LETTERS
https://doi.org/10.1038/s41591-019-0447-x
nature medicine
Corrected: Author Correction

End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography

Diego Ardila[1,5], Atilla P. Kiraly[1,5], Sujeeth Bharadwaj[1,5], Bokyung Choi[1,5], Joshua J. Reicher[2], Lily Peng[1], Daniel Tse[1*], Mozziyar Etemadi[3], Wenxing Ye[1], Greg Corrado[1], David P. Naidich[4] and Shravya Shetty[1]

With an estimated 160,000 deaths in 2018, lung cancer is the most common cause of cancer death in the United States[1]. Lung cancer screening using low-dose computed tomography has been shown to reduce mortality by 20–43% and is now included in US screening guidelines[1–6]. Existing challenges include inter-grader variability and high false-positive and false-negative rates[7–10]. We propose a deep learning algorithm that uses a patient's current and prior computed tomography volumes to predict the risk of lung cancer. Our model achieves a state-of-the-art performance (94.4% area under the curve) on 6,716 National Lung Cancer Screening Trial cases, and performs similarly on an independent clinical validation set of 1,139 cases. We conducted two reader studies. When prior computed tomography imaging was not available, our model outperformed all six radiologists with absolute reductions of 11% in false positives and 5% in false negatives. Where prior computed tomography imaging was available, the model performance was on-par with the same radiologists. This creates an opportunity to optimize the screening process via computer assistance and automation. While the vast majority of patients remain unscreened, we show the potential for deep learning models to increase the accuracy, consistency and adoption of lung cancer screening worldwide.

In 2013, the United States Preventive Services Task Force recommended low-dose computed tomography (LDCT) lung cancer screening in high-risk populations based on reported improved mortality in the National Lung Cancer Screening Trial (NLST)[2–5]. In 2014, the American College of Radiology published the Lung-RADS guidelines for LDCT lung cancer screening, to standardize image interpretation by radiologists and dictate management recommendations[1,6]. Evaluation is based on a variety of image findings, but primarily nodule size, density and growth[6]. At screening sites, Lung-RADS and other models such as PanCan are used to determine malignancy risk ratings that drive recommendations for clinical management[11,12]. Improving the sensitivity and specificity of lung cancer screening is imperative because of the high clinical and financial costs of missed diagnosis, late diagnosis and unnecessary biopsy procedures resulting from false negatives and false positives[5,13–17]. Despite improved consistency, persistent inter-grader variability and incomplete characterization of comprehensive imaging findings remain as limitations[7–10] of Lung-RADS. These limitations suggest opportunities for more sophisticated systems to improve performance and inter-reader consistency[18,19]. Deep learning approaches offer the exciting potential to automate more complex image analysis, detect subtle holistic imaging findings and unify methodologies for image evaluation[20].

A variety of software devices have been approved by the Food and Drug Administration (FDA) with the goal of addressing workflow efficiency and performance through augmented detection of lung nodules on lung computed tomography (CT)[21]. Clinical research has primarily focused on either nodule detection or diagnostic support for lesions manually selected by imaging experts[22–27]. Nodule detection systems were engineered with the goal of improving radiologist sensitivity in identifying nodules while minimizing costs to specificity, thereby falling into the category of computer-aided detection (CADe)[28]. This approach highlights small nodules, leaving malignancy risk evaluation and clinical decision making to the clinician. Diagnostic support for pre-identified lesions is included in computer-aided diagnosis (CADx) platforms, which are primarily aimed at improving specificity. CADx has gained greater interest and even first regulatory approvals in other areas of radiology, though not in lung cancer at the time of manuscript preparation[29].

To move beyond the limitations of prior CADe and CADx approaches, we aimed to build an end-to-end approach performing both localization and lung cancer risk categorization tasks using the input CT data alone. More specifically, we were interested in replicating a more complete part of a radiologist's workflow, including full assessment of LDCT volume, focus on regions of concern, comparison to prior imaging when available and calibration against biopsy-confirmed outcomes.

Another important high-level decision in our approach was to learn features using deep convolutional neural networks (CNN), rather than using hand-engineered features such as texture features or specific Hounsfield unit values. We chose to learn features because this approach has repeatedly been shown superior to hand-engineered features in many open computer vision competitions in the past five years[30,31], including the Kaggle 2017 Data Science Bowl which used NLST data[32].

