Article

International evaluation of an AI system for breast cancer screening

https://doi.org/10.1038/s41586-019-1799-6

Received: 27 July 2019

Accepted: 5 November 2019

Published online: 1 January 2020

Scott Mayer McKinney[1,14]*, Marcin Sieniek[1,14], Varun Godbole[1,14], Jonathan Godwin[2,14], Natasha Antropova[2], Hutan Ashrafian[3,4], Trevor Back[2], Mary Chesus[2], Greg C. Corrado[1], Ara Darzi[3,4,5], Mozziyar Etemadi[6], Florencia Garcia-Vicente[6], Fiona J. Gilbert[7], Mark Halling-Brown[8], Demis Hassabis[2], Sunny Jansen[9], Alan Karthikesalingam[10], Christopher J. Kelly[10], Dominic King[10], Joseph R. Ledsam[2], David Melnick[6], Hormuz Mostofi[1], Lily Peng[1], Joshua Jay Reicher[11], Bernardino Romera-Paredes[2], Richard Sidebottom[12,13], Mustafa Suleyman[2], Daniel Tse[1]*, Kenneth C. Young[8], Jeffrey De Fauw[2,15] & Shravya Shetty[1,15]*

Screening mammography aims to identify breast cancer at earlier stages of the disease, when treatment can be more successful[1]. Despite the existence of screening programmes worldwide, the interpretation of mammograms is affected by high rates of false positives and false negatives[2]. Here we present an artificial intelligence (AI) system that is capable of surpassing human experts in breast cancer prediction. To assess its performance in the clinical setting, we curated a large representative dataset from the UK and a large enriched dataset from the USA. We show an absolute reduction of 5.7% and 1.2% (USA and UK) in false positives and 9.4% and 2.7% in false negatives. We provide evidence of the ability of the system to generalize from the UK to the USA. In an independent study of six radiologists, the AI system outperformed all of the human readers: the area under the receiver operating characteristic curve (AUC-ROC) for the AI system was greater than the AUC-ROC for the average radiologist by an absolute margin of 11.5%. We ran a simulation in which the AI system participated in the double-reading process that is used in the UK, and found that the AI system maintained non-inferior performance and reduced the workload of the second reader by 88%. This robust assessment of the AI system paves the way for clinical trials to improve the accuracy and efficiency of breast cancer screening.

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

Despite the widespread adoption of mammography, interpretation of these images remains challenging. The accuracy achieved by experts in cancer detection varies widely, 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 that are missed at screening may not be identified until they are more advanced and less amenable to treatment[14].

AI may be uniquely poised to help with this challenge. Studies have demonstrated the ability of AI to meet or exceed the performance of human experts on several tasks of medical-image analysis[15–19]. As a shortage of mammography professionals threatens the 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 several assistive tools have been approved for medical use[24]. Despite early promise[25,26], this generation of software failed to improve the performance of readers 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 characterized systems for breast cancer prediction with stand-alone performance that approaches that of human experts[29,30]. However, the existing work has several limitations. Most studies are based on small, enriched datasets with limited follow-up, and few have compared performance to readers in actual clinical practice—instead relying on laboratory-based simulations of the reading environment. So far 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 that are no longer than 12 months[29,30,32,33]

[1]Google Health, Palo Alto, CA, USA. [2]DeepMind, London, UK. [3]Department of Surgery and Cancer, Imperial College London, London, UK. [4]Institute of Global Health Innovation, Imperial College London, London, UK. [5]Cancer Research UK Imperial Centre, Imperial College London, London, UK. [6]Northwestern Medicine, Chicago, IL, USA. [7]Department of Radiology, Cambridge Biomedical Research Centre, University of Cambridge, Cambridge, UK. [8]Royal Surrey County Hospital, Guildford, UK. [9]Verily Life Sciences, South San Francisco, CA, USA. [10]Google Health, London, UK. [11]Stanford Health Care and Palo Alto Veterans Affairs, Palo Alto, CA, USA. [12]The Royal Marsden Hospital, London, UK. [13]Thirlestaine Breast Centre, Cheltenham, UK. [14]These authors contributed equally: Scott Mayer McKinney, Marcin T. Sieniek, Varun Godbole, Jonathan Godwin. [15]These authors jointly supervised this work: Jeffrey De Fauw, Shravya Shetty. *e-mail: scottmayer@google.com; tsed@google.com; sshetty@google.com

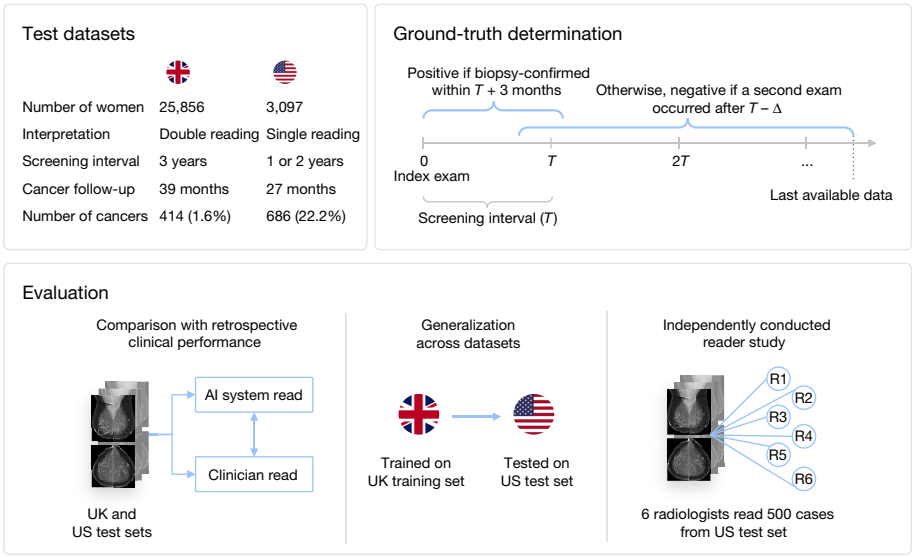


Fig. 1 | Development of an AI system to detect cancer in screening mammograms. Datasets representative of the UK and US breast cancer screening populations were curated from three screening centres in the UK and one centre in the USA. Outcomes were derived from the biopsy record and longitudinal follow-up. An AI system was trained to identify the presence of breast cancer from a set of screening mammograms, and was evaluated in three primary ways: first, AI predictions were compared with the historical decisions made in clinical practice; second, to evaluate the generalizability across populations, a version of the AI system was developed using only the UK data and retested on the US data; and finally, the performance of the AI system was compared to that of six independent radiologists using a subset of the US test set.

means that more subtle cancers that are 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 the USA. We compare the predictions of the system to those made by readers in routine clinical practice and show that performance exceeds that of individual radiologists. These observations are confirmed with an independently conducted reader study. Furthermore, we show how this system might be integrated into screening workflows, and provide evidence that the system can generalize across continents. Figure 1 shows an overview of the project.

Datasets from cancer screening programmes

A deep learning model for identifying breast cancer in screening mammograms was developed and evaluated using two large datasets from the UK and the USA. We report results on test sets that were not used to train or tune the AI system.

The UK test set consisted of screening mammograms that were collected between 2012 and 2015 from 25,856 women at two screening centres in England, where women are screened every three years. It included 785 women who had a biopsy, and 414 women with cancer that was diagnosed within 39 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 USA, where women are screened every one to two years, consisted of screening mammograms that were collected between 2001 and 2018 from 3,097 women at one academic medical centre. We included images from all 1,511 women who were 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 of imaging.

Breast cancer outcome was determined on the basis of multiple years of follow-up (Fig. 1). We chose the follow-up duration on the basis of the screening interval in the country of origin for each dataset. In a similar manner to previous work[34], we augmented each interval with a three-month buffer to account for variability in scheduling and latency of follow-up. Cases that were designated as cancer-positive were accompanied by a biopsy-confirmed diagnosis within the follow-up period. Cases labelled 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 outcomes to evaluate the predictions of the AI system as well as the original decisions made by readers in the course of clinical practice. Human performance was computed on the basis of the clinician's decision to recall the patient for further diagnostic investigation. The receiver operating characteristic (ROC) curve of the AI system is shown in Fig. 2.

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

Compared to the first reader, the AI system demonstrated a statistically significant improvement in absolute specificity of 1.2% (95% confidence interval (CI) 0.29%, 2.1%; $P = 0.0096$ for superiority) and an improvement in absolute sensitivity of 2.7% (95% CI −3%, 8.5%; $P = 0.004$ for non-inferiority at a pre-specified 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 ($P = 0.02$). Likewise, the AI system showed non-inferiority (at a 5% margin) to the consensus judgment for specificity ($P < 0.001$) and sensitivity ($P = 0.0039$).

