nature

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

## Supplementary information

# International evaluation of an AI system for breast cancer screening

In the format provided by the authors and unedited

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

## Supplementary Methods

**Region-of-interest (ROI) annotations.** In addition to the case-level cancer labels described in Methods, we used region-of-interest (ROI) annotations from three sources for algorithm development. First, the UK dataset included 8,277 ROIs on 7,672 images from 3,871 women annotated by OPTIMAM with rectangular ROIs indicating the location of subsequent biopsies. Secondly, using our own web-based platform, we collected additional ROI annotations from patients with subsequent cancer diagnoses on the UK and US datasets. For each image, a US-board-certified mammographer was asked to draw a rectangular ROI around the area(s) they suspected to contain the primary tumor based on available metadata. Through this process, we collected 2,156 ROIs on 2,145 images from 892 women from the UK data and 3,549 ROIs on 1,917 images from 694 women from the US data. Finally, we supplemented the training with 2,073 ROI annotations on 1,939 images from 1,076 women from the publicly available CBIS-DDSM dataset[1].

**Overview of the AI system.** The AI system presented here consists of an ensemble of three deep learning models, each operating on a different level of analysis. 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. Schematics of the model architectures are shown in Supplementary Figure 3.

All models were trained with data augmentation applied to each image. Random transformations included elastic deformation, shearing, rescaling, translation, and flipping. Predictions from each model are the average of those resulting from 3 stochastic training runs. All models were implemented in the Tensorflow library. When possible, neural network parameters were initialized with values derived from ImageNet pretraining. Model training took place on dedicated machine learning hardware: Google Tensor Processing Units.

**Lesion model.** The lesion model is a two-stage architecture focusing on individual lesions suggestive of cancer. In the first stage, a detector identifies suspicious regions in all four images. Next, each region is fed to a classifier to produce a lesion-level score. Lesion-level cancer risk scores are then accumulated to produce a case-level score.

Suspicious regions were identified using a RetinaNet[2] object detection mode applied to individual images. The same model was used for all four mammography views. The model was trained to identify breast tissue representing biopsy-confirmed cancer using the ROI annotations described above. Inputs to the model were full screening images, first center-cropped to 4096x4096 then downsampled to 2048x2048.

For each image, the object detection model produced a collection of rectangular bounding boxes with associated confidence scores. Regions of size 512x512 were extracted from the 10

locations with the highest scores among all four mammogram images. These regions, along with a reference region drawn from the contralateral breast were resampled to 409x409 and fed to a MobileNetV2[3] shared feature extractor. Locating the reference patch required a coarse registration of the two images in order to identify the corresponding region. An encoding of the image view, laterality and detection coordinates, as well as the patient's age, were concatenated with the image features for binary classification. The classifier shared parameters across regions.

Each of the 5 patches were augmented independently. Each patch was 409x409 in size, and was augmented with random horizontal and vertical flips, random elastic deformation, crop and shears applied.

The malignancy predictions for each crop are computed independently and then combined into a case-level score using the noisy-OR operation. This operation models the likelihood that a case has a malignancy as the complement of the probability that all individual lesions are not malignant, assuming independence among the lesions. This enables us to train lesion-level malignancy scores using case-level labels.

The second-stage model was trained with inputs generated by a fixed detection model, with supervision from case-level cancer labels. Although 10 crops were used for inference, at training time 5 crops were randomly sampled with probability proportional to their detection scores. A focal loss function was optimized with stochastic gradient descent with momentum using mini-batches of size 4. The learning rate was modulated from its initial value using cosine decay. Mini-batches were constructed to contain positive and negative examples in equal proportion. The UK and US datasets were pooled for initial training, and dataset-specific models were subsequently fine-tuned for each dataset. (However, for the generalisability experiment, only UK data was used in training.) The parameters that achieved maximum validation AUC were selected for test set inference. When running on the test set, examples were randomly perturbed using the same augmentations as during training; the final predictions were the average of 500 runs.