There were three key components in our new approach (Fig. 1). First, we constructed a three-dimensional (3D) CNN model that performs end-to-end analysis of whole-CT volumes, using LDCT

[1]Google AI, Mountain View, CA, USA. [2]Stanford Health Care and Palo Alto Veterans Affairs, Palo Alto, CA, USA. [3]Northwestern Medicine, Chicago, IL, USA. [4]New York University-Langone Medical Center, Center for Biological Imaging, New York City, NY, USA. [5]These authors contributed equally: Diego Ardila, Atilla P. Kiraly, Sujeeth Bharadwaj, Bokyung Choi. *e-mail: tsed@google.com

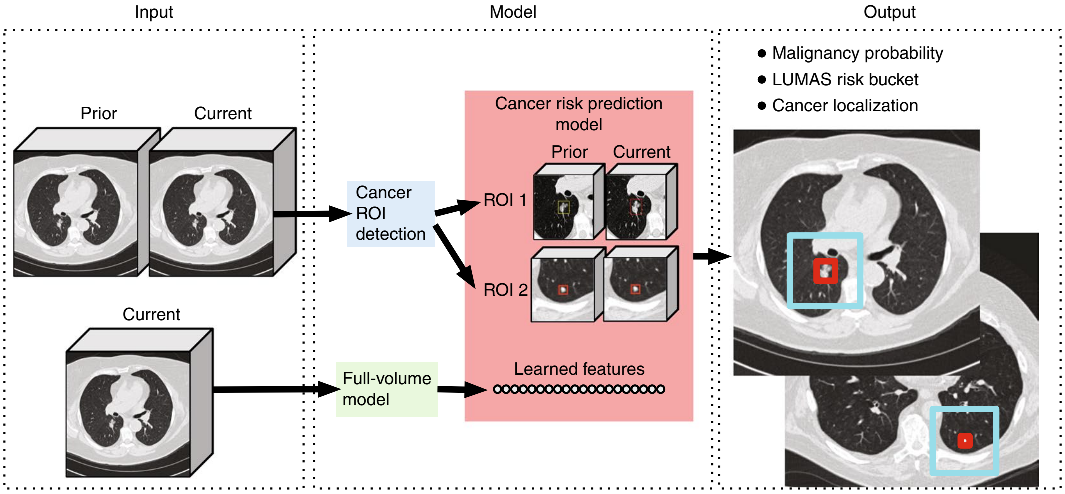


Fig. 1 | Overall modeling framework. For each patient, the model uses a primary LDCT volume and, if available, a prior LDCT volume as input. The model then analyzes suspicious and volumetric ROIs as well as the whole-LDCT volume and outputs an overall malignancy prediction for the case, a risk bucket score (LUMAS) and localization for predicted cancerous nodules.

volumes with pathology-confirmed cancer as training data (the 'full-volume model').

Second, we trained a CNN region-of-interest (ROI) detection model to detect 3D cancer candidate regions in the CT volume (the 'cancer ROI detection model'). We collected additional bounding box labels to train this model.

Third, we developed a CNN cancer risk prediction model that operates on outputs from both the cancer ROI detection model and full-volume model. This can also incorporate regions from a patient's previous scans, which is accomplished by assessing regions in prior scans corresponding to the cancer candidate regions in the current scan, and then assigning a case-level malignancy score (we use the term 'case' to refer to a single patient visit, which could contain multiple CT volumes). This component was also trained on case-level, pathology-confirmed cancer labels (see Methods, Model development and training).

To complete this study, multiple datasets were acquired and various clinical evaluations were performed, as follows.

A deep learning model for analysis of malignancy risk in lung cancer screening CTs was developed from a NLST dataset consisting of 42,290 CT cases from 14,851 patients, 578 of whom developed biopsy-confirmed cancer within the 1-year follow-up period. This represents the entire publicly available dataset provided by the National Institutes of Health (there were 26,722 patients in NLST). Details of how this dataset was selected from the entire NLST screening arm and the inclusion/exclusion criteria are given in Extended Data Fig. 1. Patients were randomly assigned into one of three sets: a training set (70%), a tuning set (15%) and a test set (15%). All CT scan volumes from each patient were then placed into the corresponding set based on this patient assignment. An individual volume was considered cancer-positive if the result of a biopsy or surgical resection was positive during the screening study year, and considered cancer-negative if the patient was cancer-free in the 1-year follow-up screen. Supplementary Tables 1, 2 and 3 contain information on demographics and cancer staging, CT model manufacturer and nodule characteristics for all NLST subsets.

On the test dataset, for 6,716 cases (86 cancer-positives) the model achieved an area under the receiver operating characteristic of 94.4% (95% confidence interval, 91.1–97.3) (see Methods, Statistical analysis). For comparison with radiologists, we then thresholded the model's predictions at three different cutoffs to produce four different lung malignancy scores (LUMAS). These thresholds were chosen so that LUMAS scores corresponded with the probability of malignancy in Lung-RADS buckets 1/2, 3+, 4A+ and 4B/X on the tuning set[33] (see Methods, Operating point selection, for more detail on model score thresholding for LUMAS). Buckets 1 and 2 were combined, as they have the same management recommendation: referral to continued annual screening.

