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.