**Breast model.** Images were resized to the same pixel width and height, random elastic deformation, crop, shearing and rotations applied and the right breast images flipped horizontally. The breast model applies a ResNet-v2-50[4] (‘Image Feature Extractor’) with shared weights to each image (resized to 4096 x 3328) in the study. The spatial feature vectors resulting from the Image Feature Extractor are concatenated per-breast (RMLO with RCC, LMLO with LCC).

After concatenation, an additional neural network is applied to predict breast-level cancer score. This consists of 4 residual blocks with bottleneck layers with a spatial downscaling of 4, then 4 residual blocks with downscaling of 2, and then a final 4 residual blocks with downscaling 2. Finally 1x1 convolutions, followed by average pooling was applied to reach a per breast

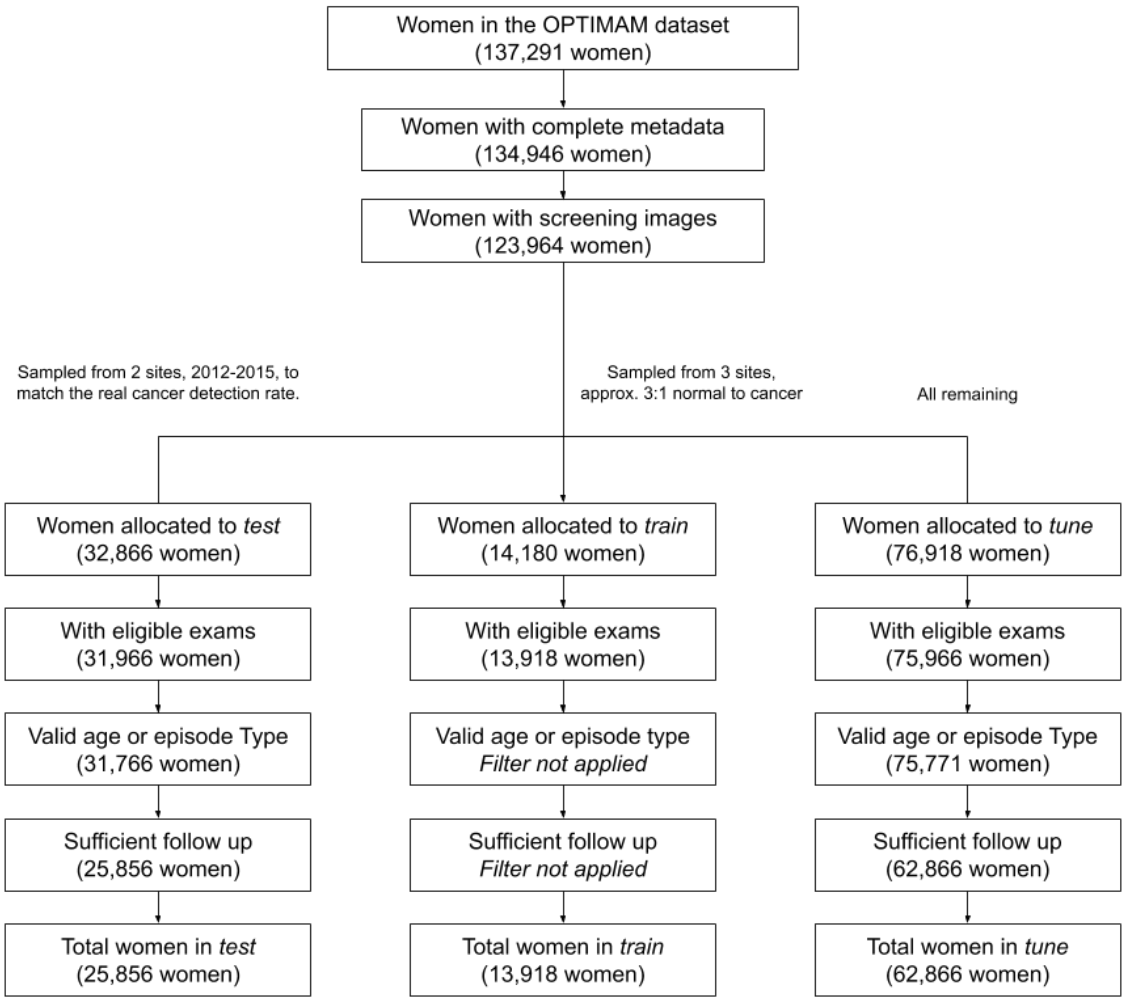
probability score. The weights between breast feature extractors were shared. A case level score was generated by taking the max of the per-breast classifications.