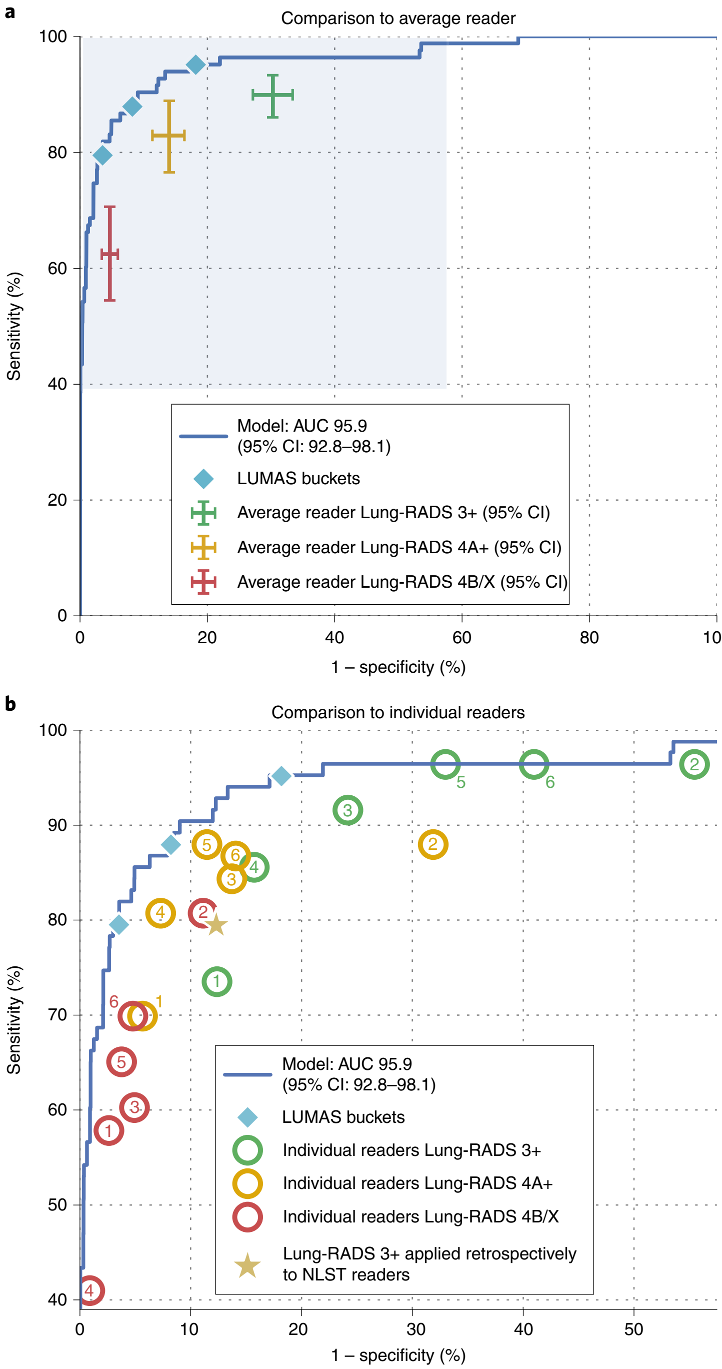
We conducted a two-part retrospective reader study with six US board-certified radiologists (average of 8 years clinical experience, range 4–20 years). In the first part, the radiologists graded a single-screening CT volume. Readers were given access to associated patient demographics and clinical history, while the deep learning model did not have access to this information. Additionally, while the volumes were resampled for the model, the readers assessed the full-resolution original CT cases. Neither the radiologists nor the model had access to previous screening CT volumes from the patient (see Methods, Reader studies). Radiologists reviewed a subset of the test dataset consisting of 507 patients (83 cancer-positives). On this subset of the test set, the model's area under the curve (AUC) was 95.9 (95% confidence interval, 92.8–98.1). This AUC, and the sensitivity/specificity for LUMAS and radiologists, are presented in Fig. 2a,b. The performance of all six radiologists trended at or below the model's receiver operating curve (Fig. 2b).

We compared the model to the average reader performance by measuring the sensitivity and specificity for each LUMAS score and its corresponding Lung-RADS risk bucket (see Methods, Reader studies). The model achieved significantly better sensitivity ($P < 0.05$ for all three thresholds) and better specificity ($P < 0.05$ for two of three thresholds) than the average radiologist (Fig. 2c,d). For instance, comparison of the operating point of LUMAS 3+ to Lung-RADS 3+ yielded a statistically significant specificity boost of 11.6% (95% confidence interval, 7.8–15.1) and a sensitivity boost of 5.2% (95% confidence interval, 0.38–9.9).

We present an alternative methodology for comparison in Supplementary Table 4a,b where, rather than using LUMAS, we set the model sensitivity to match the average reader, compared specificity and then matched specificity to compare sensitivity.

Extended Data Fig. 2 shows the same analysis presented in this section, except that the results have been reweighted to take into account the sampling from the total of 26,722 patients in the NLST screening arm.