In the standard screening protocol in the USA, each mammogram is interpreted by a single radiologist. We used the BI-RADS[35] score that was assigned to each case in the original screening context as a proxy for human cancer prediction (see Methods section 'Interpreting clinical reads'). Compared to the typical reader, the AI system demonstrated statistically significant improvements in absolute specificity of 5.7%

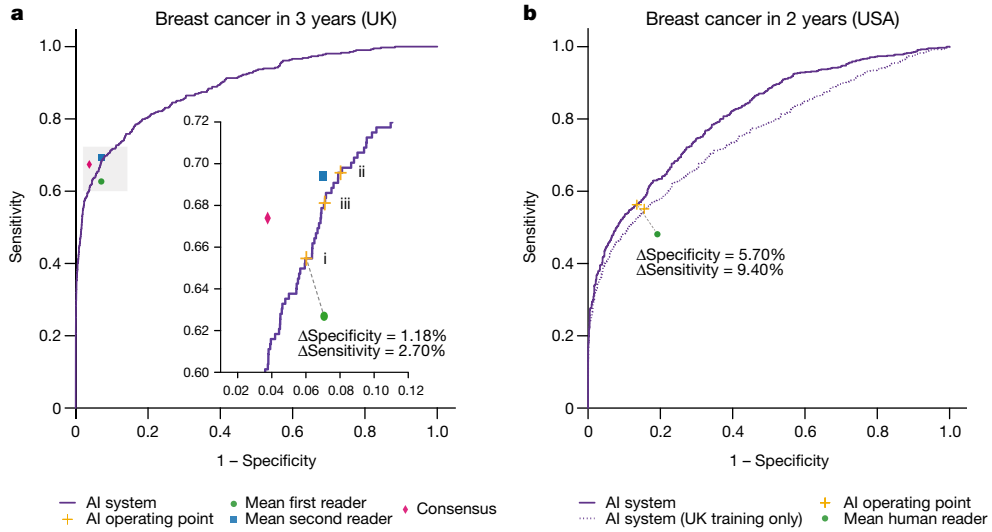


Fig. 2 | Performance of the AI system and clinical readers in breast cancer prediction. a, The ROC curve of the AI system on the UK screening data. The AUC is 0.889 (95% CI 0.871, 0.907; $n = 25{,}856$ patients). Also shown are the sensitivity and specificity pairs for the human decisions made in clinical practice. Cases were considered positive if they received a biopsy-confirmed diagnosis of cancer within 39 months of screening. The consensus decision represents the standard of care in the UK, and will involve input from between two and three expert readers. The inset shows a magnification of the grey 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**, The ROC curve of the AI system on the US screening data. When trained on both datasets (solid curve), the AUC is 0.8107 (95% CI 0.791, 0.831; $n = 3{,}097$ patients). When trained on only the UK dataset (dotted curve), the AUC is 0.757 (95% CI 0.732, 0.780). 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 diagnosis of cancer within 27 months of screening. AI system operating points were chosen, using a separate validation dataset, to exceed the sensitivity and specificity of the average reader. Negative cases were upweighted to account for the sampling protocol (see Methods section 'Inverse probability weighting'). Extended Data Figure 1 shows an unweighted analysis. See Extended Data Table 2a for statistical comparisons of sensitivity and specificity.

(95% CI 2.6%, 8.6%; $P < 0.001$) and in absolute sensitivity of 9.4% (95% CI 4.5%, 13.9%; $P < 0.001$; Extended Data Table 2a).

Generalization across populations

To evaluate the ability of the AI system to generalize across populations and screening settings, we trained the same architecture using only the UK dataset and applied it to the US test set (Fig. 2b). Even without exposure to the US training data, the ROC curve of the AI system encompasses the point that indicates the average performance of US radiologists. Again, the AI system showed improved specificity (+3.5%, $P = 0.0212$) and sensitivity (+8.1%, $P = 0.0006$; Extended Data Table 2b) compared with radiologists.

Comparison with a reader study

In a reader study that was conducted by an external clinical research organization, six US-board-certified radiologists who were compliant with the requirements of the Mammography Quality Standards Act (MQSA) interpreted 500 mammograms that were randomly sampled from the US test set. Where data were available, readers were equipped with contextual information typically available in the clinical setting, including the patient's age, breast cancer history, and previous 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, and 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 (Fig. 3).

The AI system exceeded the average performance of radiologists by a significant margin (change in area under curve (ΔAUC) = +0.115, 95% CI 0.055, 0.175; $P = 0.0002$). Similar results were observed when a follow-up period of one year was used instead of 27 months (Fig. 3c, Extended Data Fig. 2).

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

We used multi-localization receiver operating characteristic (mLROC) analysis[37] to compare the ability of the readers and the AI system to identify malignant lesions within each case (see Methods section 'Localization analysis').

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

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

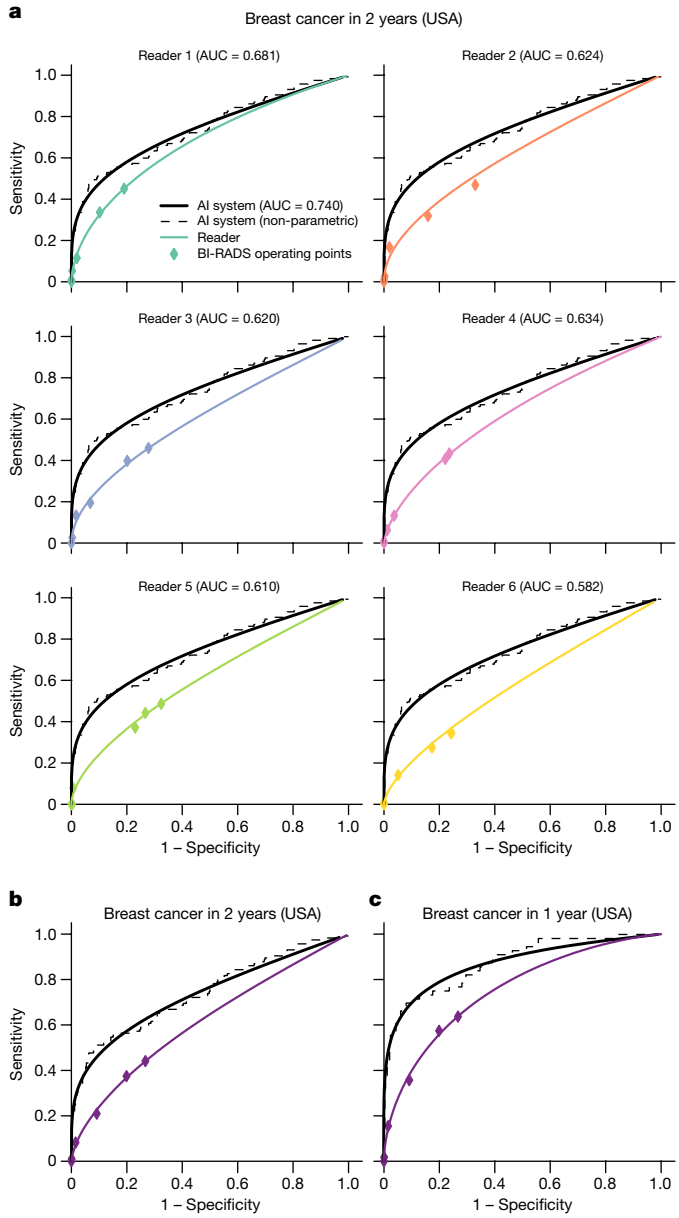


Fig. 3 | Performance of the AI system in breast cancer prediction compared to six independent readers. a, Six readers rated each case ($n = 465$) using the six-point BI-RADS scale. A fitted ROC curve for each of the readers is compared to the ROC curve of the AI system (see Methods section 'Statistical analysis'). For reference, a non-parametric ROC curve is presented in tandem. Cases were considered positive ($n = 113$) if they received a pathology-confirmed diagnosis of cancer within 27 months of the time of screening. Note that this sample of cases was enriched for patients who 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), whereas the AUC for the AI system was 0.740 (95% CI 0.696, 0.794). The AI system exceeded human performance by a significant margin (ΔAUC = +0.115, 95% CI 0.055, 0.175; $P = 0.0002$ by two-sided ORH method (see Methods section 'Statistical analysis')). For results using a 12-month interval, see Extended Data Fig. 2. **b**, Pooled results from all six readers from **a**. **c**, Pooled results ($n = 408$) from all 6 readers using a 12-month interval for cancer definition. Cases were considered positive ($n = 56$) if they received a pathology-confirmed cancer diagnosis within one year (Extended Data Table 3).

To identify normal cases with high confidence, we used a very-low decision threshold. For the UK data, we achieved a negative predictive value (NPV) of 99.99% while retaining a specificity of 41.15%. Similarly, for 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 dismiss 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 USA–UK). By comparison, consensus double reading in our UK dataset included one cancer in every 182 cases that were deemed normal.

To identify cancer cases with high confidence, we used a very-high decision threshold. For the UK data, we achieved a positive predictive value (PPV) of 85.6% while retaining a sensitivity of 41.2%. Similarly, for 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 prioritize 30–40% of cancer cases, with approximately five out of six follow-ups leading to a diagnosis of cancer. By comparison, in our study only 22.8% of UK cases that were recalled by consensus double reading and 4.9% of US cases that were 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 whereas 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 that was missed by all six radiologists, but correctly identified by the AI system. Figure 4b shows a sample cancer case that was caught by all six radiologists, but missed by the AI system. Although 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 (Supplementary Table 3). 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 sensitivity advantage of the AI system is concentrated on the identification of invasive cancers (for example, invasive lobular or ductal carcinoma) rather than in situ cancer (for example, 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 that is represented in this dataset (Extended Data Table 6c).

