nature research

Corresponding author(s): Daniel Tse

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.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	eUnity: FDA-approved fully featured PACS viewer. Used to collect reader study results. MAPLE: Internal labeling tool. Used to collect localization ground truth.
Data analysis	Colab: Internal version of Colab which is an iPython notebook viewer Pandas: Internal fork of open source library Pandas which is a framework for tabular data Matplotlib: Internal fork of open source library Matplotlib which is for making plots sklearn: Internal fork of open source library Scikit-Learn which we used for metrics such as AUC Tensorflow: Internal fork of open source library used to train machine learning models Apache Beam: Internal fork of open source library used for large scale batch processing Tensorflow object detection API: https://github.com/tensorflow/models/tree/master/research/object_detection Inflated Inception: https://github.com/deepmind/kinetics-i3d

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 upon request. 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

We used three datasets which are publicly accessible:

LUNA: https://luna16.grand-challenge.org/data/
LIDC: https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI
NLST: https://biometry.nci.nih.gov/cdas/learn/nlst/images/

The dataset from Northwestern was used under license for the current study, and so is not publicly available. The data, or a test subset, may be available from Northwestern Medicine subject to ethical approvals.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	The first step in determining the sample size was the size of the test set we decided to use for the dataset from the National Lung Cancer Screening Trial (NLST). We had to balance having enough data to train the algorithm while having enough data to validate the algorithm. We used a 70% training (29,541 cases, 401 cancer positive), 15% tuning (6,309 cases, 100 cancer positive), 15% testing (6,729 cases, 87 cancer positive) split which is a standard way of splitting datasets for deep learning research. We believe this sample size was sufficient for the test set because the test set represents all 33 sites in the NLST trial, it contains all 4 stages of cancer, and all CT manufacturers present in the trial. For our independent dataset, the medical institution returned all available cases after NLST publication related to lung cancer screening. We used all cases where we could arrive at a clear conclusion about the cancer outcome. For our reader studies, we used positive enrichment by taking all cases within the test set with a same-year positive cancer diagnosis or biopsy, and then randomly sampling negatives. We believe the sample of negatives was sufficient as it was 5x larger than the number of positives used and we were able to see statistically significant improvements in performance for specificity in both reader studies.
Data exclusions	We excluded data only when it made subsequent analysis not possible: We excluded 3 studies that were not gradable as determined by our readers as there would be no way of making a reader-model comparison since no reader grade was returned. Cases where neither reader found a bounding box suspicious for malignancy in the volume were excluded from the localization analysis since there was no bounding box to compare to. There were a small number of patients in the independent dataset where either there were no images or it was not possible to assess ground truth due to insufficient follow-up, for instance the image was suspicious for cancer but was missing a biopsy confirmation.
Replication	We replicated the high performance of our model on a completely independent dataset from an academic medical center, with different scan parameters, and from a disjoint time period.
Randomization	For NLST, we randomly split patients into the train, tune, or test split. All imaging and metadata from each patient was associated with the same split as the patient. For the reader study, we randomly selected negative cases from the test set. After a random selection of cases we randomly chose one volume from each patient to avoid having the same patient twice in the reader study.
Blinding	We held out the data from the test set and did not give anyone in the research group access to the images until we froze our choice of model and produced the test set results. We have done only one previous evaluation on the test set for an abstract for RSNA-2018 (using a different

model). In that case we only ran the model on the test set once, withholding access otherwise. No one on the model development team has been allowed to inspect the model's performance on the test set at any point.

Reporting for specific materials, systems and methods

Materials & experimental systems

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

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	For NLST, the patient population characteristics are best described in the original NLST publication: The National Lung Screening Trial: Overview and Study Design https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3009383/ For our independent dataset, we included all patients from the center who underwent lung cancer screening.
Recruitment	All participants enrolling in NLST signed an informed consent developed and were approved by the screening centers' institutional review boards (IRBs), the National Cancer Institute (NCI) IRB, and the Westat IRB. Additional details regarding cases in the dataset are available through the National Institutes of Health Cancer Data Access System. The independent dataset was gathered retrospectively under approval from the Northwestern University IRB