In the second part, CT volumes from both the current and previous year were available to the model and the same six radiologists. Comparison with previous scans to assess interval growth is an important component of Lung-RADS[34]. Readers



c

Risk buckets		Sensitivity	Delta
1,2 versus 3+	Average reader	90.0 (86.1, 93.4)	+5.2*
	Model	95.2 (89.9, 98.9)	(.4, 9.8) $P = 0.0386$
1,2,3 versus 4A+	Average reader	82.9 (76.6, 89.4)	+7.4*
	Model	90.4 (83.3, 96.3)	(1.7, 12.9) $P = 0.0114$
1,2,3,4A versus 4B/X	Average reader	62.5 (54.4, 70.7)	+17.1*
	Model	79.5 (70.8, 88.2)	(9.0, 24.5) $P < 1 \times 10^{-4}$

d

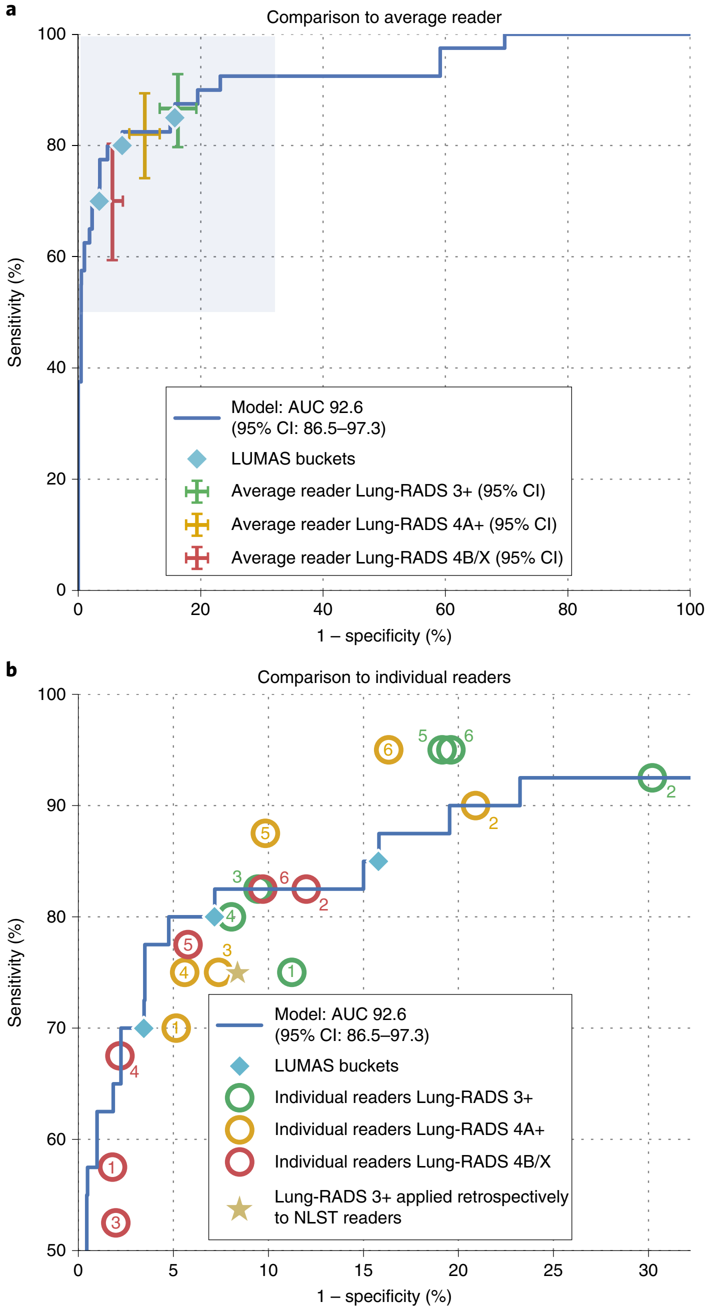
Risk buckets		Specificity	Delta
1,2 versus 3+	Average reader	69.7 (66.6, 72.8)	+11.6*
	Model	81.3 (77.3, 84.9)	(7.8, 15.1) $P < 1 \times 10^{-4}$
1,2,3 versus 4A+	Average reader	86.0 (83.4, 88.4)	+5.0*
	Model	91.0 (88.1, 93.9)	(1.7, 12.9) $P = 0.0008$
1,2,3,4A versus 4B/X	Average reader	95.3 (94.0, 96.6)	+1.1*
	Model	96.5 (94.6, 98.2)	(–0.4, 2.6) $P = 0.143$

e

Risk buckets		
1,2 versus 3+	Hit@1	73/74
	Hit@2	74/74
1,2,3 versus 4A+	Hit@1	72/73
	Hit@2	73/73
1,2,3,4A versus 4B/X	Hit@1	62/63
	Hit@2	63/63

Fig. 2 | Results from the reader study—lung cancer screening on a single CT volume. a–**e**, Performance of radiologists and model in predicting malignancy using single screening CT volumes. Model performance shown in the AUC and summary tables is based on case-level malignancy score. LUMAS buckets refers to operating points selected to match the predicted probability of cancer for Lung-RADS 3+, 4A+ and 4B/X. **a**, Performance of model (blue line) versus average radiologist for various Lung-RADS categories (crosses) using a single CT volume. The length of the crosses represents the confidence Intervals (CIs). The area highlighted in blue is magnified in **b** to show the performance of each of the six radiologists at various Lung-RADS risk buckets. **c**, Sensitivity comparison between model and average radiologist. **d**, Specificity comparison between model and average radiologist. Both sensitivity and specificity analyses were conducted with *n* = 507 volumes from 507 patients, with *P* values computed using a two-sided permutation test with 10,000 random resamplings of the data. **e**, Hit rate localization analysis used to measure how often the model correctly localized a cancerous lesion.