Discussion

In this study we present an AI system that outperforms radiologists on a clinically relevant task of breast cancer identification. These results held across two large datasets that are representative of different 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 non-inferior. 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 non-inferior to the second reader and consensus opinion.

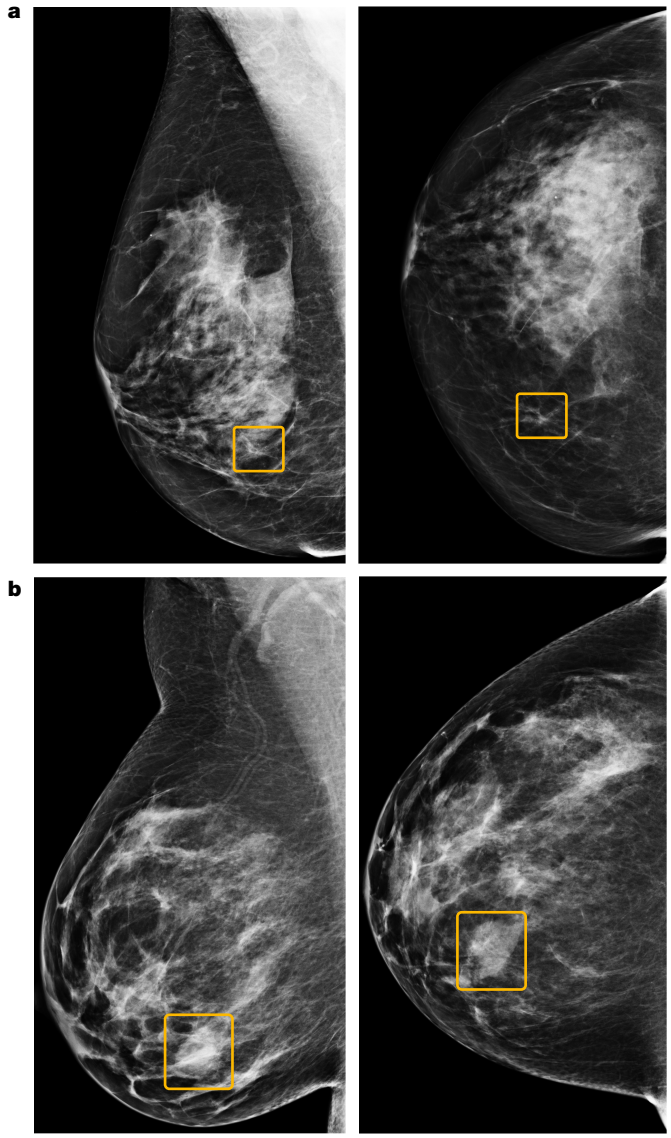


Fig. 4 | Discrepancies between the AI system and human readers. a, A sample cancer case that was missed by all six readers in the US reader study, but correctly identified by the AI system. The malignancy, outlined in yellow, is a small, irregular mass with associated microcalcifications in the lower inner right breast. **b**, A sample cancer case that was caught by all six readers in the US reader study, but missed by the AI system. The malignancy is a dense mass in the lower inner right breast. Left, mediolateral oblique view; right, craniocaudal view.

In the USA, the AI system exhibited specificity and sensitivity superior to that of radiologists practising in an academic medical centre. This trend was confirmed in an externally conducted reader study, which showed that the scores of the AI system stratified cases better than the BI-RADS ratings (the standard scale for mammography assessment in the USA) that were assigned by each of the six readers.

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

These comparisons are not without limitations. Although 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 dataset, which was drawn from a single screening centre and enriched for cancer cases.

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

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

To obtain high-quality ground-truth labels, we used extended follow-up intervals that were chosen to encompass a subsequent round of screening in each country. Although there is some precedent in clinical trials[34] and targeted cohort studies[42], this step is not usually taken during 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. As they are gatekeepers for biopsy, asymptomatic cases will only receive a cancer diagnosis if a mammogram raises the suspicions of a reader. A longer follow-up interval decouples the ground-truth labels from reader opinions (Extended Data Fig. 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 that are 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 Fig. 5).

We present early evidence of the ability of the AI system to generalize 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 could benefit from fine-tuning with local data.

The optimal use of the AI system within clinical workflows remains to be determined. The specificity advantage exhibited by the system suggests that it could help to 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 localization performance of the AI system suggests it holds early promise for flagging suspicious regions for review by experts. Notably, the additional cancers identified by the AI system tended to be invasive rather than in situ disease.

Beyond improving 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 a similar level of 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 shortages in countries such as the UK[43]. Prospective clinical studies will be required to understand the full extent to which this technology can benefit patient care.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1799-6.

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. The Canadian Task Force on Preventive Health Care. 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. J. Am. Med. Assoc*. **314**, 1599–1614 (2015).
8. Siu, A. L. 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* (US Food and Drug Administration, 2019; accessed 16 July 2019); http://www.fda.gov/radiation-emitting-products/mqsa-insights/mqsa-national-statistics
10. Cancer Research UK. *Breast Screening* (CRUK, 2017; accessed 26 July 2019); https://www.cancerresearchuk.org/about-cancer/breast-cancer/screening/breast-screening
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 Intern. Med*. **174**, 954–961 (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. *J. Am. Med. Assoc*. **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. *Radiat. Med*. **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. *Med. Phys*. **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*. **111**, 916–922 (2019).
30. Wu, N. et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans. Med. Imaging* https://doi.org/10.1109/TMI.2019.2945514 (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. D'Orsi, C. J. et al. *ACR BI-RADS Atlas: Breast Imaging Reporting and Data System* (American College of Radiology, 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**, 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. Royal College of Radiologists. *The Breast Imaging and Diagnostic Workforce in the United Kingdom* (RCR, 2016; accessed 22 July 2019); https://www.rcr.ac.uk/publication/breast-imaging-and-diagnostic-workforce-united-kingdom

Methods

Ethical approval

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

The UK dataset

The UK dataset was collected from three breast screening sites in the UK 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 three 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 follow-up. When there is disagreement, an arbitration process takes place.

The data were initially compiled by OPTIMAM (Cancer Research UK) between 2010 and 2018, from St George's Hospital (London), Jarvis Breast Centre (Guildford) and Addenbrooke's Hospital (Cambridge). The collected data included screening and follow-up mammograms (comprising mediolateral oblique and craniocaudal views of the left and right breasts), all radiologist opinions (including the arbitration result, if applicable) and the 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 women had 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 below the age of 47 at the time of the screen were excluded from validation and test sets, leaving 121,455 women. Finally, women for whom there was no exam with sufficient follow-up were excluded from validation and test sets. This last step resulted in the exclusion of 5,990 of 31,766 test-set cases (19%); see Supplementary Fig. 1.

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

The final test set consisted of 25,856 women (see Supplementary Fig. 1). When compared to the UK national breast cancer screening service, we observed a very similar distribution of cancer prevalence, age and, cancer type (see Extended Data Table 1a). Digital mammograms were acquired predominantly on devices manufactured by Hologic (95%), followed by General Electric (4%) and Siemens (1%).

The US dataset

The US dataset was collected from Northwestern Memorial Hospital (Chicago) between 2001 and 2018. In the USA, each screening mammogram is typically read by a single radiologist, and screens are conducted annually or biannually. The breast radiologists at this hospital receive fellowship training and only interpret breast-imaging studies. Their experience levels ranged from 1 to 30 years. The American College of Radiology (ACR) recommends that women start routine screening at the age of 40; other organizations, including the United States Preventive Services Task Force (USPSTF), recommend that screening begins at the age of 50 for women with an average risk of breast cancer[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 used 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. The data-processing steps involved in constructing the dataset are summarized in Supplementary Fig. 2.

Among women with a completed mammogram order, we collected records from all women with a pathology report that contained the term 'breast'. Among women that lacked such a pathology report, those 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 de-identification and transfer, women were excluded if their metadata were unavailable or corrupted. The women in the dataset were split randomly among train (55%), validation (15%) and test (30%) sets. For testing, a single case was chosen for each woman, following a similar procedure as for 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 and craniocaudal views of the left and right breasts), acquired for screening intent. Again, the vast majority of the studies were acquired using Hologic (including Loradbranded) devices (99%); the other 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 that involved 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 confirmed by biopsy. 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.

Randomization and blinding

Patients were randomized into training, validation, and test sets by applying a hash function to the de-identified medical record number. Set assignment was based on the value of the resulting integer modulo 100. For the UK data, values of 0–9 were reserved for the test set. For the US data, values of 0–29 were reserved for the test set. Test set sizes were chosen to produce, in expectation, a sufficient number of positives to power statistical comparisons on the metric of sensitivity.

The US and UK test sets were held back from AI system development, which only took place on the training and validation sets. Investigators did not access test set data until models, hyperparameters, and operating point thresholds were finalized. None of the readers who interpreted the images had knowledge of any aspect of the AI system.

