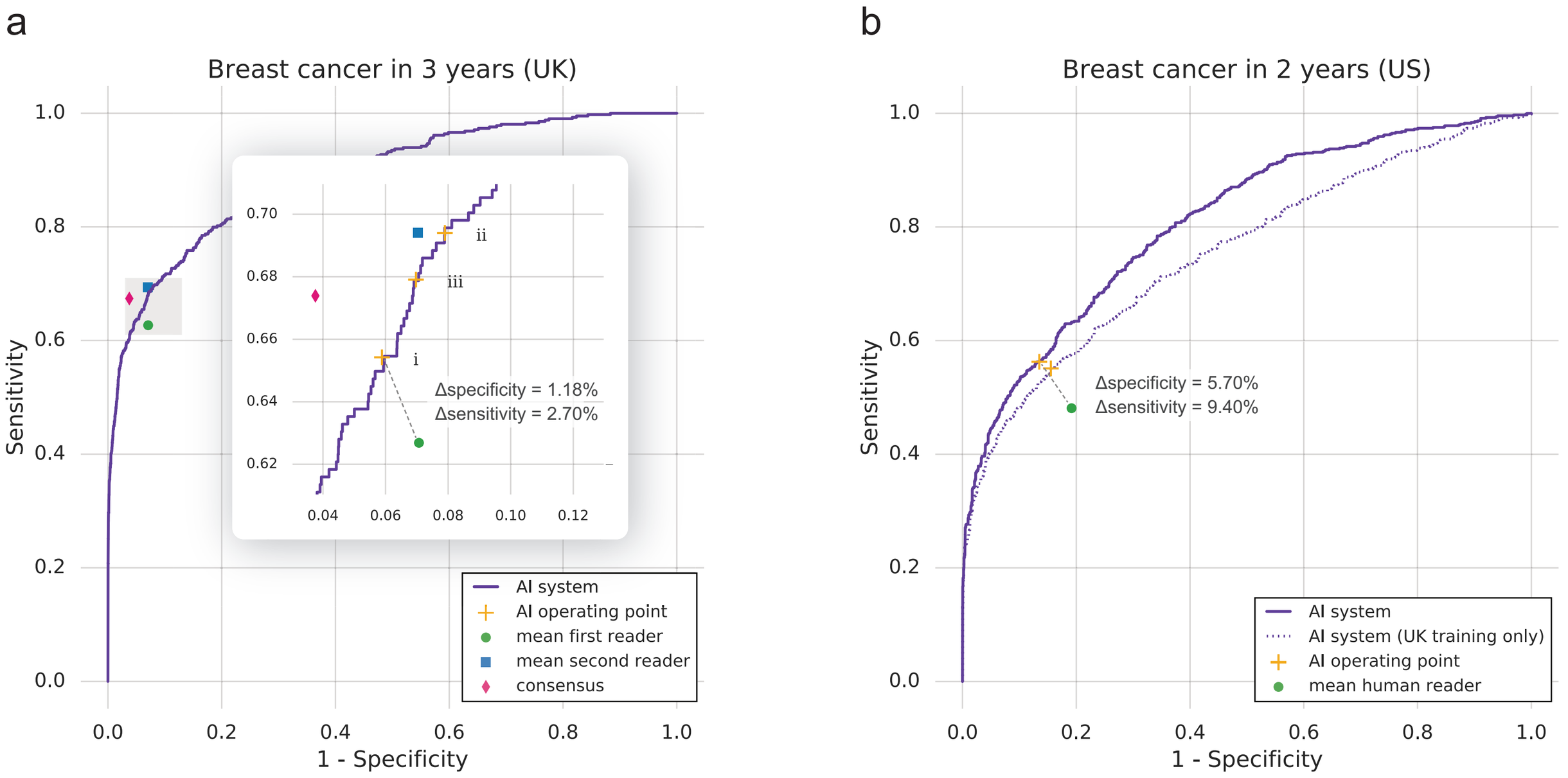
Trained on UK training set

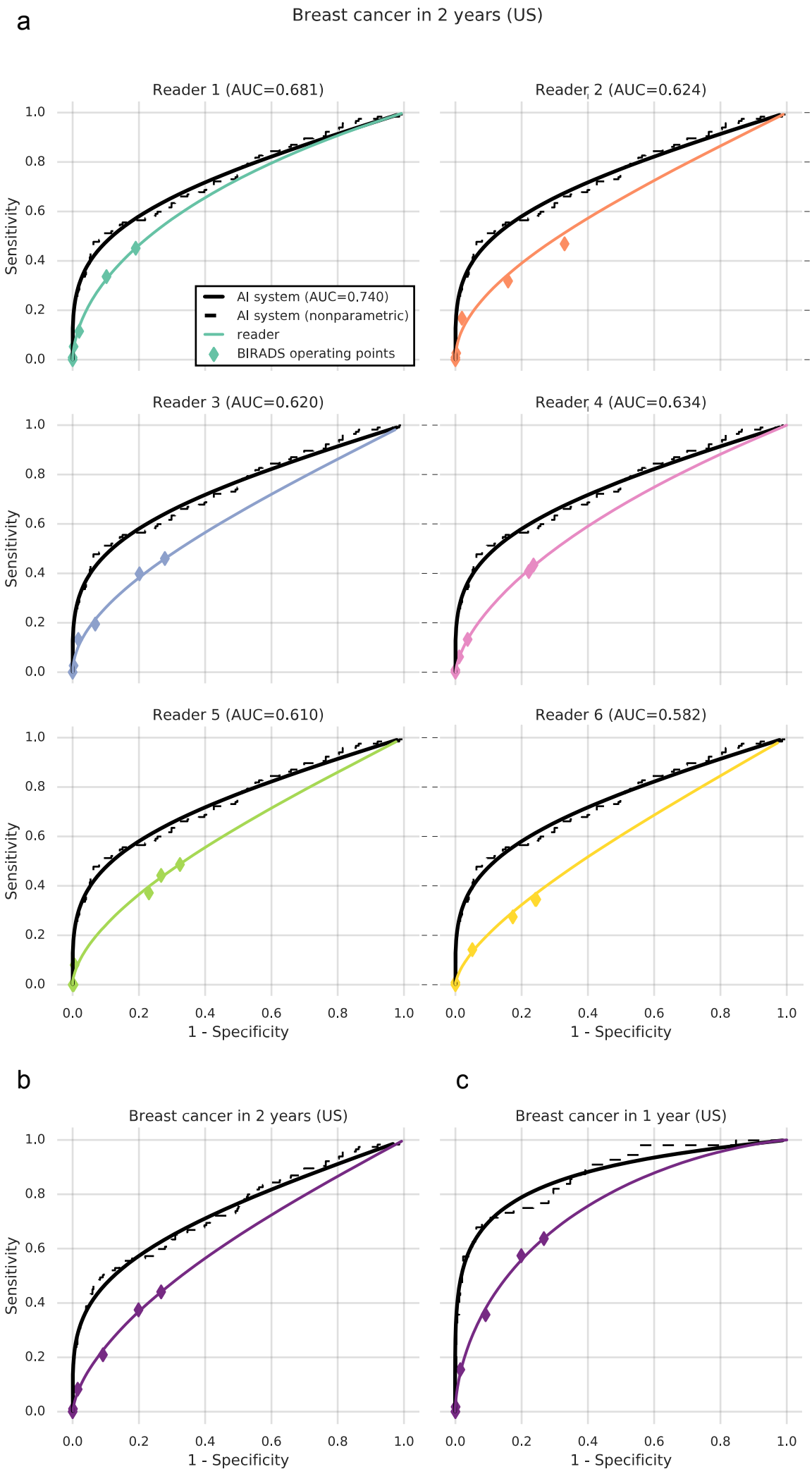
Tested on US test set

Independently-conducted reader study

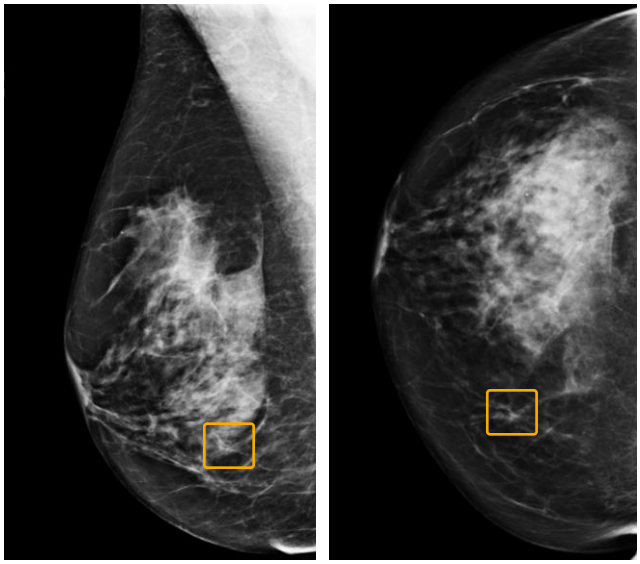


6 radiologists read 500 cases from US test set


a
Breast cancer in 3 years (UK)
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0
0.70
0.68
0.66
0.64
0.62
0.04
0.06
0.08
0.10
0.12
i
ii
iii
Δspecificity = 1.18%
Δsensitivity = 2.70%
AI system
AI operating point
mean first reader
mean second reader
consensus
b
Breast cancer in 2 years (US)
Sensitivity
1 - Specificity
Δspecificity = 5.70%
Δsensitivity = 9.40%
AI system
AI system (UK training only)
AI operating point
mean human reader


a
Breast cancer in 2 years (US)
Reader 1 (AUC=0.681)
Reader 2 (AUC=0.624)
Reader 3 (AUC=0.620)
Reader 4 (AUC=0.634)
Reader 5 (AUC=0.610)
Reader 6 (AUC=0.582)
AI system (AUC=0.740)
AI system (nonparametric)
reader
BIRADS operating points
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0
b
Breast cancer in 2 years (US)
c
Breast cancer in 1 year (US)

a



b