The model was trained end-to-end using model parallelism for 120,000 steps using the Adam optimizer. The learning rate was initially 1e-4 and divided by 2 every 1,875 steps.

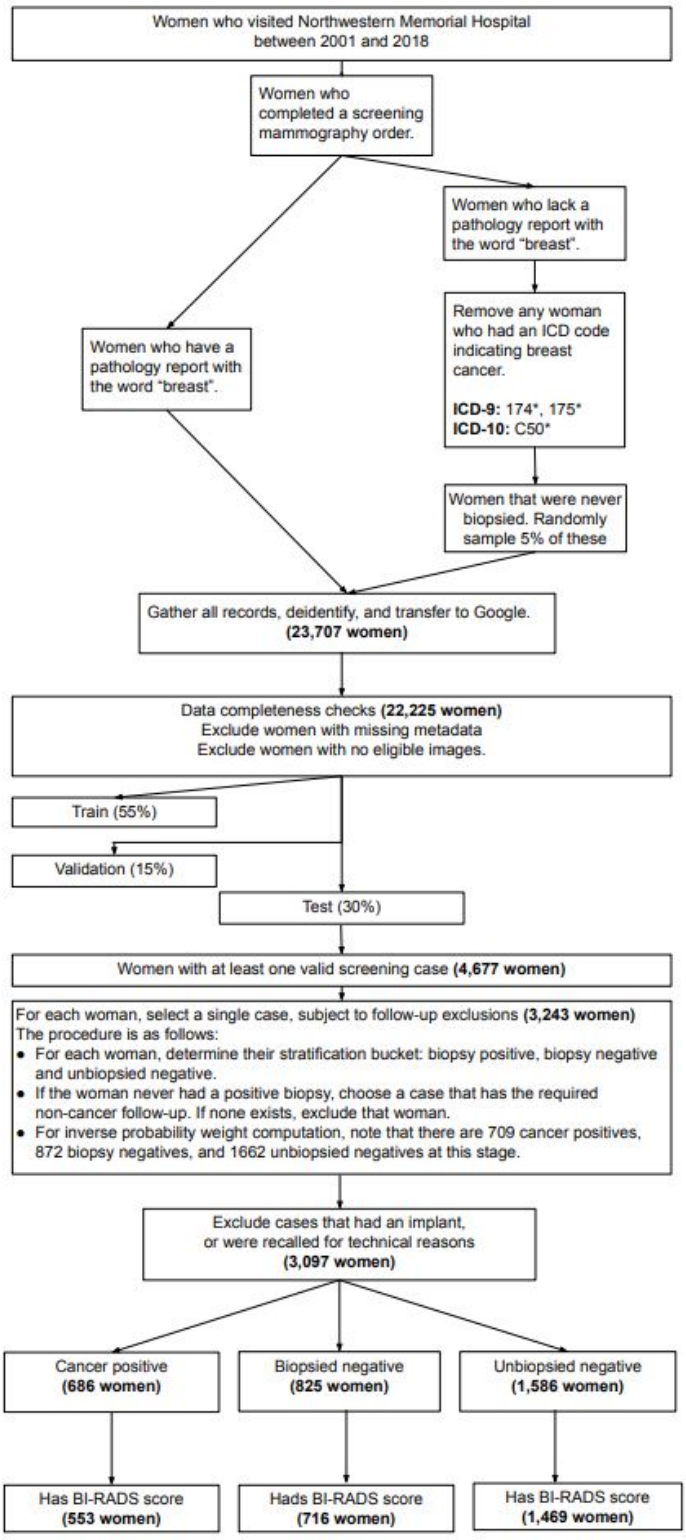
**Case model.** For the case model, each image was first cropped to 4096x4096 and then resized to 2048x2048, and was augmented with random horizontal and vertical flips, random elastic deformation, crop and shears applied. The two images corresponding to each breast were also randomly swapped with regards to their position in the model.