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 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 that results when negative cases are randomly selected, we estimated additional factors by Monte Carlo simulation. Choosing one case per patient with our preferential sampling mechanism yielded 872 cases that were biopsied within 27 months, and 1,662 cases that were not (Supplementary Fig. 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 downweighted by 557.54/872 = 0.64. Cases that were not biopsied were upweighted by another 2,056.46/1,662 = 1.24, leading to a final weight of 19.04 × 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 outcome 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 harmonize this categorization 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. Women who received a biopsy that found any of these malignant pathologies were 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 were considered to be cancerous.

Interpreting clinical reads

In the UK screening setting, readers categorize mammograms from asymptomatic women as normal or abnormal, with a third option for technical recall owing to inadequate image quality. An abnormal result at the conclusion of the double-reading process results in further diagnostic assessment. 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 to be 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 later be refined on the basis of 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 the recommendation was based on mammographic findings, not on technical grounds or patient symptoms alone. Cases of technical recall were excluded from analysis, as the images were presumed to be 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 that is 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. By contrast, only the age of the patient 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 the Supplementary Methods and Supplementary Fig. 3.

Selection of operating points

The AI system natively produces a continuous score that represents the likelihood of cancer being 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 that separates 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 for both sensitivity and specificity.

The US dataset contains a single operating point for comparison, which corresponds to the 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 for both sensitivity and specificity.

Reader study

For the reader study, six 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 7b). Two of them were fellowship-trained in breast imaging. The sample of cases was stratified to contain 50% normal cases, 25% biopsy-confirmed negative cases and 25% biopsy-confirmed positive cases. A detailed description of the case composition of the reader study 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 using the values 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 the values 1, 2, 3, 4A, 4B, 4C or 5. Readers also gave a finer-grained score between 0 and 100 that was indicative of their suspicion that the case contains a malignancy.

In addition to the four standard mammographic screening images, clinical context was provided to better simulate the screening setting. Readers were presented with the preamble of the de-identified radiology report that was produced by the radiologist who originally interpreted the study. This contained information such as the age of the patient and their family history of cancer. The information was manually reviewed to ensure that no impression or findings were included.

Where possible (in 43% of cases), previous imaging was made available to the readers. Readers could review up to four sets of previous

screening exams that were acquired between 1 and 4 years earlier, accompanied by de-identified radiologist reports. If prior imaging was available, the study was read twice by each reader—first without the prior information, and then immediately after, with the prior information present. The system ensured that readers could not update their initial assessment after the prior information was presented. For cases for which previous exams were available, the final reader assessment (given after having reviewed the prior exams) was used for the analysis.

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

Localization 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 localizations, 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 in which 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 owing 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 ROI annotations surrounding suspicious findings in all cases to which they assigned a BI-RADS score of 3 or higher. A limit of six 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 produces a ranked list of rectangular bounding boxes for each case. To conduct a fair comparison, we allowed only the top two bounding boxes from the AI system to match the number of ROIs produced by the readers.

To compare the localization 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 localization 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 case-wide score exceeded this threshold and at least one culprit area was correctly localized in any of the four mammogram views. Correct localization 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 centre of their marking falls within the boundary of a ground-truth annotation[48]. This is potentially problematic as it does not properly penalize predicted bounding boxes that are so large as to be non-specific, but whose centre 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 ROI annotations of vastly different sizes depending on subjective factors of the annotator (see Supplementary Fig. 4). Similar work in three-dimensional chest computed tomography[18] used any pixel overlap to qualify for correct localization. Likewise, an FDA-approved software device for the detection of wrist fractures 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 the stand-alone performance of the AI system, the AUC-ROC was estimated using the normalized Wilcoxon (Mann–Whitney) *U* statistic[50]. This is the standard non-parametric method used by most modern software libraries. For the UK dataset, non-parametric 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 1,000 replications.

For 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 pseudo-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 non-inferiority[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 sensitivity and specificity of the AI system with those of the pool of radiologists. Confidence intervals on the difference were generated with the bootstrap method with 1,000 replications. A *P* value on the difference was generated through the use of a permutation test[58]. In each of 10,000 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 Fig. 3), to avoid bias from non-parametric analysis[59] we fitted parametric ROC curves to the data using the proper binormal model[60]. This issue was not alleviated by using the readers' ratings for their suspicion of malignancy, which showed very strong correspondence with the BI-RADS scores (Supplementary Fig. 5). As BI-RADS is used in actual screening practice, we chose to focus on these scores for their superior clinical relevance. In a similar fashion, we fitted 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 that are standard in the radiology community[61]. More specifically, we compared the AUC-ROC and pAUC-mLROC for the AI system to those of the average radiologist using the ORH procedure[62,63]. 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 stand-alone algorithm[61]. The jackknife method was used to estimate the covariance terms in the model. Computation of *P* values and confidence intervals 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 (https://cran.r-project.org/web/packages/RJafroc/index.html).

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 against 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 pre-specified before the test set was inspected. We used a statistical significance threshold of 0.05. All seven *P* values survived correction for multiple comparisons using the Holm–Bonferroni method[64].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

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/.

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 support 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).

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. Ellis, I. O. et al. *Pathology Reporting of Breast Disease in Surgical Excision Specimens Incorporating the Dataset for Histological Reporting of Breast Cancer, June 2016* (Royal College of Pathologists, accessed 22 July 2019); https://www.rcpath.org/resourceLibrary/g148-breastdataset-hires-jun16-pdf.html
47. Chakraborty, D. P. & Yoon, H.-J. Operating characteristics predicted by models for diagnostic tasks involving lesion localization. *Med. Phys*. **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. US Food and Drug Administration. *Evaluation of Automatic Class III Designation for OsteoDetect* (FDA, 2018; accessed 2 October 2019); https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180005.pdf
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, 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. *J. Math. Psychol*. **43**, 1–33 (1999).
61. Chakraborty, D. P. *Observer Performance Methods for Diagnostic Imaging: Foundations, Modeling, and Applications with R-Based Examples* (CRC, 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. *Commun. Stat. Simul. Comput*. **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. 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).
65. NHS Digital. *Breast Screening Programme* (NHS, accessed 17 July 2019); https://digital.nhs.uk/data-and-information/publications/statistical/breast-screening-programme

Acknowledgements 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; S. Tymms and S. Steer for providing patient perspectives; R. Wilson for providing a clinical perspective; all members of the Etemadi Research Group for their efforts in data aggregation and de-identification; and members of the Northwestern Medicine leadership, without whom this work would not have been possible (M. Schumacher, C. Christensen, D. King and C. Hogue). We also thank everyone at NMIT for their efforts, including M. Lombardi, D. Fridi, P. Lendman, B. Slavicek, S. Xinos, B. Milfajt and others; V. Cornelius, who provided advice on statistical planning; R. West and T. Saensuksopa for assistance with data visualization; A. Eslami and O. Ronneberger for expertise in machine learning; H. Forbes and C. Zaleski for assistance with project management; J. Wong and F. Tan for coordinating labelling resources; R. Ahmed, R. Pilgrim, A. Phalen and M. Bawn for work on partnership formation; R. Eng, V. Dhir and R. Shah for data annotation and interpretation; C. Chen for critically reading the manuscript; D. Ardila for infrastructure development; C. Hughes and D. Moitinho de Almeida for early engineering work; and J. Yoshimi, X. Ji, W. Chen, T. Daly, H. Doan, E. Lindley and Q. Duong for development of the labelling infrastructure. A.D. and F.J.G. receive funding from the National Institute for Health Research (Senior Investigator award). Infrastructure support for this research was provided by the NIHR Imperial Biomedical Research Centre (BRC). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