graded 308 volumes from the first reader study that were not from the initial baseline NLST prevalence screening; all of the cases in this subset had prior scans available (see Methods, Reader study—lung cancer screening using current and prior CT volume). On this subset, the model's AUC was 92.6% (95% confidence interval, 86.5–97.3). Notably, both the reader and model performance dropped relative to the first part of the reader study as a result of dropping the CTs from the baseline year. We performed the same comparison as in the previous reader study (Fig. 3). LUMAS showed statistically significant improved specificity for the 4A+ bucket, and otherwise matched the average reader sensitivity and specificity (Fig. 3c,d).



c

Risk buckets		Sensitivity	Delta
1,2 versus 3+	Average reader	86.7 (79.7, 92.9)	+0.8
	Model	87.5 (76.5, 97.2)	(.4, 9.8) $P = 0.0386$
1,2,3 versus 4A+	Average reader	82.1 (74.1, 89.4)	+0.4
	Model	82.5 (69.0, 93.9)	(−10.3, 10.3) $P = 0.975$
1,2,3,4A versus 4B/X	Average reader	70.0 (59.4, 80.3)	+2.5
	Model	72.5 (58.8, 85.7)	(−6.8, 12.3) $P = 0.665$

d

Risk buckets		Specificity	Delta
1,2 versus 3+	Average reader	83.7 (81.1, 87.0)	+0.5
	Model	84.2 (77.7, 88.1)	(−3.7, 4.6) $P = 0.8172$
1,2,3 versus 4A+	Average reader	89.1 (87.1, 91.9)	+3.4*
	Model	92.7 (87.8, 97.4)	(0.7, 6.6) $P = 0.0156$
1,2,3,4A versus 4B/X	Average reader	95.3 (92.0, 96.2)	+1.9
	Model	96.5 (93.2, 98.4)	(−0.1, 4.1) $P = 0.0527$

e

Risk buckets		
1,2 versus 3+	Hit@1	31/31
	Hit@2	31/31
1,2,3 versus 4A+	Hit@1	31/31
	Hit@2	31/31
1,2,3,4A versus 4B/X	Hit@1	27/27
	Hit@2	27/27

Fig. 3 | Results from the reader study—lung cancer screening using current and prior CT volume. a–e, Model performance in the AUC curve and summary tables is based on case-level malignancy score. The term 'LUMAS buckets' refers to operating points selected to represent sensitivity/specificity at the 3+, 4A+ and 4B/X thresholds. **a**, Performance of model (blue line) versus average radiologist at various Lung-RADS categories (crosses) using a CT volume and a prior CT volume per patient. The length of the crosses represents the 95% confidence interval. The area highlighted in blue is magnified in **b** to show the performance of each of the six radiologists at various Lung-RADS categories in this reader study. **c**, Sensitivity comparison between model and average radiologist. **d**, Specificity comparison between model and average radiologist. Both sensitivity and specificity analyses were conducted with $n = 308$ volumes from 308 patients, with P values computed using a two-sided permutation test with 10,000 random resamplings of the data. **e**, Hit rate localization analysis to measure how often the model correctly localized a cancerous lesion.

We present an alternative methodology for comparison in Supplementary Table 4c,d where, rather than using LUMAS, we set the model sensitivity to match the average reader, compared specificity and then matched specificity to compare sensitivity.

Extended Data Fig. 3 shows the same analysis presented in this section, except the results have been reweighted to take into account the sampling from the total 26,722 patients in the NLST screening arm.

Application of the model to all 6,716 cases (86 cancer-positives) in the held-out NLST test set yielded an overall AUC of 94.4% (95% confidence interval, 91.1–97.3). A total of 2,302 cases from the baseline year did not have prior volumes available, but in all other cases readers and the model had access to both current and prior year volumes. We followed an earlier algorithmic methodology[33] to estimate Lung-RADS performance from NLST nodule annotations. Because the nodule annotations in NLST do not contain all of the findings needed by the Lung-RADS guidelines (see Methods, Retrospective application of model to NLST), for comparison of the model to this Lung-RADS estimate we chose a different operating point. For the 1-year cancer outcomes data, we found a boost in specificity (5.0%; 95% confidence interval, 4.2–5.7). We also analyzed the model's performance for a longer-term endpoint, cancer within 2 years, resulting in an AUC of 87.3% (95% confidence interval, 83.2–90.9). For this endpoint, the model yielded improvements in both sensitivity (9.5%; 95% confidence interval, 2.5–16.4) and specificity (5.1%; 95% confidence interval, 4.4–5.9) relative to retrospective-Lung-RADS.