This case-level model applies a ResNet-v1-50[5] feature extractor with shared weights to each image. Feature vectors from each of the four images in the case were concatenated and then fed through a hidden layer of size 512 before binary classification. The UK and US datasets were pooled for training. However, for the generalisability experiment, only UK data was used. Unlike the lesion model, no distinct fine-tuning phases took place, but checkpoints that achieved maximum validation AUC were chosen for each dataset from a single training run. The parameters of the ResNet model were initialized from the backbone of the object detection model used to generate input for the Lesion Model. A focal loss function was optimized with stochastic gradient descent with momentum using mini-batches of size 2. The learning rate was modulated from its initial value using cosine decay. Mini-batches were constructed to contain positive and negative examples in equal proportion. When running on the test set, examples were randomly perturbed using the same augmentations as during training; the final predictions were the average of 500 runs.

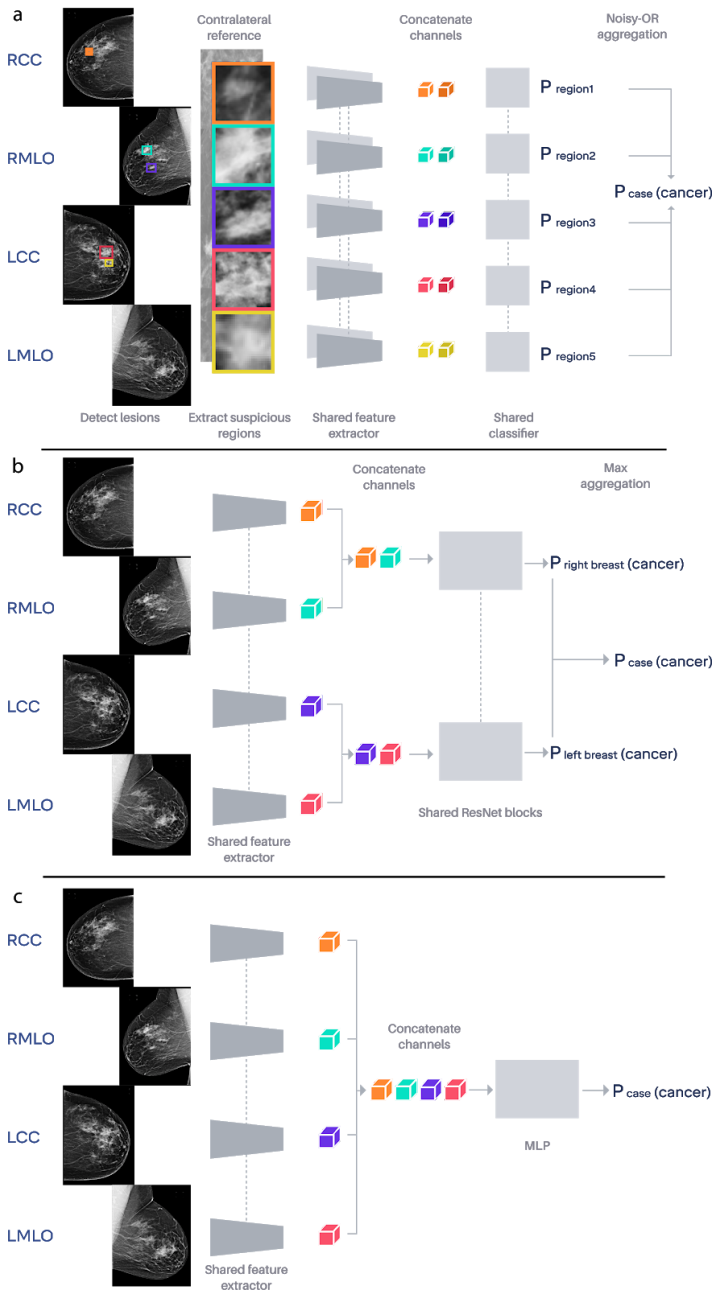
## Supplementary Figures



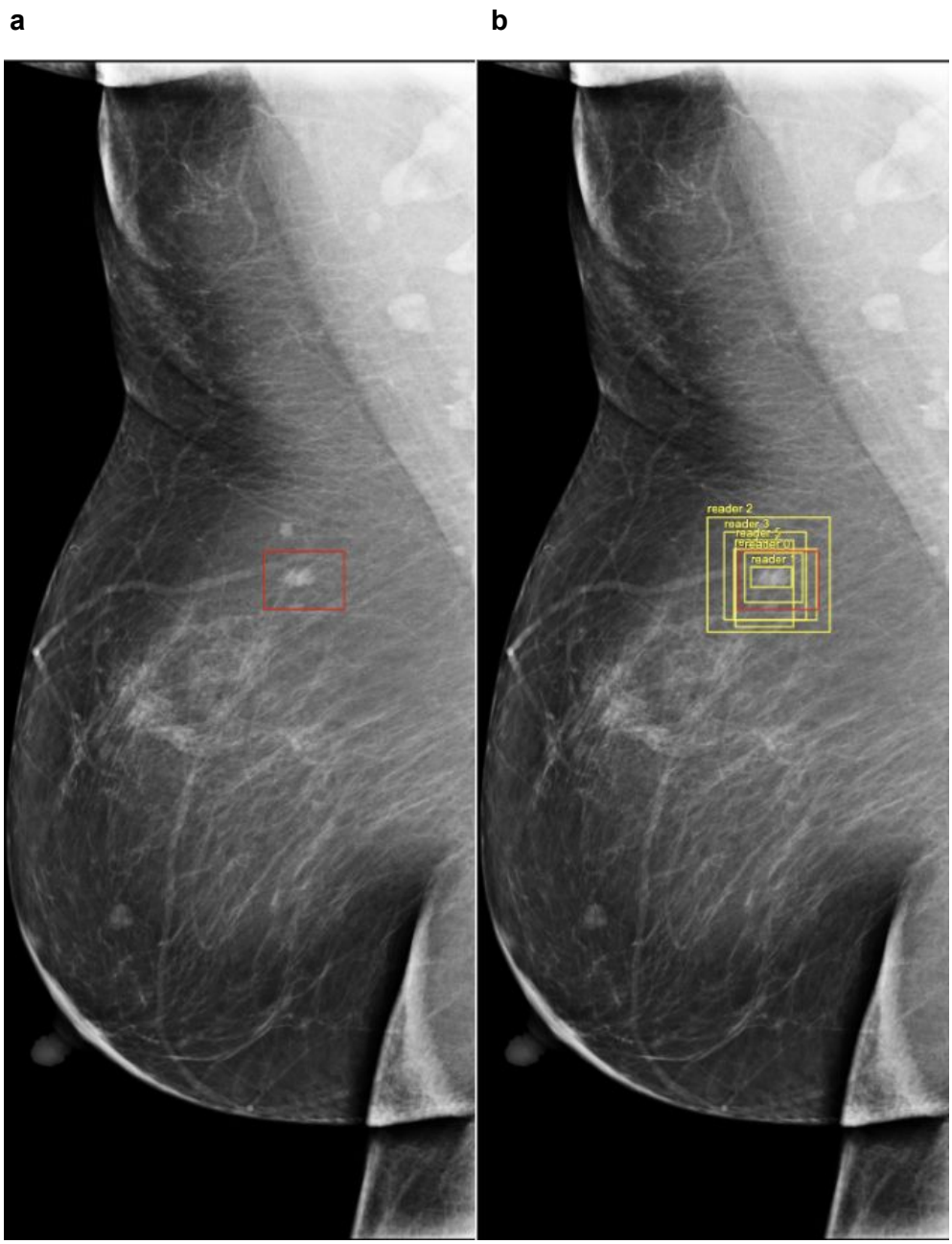

**Supplementary Figure 1 | STARD diagram describing the inclusion/exclusion criteria for the UK dataset.**



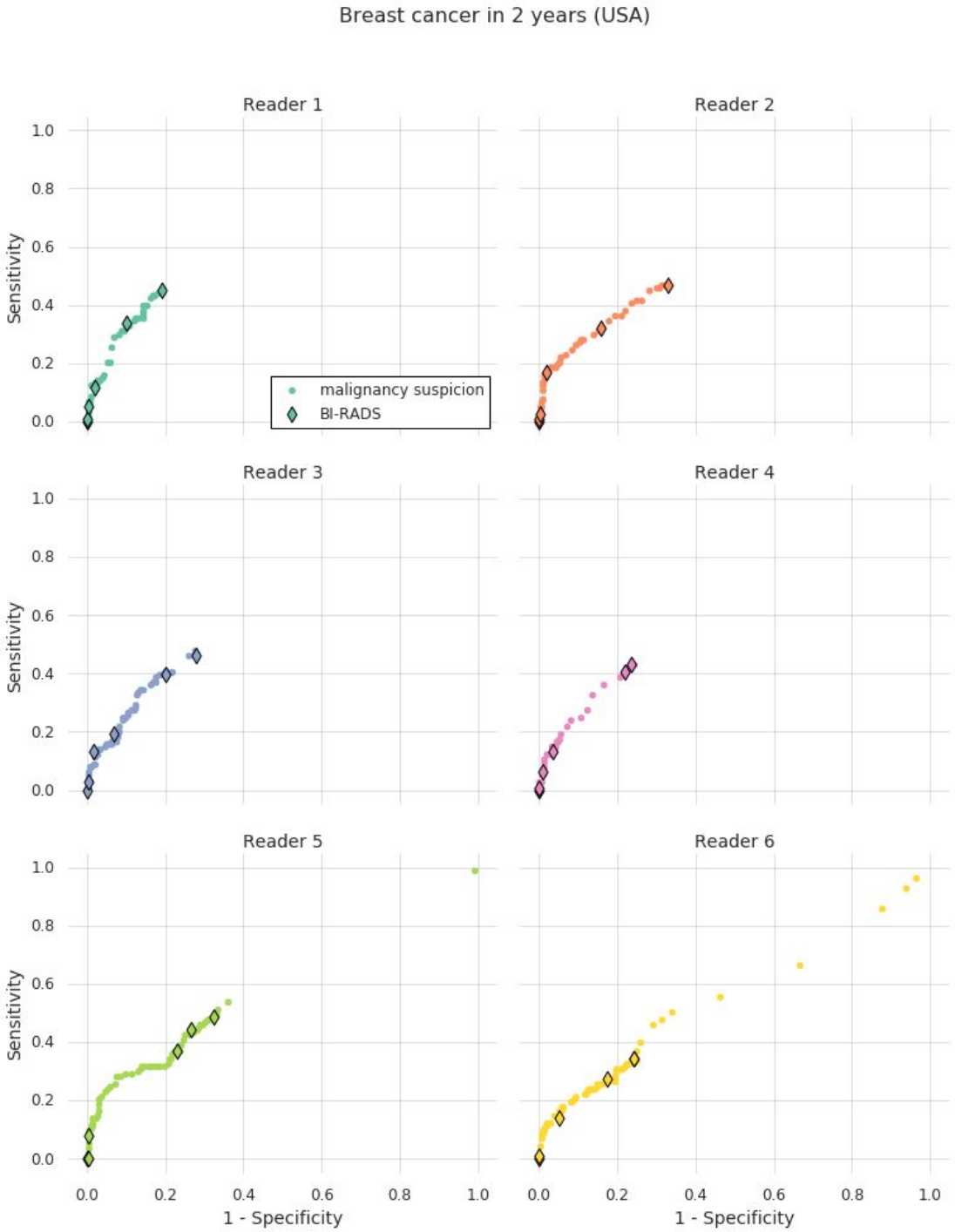

**Supplementary Figure 2 | STARD diagram describing the inclusion/exclusion criteria for the US dataset.**




**Supplementary Figure 3 | Deep learning architectures. a**, The lesion model starts by applying a detector to identify suspicious regions in all 4 mammogram views. The top-scoring regions are extracted and sent through a shared feature extractor along with a corresponding contralateral region. The image features are concatenated and sent through a shared classifier. The classifier outputs across 10 regions are finally aggregated using a noisy-OR operation to produce the case-level score. For illustration purposes, we depict 5 regions here, but 10 are used for the final inference. **b**, The breast model applies a shared feature extractor for each breast independently to generate four image features. The two features for each breast are concatenated and sent to a shared classifier. The outputs of the classifier for both breasts are aggregated using the max operation to produce the case-level score. **c**, The case model applies a shared feature extractor across all 4 image views. Flattened features are concatenated and then sent to a classifier to produce the case-level score.