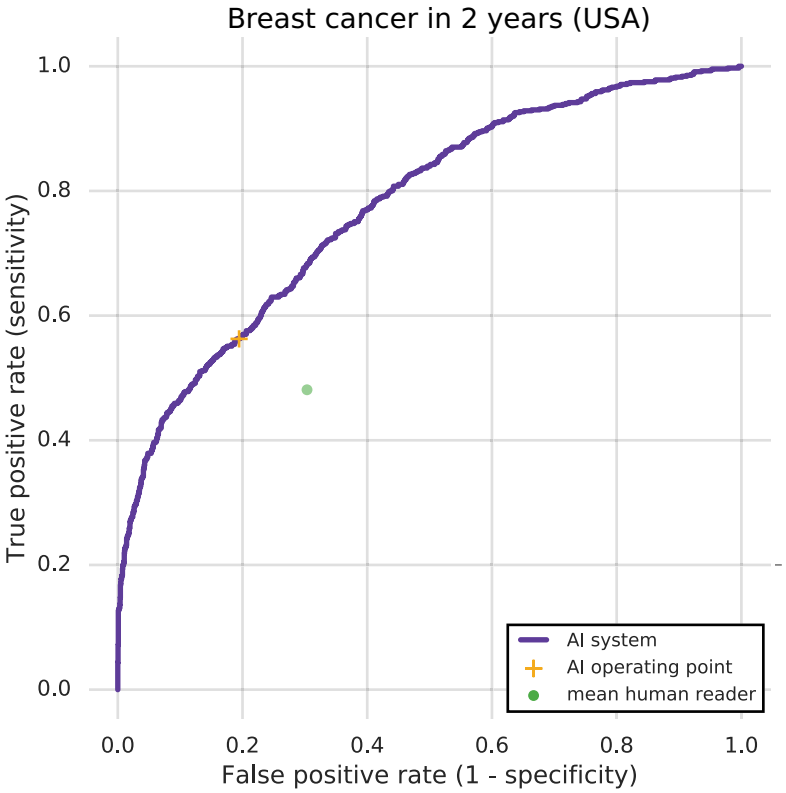
Author contributions A.K., A.D., D.H., D.K., H.M., G.C.C., J.D.F., J.R.L., K.C.Y., L.P., M.H.-B., M. Sieniek, M. Suleyman, R.S., S.M.M., S.S. and T.B. contributed to the conception of the study; A.K., B.R.-P., C.J.K., D.H., D.T., F.J.G., J.D.F., J.R.L., K.C.Y., L.P., M.H.-B., M.C., M.E., M. Sieniek, M. Suleyman, N.A., R.S., S.J., S.M.M., S.S., T.B. and V.G. contributed to study design; D.M., D.T., F.G.-V., G.C.C., H.M., J.D.F., J.G., K.C.Y., L.P., M.H.-B., M.C., M.E., M. Sieniek, S.M.M., S.S. and V.G. contributed to acquisition of the data; A.K., A.D., B.R.-P., C.J.K., F.J.G., H.A., J.D.F., J.G., J.J.R., M. Suleyman, N.A., R.S., S.J., S.M.M., S.S. and V.G. contributed to analysis and interpretation of the data; A.K., C.J.K., D.T., F.J.G., J.D.F., J.G., J.J.R., M. Sieniek, 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. Sieniek, 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. Suleyman, 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.

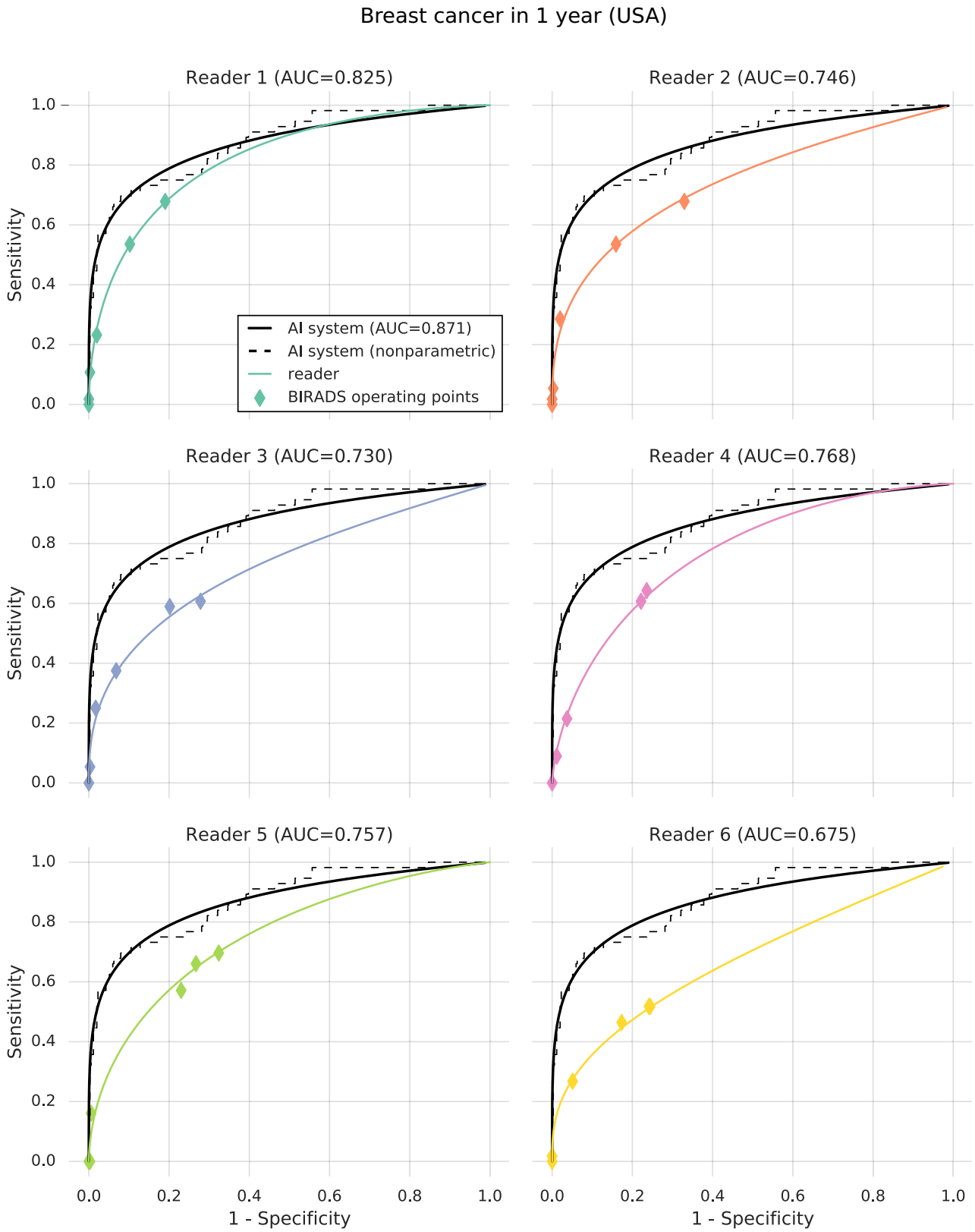
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1799-6.

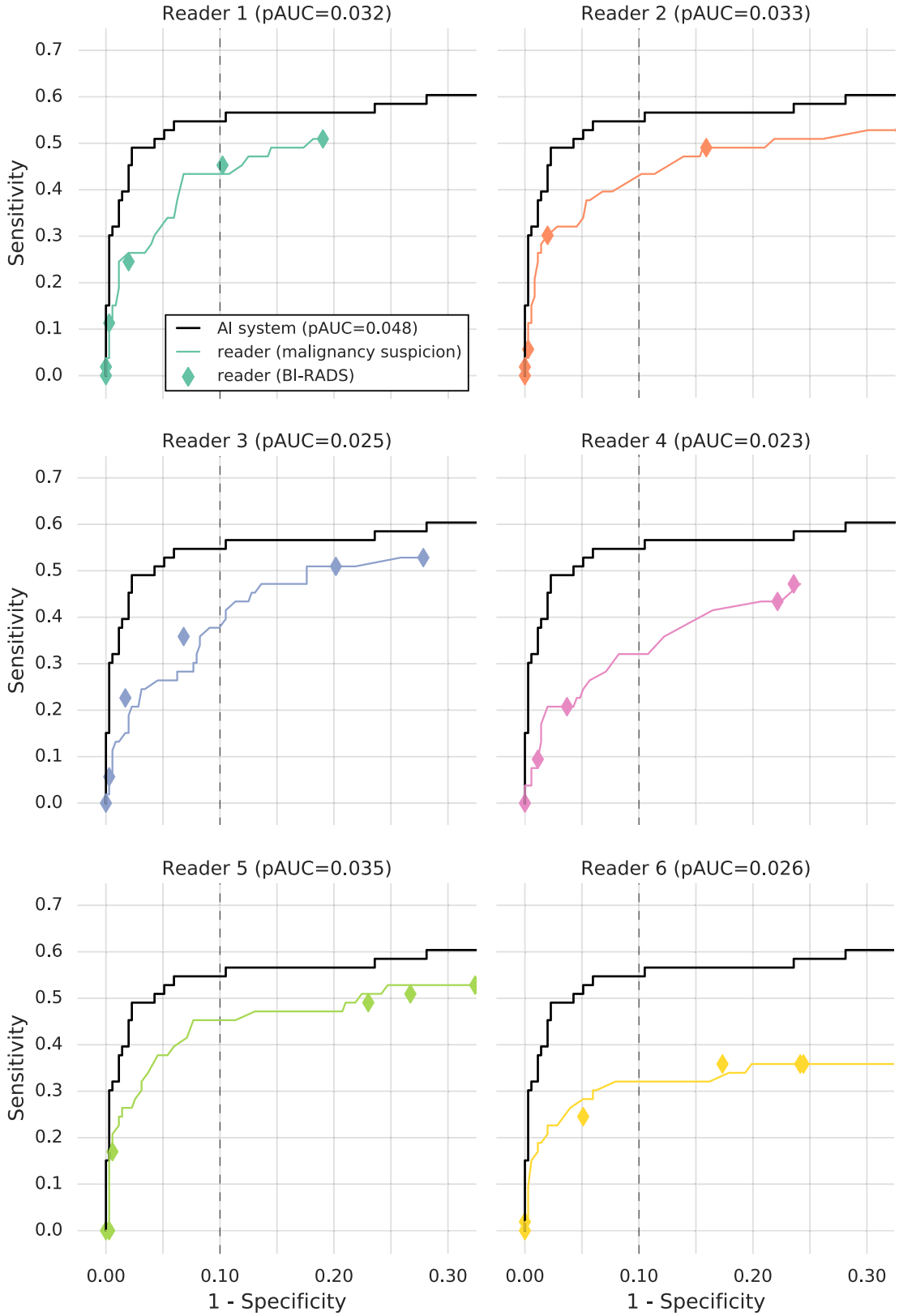
Correspondence and requests for materials should be addressed to S.M.M., D.T. or S.S.