Extended Data Fig. 4 shows the same analysis presented in this section, except that the results have been reweighted to take into account the sampling from the total 26,722 patients in the NLST screening arm.

Under insitutional review board (IRB) approval, we evaluated the model on an additional independent, fully de-identified screening dataset from a US academic medical center, resulting in an AUC of 95.5% (95% confidence interval, 88.0–98.4) (Fig. 4b and Extended Data Fig. 5a). This dataset contained 1,139 cases (27 cancer-positives) and was used to evaluate model performance for biopsy and/or surgically confirmed lung cancers. The model was not trained or tuned using this dataset. Images were not submitted for re-interpretation by radiologists (see Methods, Development and validation datasets, for more details on the dataset). We also evaluated the sensitivity and specificity of LUMAS (Fig. 4b). For LUMAS 3+, we found a sensitivity of 81.5% (95% confidence interval, 66.7–95.0) and a specificity of 89.3% (95% confidence interval, 87.5–91.2).

We performed a localization analysis to measure how often a correct cancer diagnosis was linked with a correct localization. A bounding box was produced by the model for the top two candidate lesions by malignancy risk. For the localization ground truth, each of 79 scans was labeled by two radiologists from a pool of nine. Every scan was derived from a cancer-positive patient in NLST. The radiologists were given the location and staging information from the pathology report, as well as all CT volumes from the patient's data. They were then instructed to label all malignancies with a bounding box. The highest-ranked bounding box overlapped with a malignancy in the scan labeled by our radiologists in all but one case (Figs. 2e and 3e), for a Hit@1 rate of 98%. The Hit@2 rate was 100% (see Methods, Localization analysis, for more details on the Hit metric). These findings were consistent regardless of the specific LUMAS score used to define a true-positive. For a more detailed analysis of the extent of overlap, see Extended Data Fig. 5b.

Given the perceived 'black box' nature of deep learning, an important step in evaluating clinical performance is a deeper assessment of the modeling results. We measured performance on many data subsets to show that the model's overall performance improvements were not obscuring poor performance in clinically relevant subsets. Additionally, we attempted to understand where the model's performance improvements were greatest. The full list of subsets and metrics can be seen in the Supplementary Information (see Supplementary Tables 5 and 6). The model was not statistically inferior relative to the average reader for any metric, subset or risk bucket in either part of the reader study.

Some of the subsets we analyzed were based on a patient's cancer stage when diagnosed according to NLST pathology data. In the first part of the reader study (see Reader study—lung cancer screening on a single CT volume), LUMAS 4B/X versus the average reader Lung-RADS 4B/X showed an absolute improvement in sensitivity of 24.4% (95% confidence interval, 10.4–37.2) for early-stage cancers.

Another group of subsets were based on NLST nodule size annotations. On the subset with nodules 8–15 mm, we saw an absolute improvement in sensitivity of 42.4% (95% confidence interval, 24.7–58.0) in LUMAS 4B/X versus the average reader Lung-RADS 4B/X. An example case of this type is illustrated in Extended Data Fig. 6d, annotated by one radiologist as containing a 12-mm nodule.

Further exploration of model results was completed by two additional radiologists (with 10 and 21 years of clinical experience) reviewing the 140 cases of disagreement between the radiologist consensus and the model from the first part of the reader study (without priors volumes) to evaluate possible causes of disparities (see Supplementary Information, Subjective analysis). The radiologists observed scarring in 22% of the model–reader disagreements and, in 57% of these cases, LUMAS appropriately assigned a lower risk bucket than the readers. This downgrading of scarring accounts for some of the specificity improvements in the model. An example where LUMAS downgraded risk for a cancer-negative case with scarring is shown in Extended Data Fig. 6c (See Supplementary Information, Subjective analysis, for more details of the analysis).