**Supplementary Figure 4 | Example reader localizations. a**, One biopsy-confirmed malignant lesion including faint spiculation, highlighted in red, on a right MLO view. **b**, Corresponding annotations provided by six radiologists in the reader study, shown in yellow. The lesion was correctly identified by all six readers, but with varying degrees of precision. Their bounding box annotations overlap with the ground truth box with intersection-over-union (IoU) scores of 0.626, 0.176, 0.331, 0.500, 0.767, and 0.496, respectively. We chose an IoU threshold of 0.1 to account for this variation (see Methods section, 'Localization analysis').




**Supplementary Figure 5 | Correspondence between the readers' malignancy suspicion scores and their BI-RADS ratings.** In the reader study, radiologists' malignancy suspicion scores (0-100) strongly aligned with their BI-RADS ratings (a 6-point scale). Since BI-RADS is used in actual screening practice, we elected to focus our analysis on these ratings.

## Supplementary Tables

**a**

| UK positives 39 months | | First reader | |
|---|---|---|---|
| | | TP | FN |
| AI system | TP | 220 | 45* |
| | FN | 34* | 108 |

| UK negatives 39 months | | First reader | |
|---|---|---|---|
| | | TN | FP |
| AI system | TN | 22,165 | 1,439 |
| | FP | 1,175 | 336 |

| UK positives 39 months | | Consensus | |
|---|---|---|---|
| | | TP | FN |
| AI system | TP | 248 | 40 |
| | FN | 31 | 95 |

| UK negatives 39 months | | Consensus | |
|---|---|---|---|
| | | TN | FP |
| AI system | TN | 22,683 | 713 |
| | FP | 1,803 | 243 |

**b**

| US positives 27 months | | First reader | |
|---|---|---|---|
| | | TP | FN |
| AI system | TP | 197 | 121* |
| | FN | 69* | 166 |

| US negatives 27 months | | First reader | |
|---|---|---|---|
| | | TN | FP |
| AI system | TN | 1,282 | 482 |
| | FP | 240 | 181 |

**Supplementary Table 1 | Confusion matrices for AI system and human readers.** This analysis excludes technical recalls. Asterisk (*) denotes numbers featured in Extended Data Table 5. **a**, Counts for the US dataset. **b**, Counts for the UK dataset. First reader and consensus totals differ due to the exclusion of technical recalls.

| ER status | AI caught, reader missed | Reader caught, AI missed |
|---|---|---|
| *Negative* | 7 | 3 |
| *Positive* | 35 | 22 |
| *Unknown* | 79 | 44 |
| **PR status** | **AI caught, reader missed** | **Reader caught, AI missed** |
| *Negative* | 11 | 6 |
| *Positive* | 30 | 18 |
| *Unknown* | 80 | 45 |
| **HER2 status** | **AI caught, reader missed** | **Reader caught, AI missed** |
| *Negative* | 31 | 18 |
| *Borderline/equivocal* | 0 | 0 |
| *Positive* | 2 | 0 |
| *Unknown* | 88 | 51 |

**Supplementary Table 2 | Disagreement between the AI system and the interpreting clinician based on molecular markers.** The table shows cases missed by the reader by not the AI system, and vice versa, broken down by estrogen receptor (ER), progesterone receptor (PR), and HER2 status. Molecular marker data, extracted from histopathology reports, were only available for the US dataset. Note that for a sizable number of cancers, receptor status was unknown.