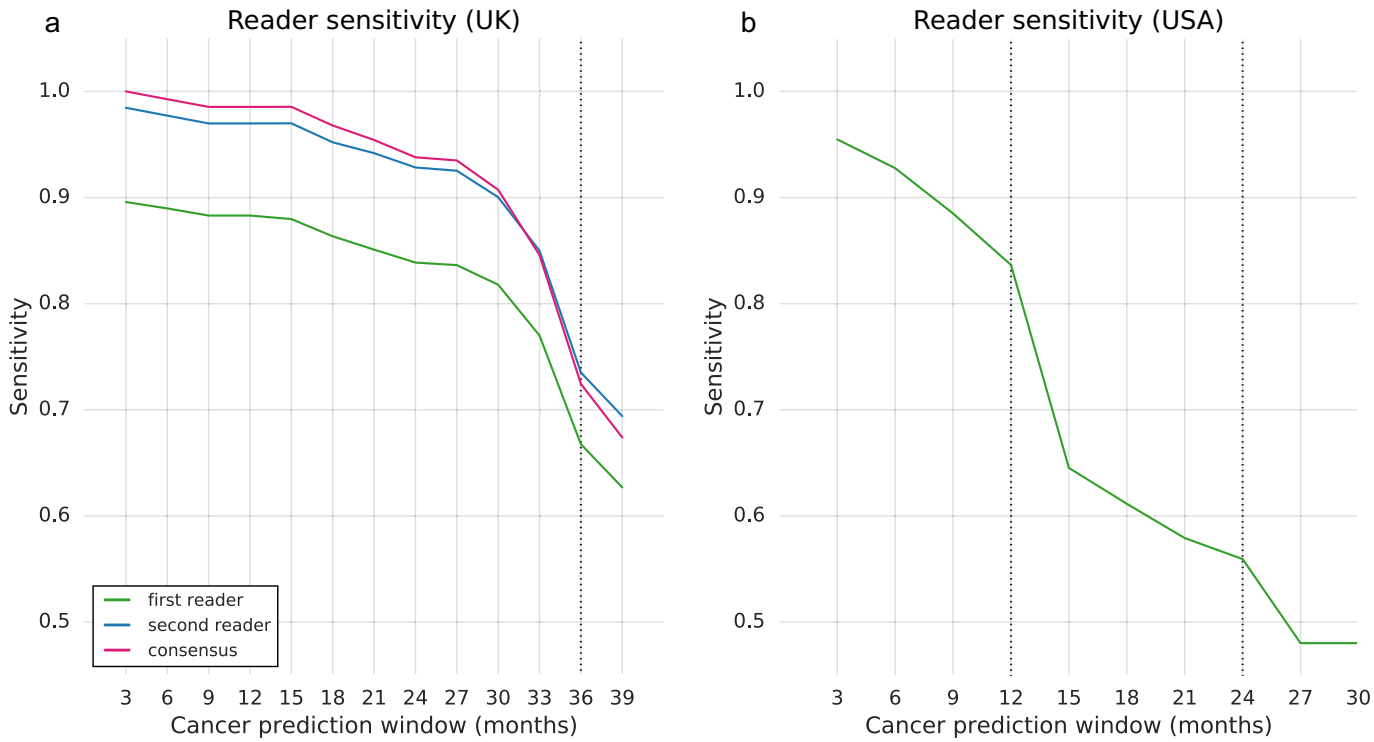
Extended Data Fig. 1 | Unweighted evaluation of breast cancer prediction on the US test set. In contrast to in Fig. 2b, the sensitivity and specificity were computed without the use of inverse probability weights to account for the spectrum enrichment of the study population. Because hard negatives are 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.



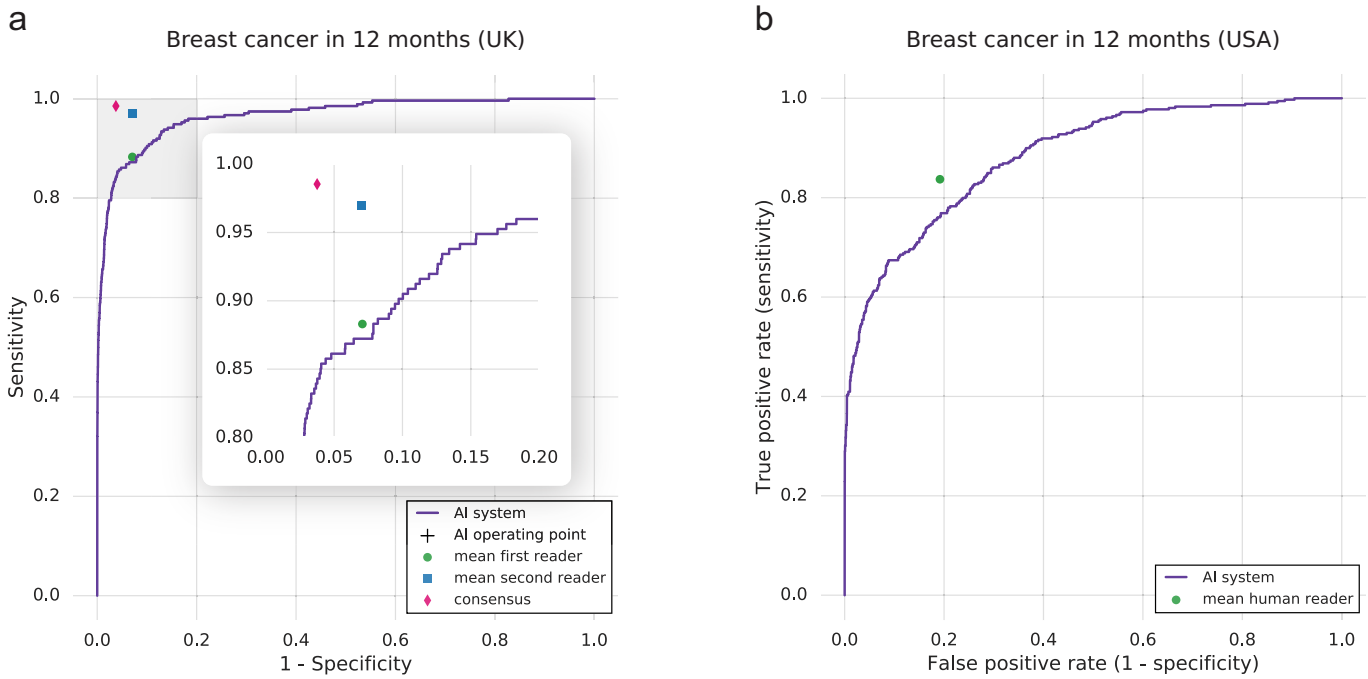
Extended Data Fig. 2 | Performance of the AI system in breast cancer prediction compared to six independent readers, with a 12-month follow-up interval for cancer-positive status. Whereas the mean reader AUC was 0.750 (s.d. 0.049), the AI system achieved an AUC of 0.871 (95% CI 0.785, 0.919). The AI system exceeded human performance by a significant margin (ΔAUC = +0.121, 95% CI 0.070, 0.173; $P = 0.0018$ by 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 who had received a negative biopsy result ($n = 119$), making it a more challenging population for screening. As 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 Fig. 3a for a comparison with a time interval that was chosen to encompass a subsequent screening exam.



Extended Data Fig. 3 | Localization (mLROC) analysis. Similar to Extended Data Fig. 2, but true positives require localization of a malignancy in any of the four mammogram views (see Methods section 'Localization 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 pAUC calculation. The mean reader pAUC was 0.029 (s.d. 0.005), whereas that of the AI system was 0.048 (95% CI 0.035, 0.061). The AI system exceeded human performance by a significant margin (ΔpAUC = +0.0192, 95% CI 0.0086, 0.0298; $P = 0.0004$ by two-sided ORH method).



Extended Data Fig. 4 | Evidence for the gatekeeper effect in retrospective datasets. a, **b**, Graphs show the change in observed reader sensitivity in the UK (**a**) and the USA (**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 the initial screening. However, steeper drops occur when the follow-up window encompasses the screening interval (36 months in the UK; 12 and 24 months in the USA). This is suggestive of what happens to reader metrics when gatekeeper bias is mitigated by another screening examination. In both graphs, the number of positives grows as the follow-up interval is extended. In the UK dataset (**a**), it increases from $n = 259$ within 3 months to $n = 402$ within 39 months. In the US dataset (**b**), it increases from $n = 221$ within $n = 3$ months to 553 within 39 months.



c

test dataset	human benchmark	metric	clinical decision (%)	*n*
UK	first reader	sensitivity	88.3	265
		specificity	92.9	25,116
	second reader	sensitivity	97.0	265
		specificity	93.0	25,114
	consensus	sensitivity	98.5	274
		specificity	96.2	25,443
USA	reader	sensitivity	83.7	359
		specificity	80.8	2,185

Extended Data Fig. 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 skewed by the gatekeeper effect (Extended Data Fig. 4). See Fig. 2 for comparison with longer time intervals. **a**, Performance of the AI system 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 (95% CI 0.954, 0.977). **b**, Performance of the AI system 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 (95% CI 0.859, 0.903). **c**, Reader performance. When computing reader metrics, we excluded cases for which the reader recommended repeat mammography to address technical issues. In the US data, the performance of radiologists could only be assessed on the subset of cases for which a BI-RADS grade was available.