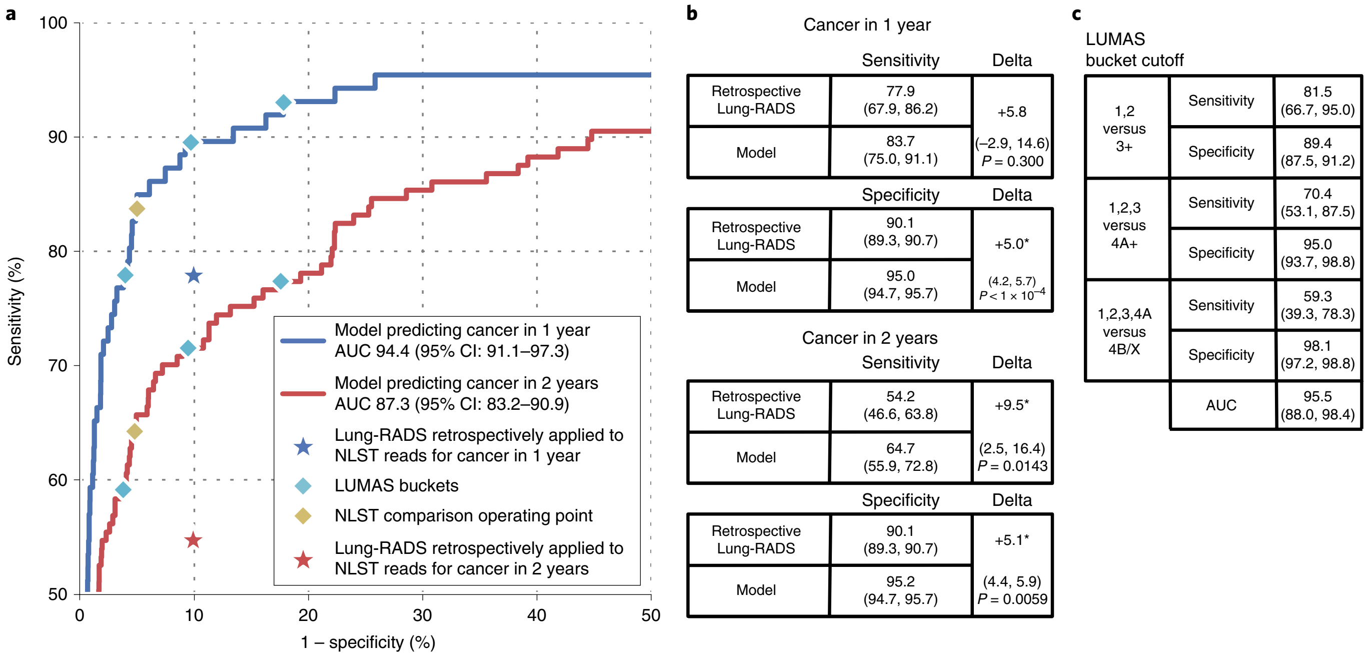
Further analysis of the model's results included examining attribution regions computed with integrated gradients[35], using three radiologists with an average of 23 years' clinical experience (range 10–38 years). Positive and negative classification regions were examined by three radiologists on a subset of examples from the test set. The attribution regions indicated that the model primarily concentrated within and on the edges of the nodule, although in some cases also on the vasculature in the parenchyma (see Supplementary Information, Subjective analysis; Extended Data Fig. 7 and example model false positives in Extended Data Fig. 8).

In summary, we used advanced deep learning techniques to train models with state-of-the-art-performance by leveraging full 3D LDCT volumes, pathology-confirmed case results and prior volumes. These models, if clinically validated, could aid clinicians in evaluating lung cancer screening exams.

Our end-to-end priors approach generates case-level malignancy risk predictions as well as localization information for LDCT lung screening volumes. The strong performance of the model at the case level has important potential clinical relevance. The observed increase in specificity could translate to fewer unnecessary follow-up procedures. Increased sensitivity in cases without priors could translate to fewer missed cancers in clinical practice, especially as more patients begin screening. For patients with prior imaging exams, the performance of the deep learning model could enable gains in workflow efficiency and consistency as assessment of prior imaging is already a key component of a specialist's workflow[36]. Given that LDCT screening is in the relatively early phases of adoption, the potential for considerable improvement in patient care in the coming years is substantial. The model's localization directs follow-up for specific lesion(s) of greatest concern. These predictions are critical for patients proceeding for further work-up and treatment, including diagnostic CT, positron emission tomography (PET)/CT or biopsy.

Malignancy risk prediction allows for the possibility of augmenting existing, manually created interpretation guidelines such as Lung-RADS, which are limited to subjective clustering and assessment to approximate cancer risk. Numerous investigations have evaluated CADx applications built to assist radiologists in classification of suspected lesions previously detected and segmented by radiologists[18]. These prior CADx studies typically report only a lesion-level classification performance, which is not comparable to this work. In contrast, the model presented performs human-independent detection and classification on full volumes. Past non-peer-reviewed efforts that have attempted direct, automated malignancy prediction from full volumes using deep learning methods reported AUCs as high as 0.88 (ref. [37]). However, these models were primarily trained and tested on smaller portions of the NLST dataset, did not evaluate the use of priors and did not report localization metrics[32,37]. We hypothesize that taking into account a larger context in our cancer risk prediction model (larger ROIs around candidate regions, whole-3D volume assessment and priors) and training on a larger portion of NLST led to superior performance.