| Reader id | No. cases | No. cancers | Sensitivity | | | Specificity | | |
|---|---|---|---|---|---|---|---|---|
| | | | AI system | First reader | Δ 95% CI | AI system | First reader | Δ 95% CI |
| 1 | 1,713 | 28 | 0.75 | 0.86 | (-0.22, 0.01) | 0.93 | 0.94 | (-0.03, 0.00) |
| 2 | 1,622 | 24 | 0.75 | 0.54 | (0.01, 0.41) | 0.93 | 0.94 | (-0.03, 0.01) |
| 3 | 1,542 | 21 | 0.71 | 0.71 | (-0.13, 0.13) | 0.94 | 0.95 | (-0.03, 0.00) |
| 4 | 1,422 | 23 | 0.65 | 0.87 | (-0.39, -0.05) | 0.95 | 0.95 | (-0.02, 0.01) |
| 5 | 1,277 | 12 | 0.83 | 0.75 | (-0.07, 0.24) | 0.94 | 0.94 | (-0.02, 0.01) |
| 6 | 1,195 | 15 | 0.60 | 0.66 | (-0.29, 0.16) | 0.96 | 0.94 | (-0.00, 0.03) |
| 7 | 1,155 | 18 | 0.72 | 0.66 | (-0.19, 0.30) | 0.94 | 0.93 | (-0.02, 0.02) |
| 8 | 1,038 | 5 | 0.80 | 0.60 | (-0.15, 0.55) | 0.94 | 0.95 | (-0.03, 0.01) |
| 9 | 1,037 | 12 | 0.83 | 0.75 | (-0.20, 0.36) | 0.95 | 0.91 | (0.01, 0.06) |
| 10 | 952 | 10 | 0.60 | 0.60 | (-0.28, 0.28) | 0.94 | 0.96 | (-0.05, -0.01) |
| 11 | 888 | 25 | 0.60 | 0.48 | (-0.01, 0.25) | 0.94 | 0.91 | (0.01, 0.06) |
| 12 | 856 | 17 | 0.72 | 0.76 | (-0.26, 0.14) | 0.93 | 0.90 | (0.01, 0.06) |
| 13 | 782 | 16 | 0.56 | 0.38 | (-0.07, 0.45) | 0.96 | 0.93 | (0.01, 0.05) |
| 14 | 732 | 19 | 0.53 | 0.53 | (-0.21, 0.21) | 0.96 | 0.90 | (0.03, 0.08) |
| 15 | 701 | 14 | 0.43 | 0.36 | (-0.17, 0.31) | 0.93 | 0.93 | (-0.03, 0.02) |
| 16 | 630 | 12 | 0.42 | 0.33 | (-0.07, 0.24) | 0.92 | 0.94 | (-0.05, 0.01) |
| 17 | 609 | 6 | 0.83 | 0.50 | (-0.04, 0.71) | 0.93 | 0.91 | (-0.01, 0.05) |
| 18 | 545 | 8 | 1.00 | 1.0 | (0.00, 0.00) | 0.94 | 0.97 | (-0.06, -0.01) |
| 19 | 540 | 9 | 0.78 | 0.89 | (-0.32, 0.10) | 0.93 | 0.90 | (-0.01, 0.10) |
| 20 | 507 | 6 | 1.00 | 1.0 | (0.00, 0.00) | 0.94 | 0.95 | (-0.04, 0.00) |

**Supplementary Table 3 | Comparison with individual clinical readers in the UK dataset.**
We compared the sensitivity and specificity of the AI system to that of the 20 individual readers most represented in the dataset. Each row represents metrics based on the subset of cases interpreted by one reader. Since ground truth cancers occurred within 39 months of examination, the sensitivities appear lower than what is traditionally reported on a 12-month interval. A synopsis of reader experience levels is presented in Extended Data Table 7.

## References


1. Lee, R. S. *et al.* A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci Data* **4**, 170177 (2017).
2. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollar, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* (2018). doi:10.1109/TPAMI.2018.2858826
3. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018). doi:10.1109/cvpr.2018.00474
4. He, K., Zhang, X., Ren, S. & Sun, J. Identity Mappings in Deep Residual Networks. *arXiv [cs.LG]* (2016).
5. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016). doi:10.1109/cvpr.2016.90