Extended Data Table 1 | Characteristics of the UK and US test sets

a

		UK test set	CI at 95%	NHS BSP
	Years	2012 to 2015	-	2011 to 2016
	Sources	2 sites in the UK	-	All UK screening sites
	No. women	25,856	-	10,257,551
	No. normals	25,588 (99.0%)	(98.8, 99.1)	10,171,074 (99.1%)
	No. cancers	268 (1.0%)	(0.9, 1.2)	86,477 (0.8%)
	Recall rate	1,235 (4.8%)	(4.5, 5.1)	427,457 (4.2%)
Age	*45 – 49*	1,707 (6.6%)	(6.2, 7.1)	832,883 (8.1%)
	50 – 52	4,399 (17.1%)	(16.4, 17.7)	1,487,366 (14.5%)
	53 – 54	2,742 (10.6%)	(10.1, 11.1)	944,823 (9.2%)
	55 – 59	6,034 (23.3%)	(22.6, 24.0)	2,139,701 (20.9%)
	60 – 64	5,457 (21.1%)	(20.4, 21.8)	2,044,746 (19.9%)
	65 – 70	4,575 (17.7%)	(17.0, 18.3)	2,217,947 (21.6%)
	>= 70	942 (3.6%)	(3.3, 4.0)	590,085 (5.8%)
Cancer type	*Invasive*	204 (76.1%)	(69.5, 81.8)	68,006 (78.6%)
	Non-invasive	58 (21.6%)	(16.2, 28.1)	17,733 (20.5%)
	Micro-invasive	-	-	654 (0.8%)
	Unknown	6 (2.2%)	(0.9, 5.6)	84 (0.1%)
Cancer size (Invasive only)	*< 10mm*	41 (20.1%)	(13.7, 28.3)	17,242 (25.4%)
	10 – 15mm	44 (21.6%)	(15.3, 30.4)	17,745 (26.1%)
	15 – 20mm	39 (19.1%)	(12.9, 27.2)	12,864 (18.9%)
	20 – 50mm	61 (29.9%)	(22.1, 38.7)	16,316 (24.0%)
	>= 50mm	13 (6.4%)	(3.1, 12.4)	1,527 (2.3%)
	Unknown	6 (2.9%)	(1.0, 7.9)	2,312 (3.4%)

b

		US test set	CI at 95%	US BCSC
	Years	2001 to 2018	-	2007 to 2013
	Sources	1 US medical center	-	6 BCSC registries
	No. women	3,097	-	1,682,504
	No. normals	2,738 (88.4%)	(87.2, 89.8)	1,672,692 (99.4%)
	No. cancers	359 (11.6%)	(10.2, 12.8)	9,812 (0.6%)
	Recall rate	929 (30.0%)	(18.4, 21.5)	195,170 (11.6%)
Age	*< 40*	181 (5.9%)	(4.8, 7.1)	41,479 (2.5%)
	40 – 49	1,259 (40.8%)	(38.6, 43.0)	448,587 (26.7%)
	50 – 59	800 (26.1%)	(24.1, 28.1)	505,816 (30.1%)
	60 – 69	598 (19.0%)	(17.3, 20.9)	396,943 (23.6%)
	>= 70	259 (8.2%)	(7.0, 9.5)	289,679 (17.3%)
Cancer type	*Invasive*	240 (66.9%)	(60.5, 72.1)	5,885 (69.0%)
	DCIS	100 (27.9%)	(22.8, 33.9)	2,644 (31.0%)
	Other	19 (5.3%)	(3.2, 8.9)	-

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 NHSBSP[65]. For comparison with national numbers, only cancers that were detected by screening are reported here. **b**, The US test set was drawn from one academic medical centre 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.
DCIS, ductal carcinoma in situ.

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

a

test dataset	human benchmark	metric	clinical decision (%)	AI decision (%)	Δ (%)	95% CI (%)	p-value	comparison	N
UK	first reader	sensitivity	62.69	65.42	2.70	(-3.0, 8.5)	0.0043	noninferiority	402
		specificity	92.93	94.12	**1.18**	(0.29, 2.08)	0.0096	superiority	25,115
	second reader	sensitivity	69.40	69.40	0.00	(-4.89, 4.89)	0.0225	noninferiority	402
		specificity	92.97	92.13	-0.84	(-1.97, 0.282)	2e-13	noninferiority	25,113
	consensus	sensitivity	67.39	68.12	0.72	(-3.49, 4.94)	0.0039	noninferiority	414
		specificity	96.24	96.24	-3.35	(-4.06, -2.63)	3e-6	noninferiority	25,442
USA	reader	sensitivity	48.10	57.50	**9.40**	(4.45, 13.85)	0.0004	superiority	553
		specificity	80.83	86.53	**5.70**	(2.62, 8.64)	0.0002	superiority	2,185

b

USA	reader	sensitivity	48.10	56.24	**8.14**	(3.54, 12.5)	0.0006	superiority	553
		specificity	80.83	84.29	**3.47**	(0.6, 5.98)	0.0212	superiority	2,185

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, on the basis of separate validation data, to match or exceed the performance of each human benchmark (see Methods section 'Selection of operating points'). These points are depicted graphically in Fig. 2. Note that the number of cases (*N*) differs from Fig. 2 because the opinion of the radiologist was not available for all images. We also note that sensitivity and specificity metrics are not easily comparable to most previous publications in breast imaging (for example, the DMIST Trial[34]), given the differences in follow-up interval. Negative cases in the US dataset were upweighted to account for the sampling protocol (see Methods section 'Inverse probability weighting'). **b**, Same columns as **a**, but using a version of the AI system that was trained exclusively on the UK dataset. It was tested on the US dataset to show generalizability of the AI across 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. Non-inferiority 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 (see Methods section 'Statistical analysis'). Quantities in bold represent estimated differences that are statistically significant for superiority; all others are statistically non-inferior at a pre-specified 5% margin.

Extended Data Table 3 | Detailed description of the case composition for the reader study

Row	Description	No. cancer cases	No. biopsied negative cases	No. normal cases	Total	Figure
1	inclusion based on 27-month outcome	125	125	250	500	-
2	manual quality review	113	119	233	465	Figure 3a,b
3	restrict to cancers in 12 months	56	119	233	408	Figure 3c Extended Data Figure 2
4	obtain ground truth localizations	53	119	233	405	Extended Data Figure 3

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 Fig. 3a, b. Row 3: for further analysis, we restricted the cancers to those that developed within 12 months. Cases in which cancer developed 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 Fig. 3c and Extended Data Fig. 2. Row 4: to perform localization 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 Fig. 3.

Extended Data Table 4 | Potential use of the AI system in two clinical applications

a

	Sensitivity (%) (*n* = 414)	Specificity (%) (*n* = 25,422)	Simulated reduction of second reader workload (%)
AI as second reader (UK)	66.66	96.26	87.98
Existing workflow (UK)	67.39	96.24	-
95% CI on the difference	(-2.68, 1.23)	(-0.13, 0.17)	-

b

Triage status	Dataset	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)	Reliability of triage decision (%) (95% CI)
Negative	UK	99.63 (98.88, 100.0) *n* = 274	41.15 (40.57, 41.72) *n* = 25,443	99.99 (NPV) (99.97, 100.0) *n* = 10,471
	USA	98.05 (96.12, 99.16) *n* = 359	34.79 (31.97, 37.60) *n* = 2,411	99.90 (NPV) (99.83, 99.96) *n* = 720
Positive	UK	41.24 (35.63, 47.08) *n* = 274	99.92 (99.89, 99.95) *n* = 25,443	85.69 (PPV) (79.66, 90.98) *n* = 132
	USA	29.80 (25.21, 34.45) *n* = 359	99.90 (99.78, 99.97) *n* = 2,411	82.41 (PPV) (65.38, 94.71) *n* = 121

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 effort of the second reader can be eliminated while maintaining the standard of care that is produced by double reading. The decision of the AI system was generated using the first reader operating point (i) shown in Fig. 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

Dataset	Cancer type	AI caught, reader missed	Reader caught, AI missed
UK	*Invasive*	31	20
	In situ	7	12
	Unknown	7	2
USA	*ILC or IDC*	83	37
	DCIS	31	27
	Other	7	5

Invasive cancer grade (UK only)	AI caught, reader missed	Reader caught, AI missed
Grade 1	10	4
Grade 2	15	13
Grade 3	6	3

Invasive primary tumour size (UK only)	AI caught, reader missed	Reader caught, AI missed
< 10mm	4	6
10 – 15mm	6	7
15 – 20mm	5	2
20 – 50mm	14	4
>= 50mm	2	1

For the UK comparison, we used the first reader operating point (i) shown in Fig. 2a. For the US comparison, we used the operating point shown in Fig. 2b. ILC, invasive lobular carcinoma; IDC, invasive ductal carcinoma; DCIS, ductal carcinoma in situ.

Extended Data Table 6 | Performance breakdown

a

Cancer type (UK first reader)			AI system	Reader	Delta (95% CI)	No. examples
Sensitivity	Invasive grade	Grade 1	81.94	73.61	8.33 (-4.71, 21.38)	72
		Grade 2	63.87	62.58	1.29 (-6.60, 9.15)	155
		Grade 4	69.36	64.52	4.84 (-3.66, 13.34)	62
		Grade unknown	25	25	-	8
	In situ grade	High grade	58.97	53.85	5.13 (-14.19, 24.45)	39
		Intermediate grade	25	75	-50.00 (-100.00, 14.82)	8
		Low grade	56	64	-8.00 (-24.194, 8.19)	25
		Grade unknown	69.23	76.92	-7.69 (-35.08, 19.70)	13
	Primary tumour size (invasive only)	< 10mm	61.81	65.46	-3.64 (-14.86, 7.59)	55
		10 – 15mm	72.73	74.55	-1.82 (-14.66, 11.02)	55
		15 – 20mm	71.42	66.07	5.36 (-3.80, 14.51)	56
		20 – 50mm	67.3	57.43	9.90 (1.90, 17.90)	101
		>= 50mm	88.24	82.35	5.88 (-13.89, 25.65)	17

b

Cancer type (US clinical radiologist)		AI system	Reader	Delta (95% CI)	No. examples
Sensitivity	ILC or IDC	57.97	45.33	12.63 (6.88, 18.39)	364
	DCIS	57.05	54.6	2.45 (-6.70, 11.60)	163
	Other	53.85	46.15	7.69 (-18.25, 33.64)	26

c

Breast density (US clinical radiologist)		AI system	Reader	Delta (95% CI)	No. examples
Sensitivity	Entirely fatty	53.84	48.71	5.12 (-12.21, 22.46)	39
	Scattered fibroglandular densities	60.41	49.58	10.8 (3.39, 18.28)	240
	Heterogeneously dense	56.11	48.1	8.01 (0.93, 15.11)	237
	Extremely dense	16.67	25	-8.33 (-44.55, 27.88)	12
	Unknown	66.67	66.67	0.00 (-92.39, 92.39)	5
Adjusted specificity	Entirely fatty	90.6	82.88	7.72 (-1.24, 17.40)	6
	Scattered fibroglandular densities	86.78	80.75	6.03 (1.57, 10.42)	149
	Heterogeneously dense	85.65	80.55	5.09 (0.76, 9.74)	831
	Extremely dense	92.18	77.1	15.07 (-1.90, 33.74)	1,061
	Unknown	95.34	93.01	2.33 (-25.36, 57.62)	73
Specificity	Entirely fatty	85.23	77.85	7.38 (-0.08, 14.85)	6
	Scattered fibroglandular densities	80.75	71	9.74 (5.92, 13.57)	149
	Heterogeneously dense	80.21	67.39	12.82 (9.38, 16.26)	831
	Extremely dense	86.3	75.34	10.96 (-2.50, 24.42)	1,061
	Unknown	66.67	50	16.67 (-38.32, 71.65)	73

The 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 first reader operating point (i) shown in Fig. 2a. Also shown is the performance of the first reader on the same subset. **b**, Sensitivity across cancer subtypes in the US data. We used the operating point shown in Fig. 2b. 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 breast density was extracted from the radiology report rendered at the time of screening, which was only available in the US dataset. We used the operating point shown in Fig. 2b. Adjusted specificities were computed using inverse probability weighting (Methods).

Extended Data Table 7 | Reader experience

a

UK	
Reads per year	**No. readers**
3,000-4,000	3
4,000-5,000	6
5,000-6,000	3
6,000-7,000	1
7,000-8,000	2
8,000+	3
Unknown	33

Years of experience	**No. readers**
5-10	4
10-15	5
15-20	4
20+	5
Unknown	33

Job title	**No. readers**
Consultant Radiologist	8
Consultant Radiographer	6
Advanced Practitioner Radiographer	4
Unknown	33

b

US reader study			
	Reads per year	**Years of experience**	**Fellowship trained**
Reader 1	5,500	12	Yes
Reader 2	4,000	7	No
Reader 3	2,000	4	No
Reader 4	3,000	12	No
Reader 5	3,500	15	Yes
Reader 6	2,500	10	No

a, Detailed information was available for 18 of the 51 readers represented in UK the test set. Reads were performed as part of routine practice and so reflect the standard of care in the UK screening programme. **b**, Experience levels of the six radiologists involved in the US reader study.

nature research

Corresponding author(s): Shravya Shetty

Last updated by author(s): Oct 23, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Dicom files were handled with the open source libraries DCMTK (https://support.dcmtk.org/docs/, version 3.6.1_20160630) and Pydicom (https://pydicom.github.io/, version v1.2.0).
Data analysis	The code used for training deep learning 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 Methods section to allow independent replication with non-proprietary libraries. Several major components of our work are available in open source repositories including Tensorflow (https://www.tensorflow.org, version 1.14.0) and the Tensorflow Object Detection API (https://github.com/tensorflow/models/tree/master/research/object_detection; Oct 15th, 2019 release). Data analysis was conducted in Python using the numpy (version v1.16.4), scipy (version 1.2.1), and scikit-learn (version 0.20.4) packages.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on 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/.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The UK 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. Women from the US cohort were split randomly between train (55%), validation (15%) and test (30%). This scheme follows machine learning convention, but errs on the side of a larger test set to power statistical comparisons and include a more representative population. The size of the reader study was selected due to time and budgetary constraints. The case list was composed of 250 negative exams, 125 biopsy-confirmed negative exams and 125 biopsy-confirmed positive exams. We sought to include sufficient positives to power statistical comparisons on the metric of sensitivity, while avoiding undue enrichment of the case mixture. Biopsy-confirmed negatives were included to make the malignancy discrimination task more difficult.
Data exclusions	UK Dataset 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 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 no exam with sufficient follow-up were excluded from validation and test. This last step resulted in the exclusion of 5,990 of 31,766 test set cases (19%). 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. 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. US 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 population of unbiopsied negative women were sampled. After de-identification 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. 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.
Replication	All attempts at replication were successful. Comparisons between AI system and human performance revealed consistent trends across three settings: a UK clinical environment, a US clinical environment, and an independent, laboratory-based reader study. Our findings persisted through numerous retrainings with random network initialization and training data iteration order. Remarkably, our findings on the US test set replicated even when we trained the AI system solely on UK data.
Randomization	Patients were randomized into training, validation, and test sets by applying a hash function to the deidentified medical record number. Assignment to each set was made based on the value of the resulting integer modulo 100. For the UK data, values of 0-9 were reserved for the test set. For the US data, values of 0-29 were reserved for the test set.
Blinding	The US and UK test sets were held back from AI system development, which only took place on the training and validation sets. Investigators did not access test set data until models, hyperparameters, and thresholds were finalized. None of the readers who interpreted the images (either in the course of clinical practice or in the context of the reader study) had knowledge of any aspect of the AI system.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	The focus of the paper is on breast cancer screening, so all individuals in the population were women from the screening populations in the US and UK. 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. Specifically, there were 25,856 women in the test set, of which 268 (1%) had breast cancer detected during screening. For many cancers in the test set, additional metadata was available. There was a rich collection of both invasive (76.1%) and non-invasive cancers (21.6%). The invasiveness of 2.2% of cancers was unknown. These cancers had a lesion size of less than 10mm to lesions greater than 50mm. 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. For all the cancers in the test set, additional metadata was available. For example, 66.9% of the cancers were invasive, 27.9% were DCIS and the rest were of an other cancer subtype.
Recruitment	Patient data were gathered retrospectively from screening practices in the UK and US. As such, they reflect natural screening populations at the sites under study. Self-selection biases associated with the choice to enroll in screening may be present, but are likely to be representative of the real-world patient population. In the UK, 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. Specifically, for this paper, the data was initially compiled by OPTIMAM, a Cancer Research UK effort, from three between the years of 2010 and 2018: 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 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 the US, the American College of Radiology, the American Cancer Society, and the US Preventive Services Task Force recommends screening every 1 or 2 years for women starting at age 40 or 50. The various US guidelines are summarized at https://www.acraccreditation.org/mammography-saves-lives/guidelines. Our US dataset was collected from Northwestern Memorial Hospital (Chicago, IL) between the years of 2001 and 2018. 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.
Ethics oversight	Use of the UK dataset for research collaborations by both commercial and non-commercial organisations received ethical approval (Research Ethics Committee reference 14/SC/0258). The US data was fully de-identified and released only after an Institutional Review Board approval (STU00206925).

Note that full information on the approval of the study protocol must also be provided in the manuscript.