While we did note a performance decrease in the with-priors subset, we also found a corresponding drop in performance for our readers. This decrease may be because patients with easy-to-spot



Cancer in 1 year

	Sensitivity	Delta
Retrospective Lung-RADS	77.9 (67.9, 86.2)	+5.8 (−2.9, 14.6) $P = 0.300$
Model	83.7 (75.0, 91.1)	

	Specificity	Delta
Retrospective Lung-RADS	90.1 (89.3, 90.7)	+5.0* (4.2, 5.7) $P < 1 \times 10^{-4}$
Model	95.0 (94.7, 95.7)	

Cancer in 2 years

	Sensitivity	Delta
Retrospective Lung-RADS	54.2 (46.6, 63.8)	+9.5* (2.5, 16.4) $P = 0.0143$
Model	64.7 (55.9, 72.8)	

	Specificity	Delta
Retrospective Lung-RADS	90.1 (89.3, 90.7)	+5.1* (4.4, 5.9) $P = 0.0059$
Model	95.2 (94.7, 95.7)	

LUMAS bucket cutoff		
1,2 versus 3+	Sensitivity	81.5 (66.7, 95.0)
	Specificity	89.4 (87.5, 91.2)
1,2,3 versus 4A+	Sensitivity	70.4 (53.1, 87.5)
	Specificity	95.0 (93.7, 98.8)
1,2,3,4A versus 4B/X	Sensitivity	59.3 (39.3, 78.3)
	Specificity	98.1 (97.2, 98.8)
	AUC	95.5 (88.0, 98.4)

Fig. 4 | Results of the full NLST and independent test sets. a, Comparison of model performance to NLST reader performance on the full NLST test set. NLST reader performance was estimated by retrospectively applying Lung-RADS 3 criteria to the NLST reads. **b**, Sensitivity and specificity comparisons between the model and Lung-RADS retrospectively applied to NLST reads. The comparison was performed on $n = 6{,}716$ cases, using a two-sided permutation test using 10,000 random resamplings of the data. **c**, Sensitivity and specificity of different LUMAS buckets on an independent dataset comprising $n = 1{,}139$ cases using the same two-sided set with 10,000 random resamplings. The full AUC plot is shown in Extended Data Fig. 5a.

cancers are diagnosed and dropped from the study in the baseline year, leaving only more subtle cancer cases.

We propose a LUMAS system in this paper, but the underlying techniques allow for broader exploration of other risk stratification methods. Incorporating these new methods into CAD systems could also address the issues of inter-grader variability in lung cancer assessment, a pattern seen in both our reader study (see Supplementary Table 7) and prior publications[38,39].

Explainability of deep learning models is still at an early stage. To begin to explore how the model evaluates risk of malignancy, we asked our clinicians to analyze a subset of cases subjectively. We hypothesize that there are advantages to the model's more consistent visualization of morphological features, such as scars and nodules, in 3D. Additionally, the model was not bound by the size guidelines in Lung-RADS, allowing for new risk categorization. We found cases where the model appeared to use features outside of the main nodule, such as the vasculature and parenchyma surrounding the nodule (see Supplementary Information, Subjective analysis). However, we do not know whether the model incorporates other abnormalities such as background emphysema in its predictions. Further examination using model attribution techniques may allow radiologists to take advantage of the same visual features used by the model to assess malignancy.

Our study did have some important limitations. While our radiologist-comparison studies were larger than in prior published work[32], they were still limited to retrospective data from the NLST dataset. Although clinical comparison metrics were limited to a small number of general (not thoracic) radiologists, lung cancer screening is commonly performed by general radiologists[40].

Another limitation resulting from initial lung cancer screening studies is the relative lack of cancer outcomes information available. In spite of this, our multi-stage modeling approach was able to leverage the natural distribution of data from the screening population using only 398 cancer-positives for training. We were also encouraged by the indicators of generalizability of our model to an independent dataset from another patient population. As we used only two datasets during testing, there is a limit to the conclusions that can be drawn about generalizability. However, the NLST test set we used represents 33 different test sites across 21 different manufacturer and model combinations. In addition, our academic medical center test set is derived from 1,039 cases, all in the years post-NLST. Further study will require testing and tuning against an even broader variability of screening data parameters to ensure generalizability.

Lastly, although we presented a methodology for choosing operating points for the model, this was primarily for the purposes of comparing reader and model performance. It is important to stress that the selection of operating points for use in clinical practice remains an ongoing area of research, potentially involving an analysis of costs and outcomes to properly trade off between sensitivity and specificity.

More robust retrospective and prospective studies will be required to ensure clinical applicability as screening programs continue to scale. In future studies we aim to explore different approaches in presenting radiologists with model output assessments, including malignancy risk calculations and localization. Correlating the performance improvements with documented improved clinical outcomes and health system costs will also be required to determine potential impact. Another opportunity would be to apply similar modeling techniques to routine diagnostic CT, aiding in the detection and management of incidental pulmonary nodules.

In addition to its application to lung cancer screening, the deep learning techniques applied in this study have considerable relevance to other types of 3D imaging data. For instance, this approach holds promise for magnetic resonance imaging, PET or other types of volumetric or multi-view problem research. Our research also has applications in workflows involving comparison with a patient's prior imaging.

Lastly, the early stage of lung cancer screening adoption led to a relative scarcity of quality ground truth data for training. While this presented a challenge during the research process, it demonstrated that it is possible for deep learning to achieve radiologist-level performance with a smaller number of positive examples. As data scarcity is a common problem in medical deep learning research, we hope these methods will translate to new opportunities for exploration, especially in rare diseases.

In conclusion, these results represent a step toward automated image evaluation via lung cancer risk malignancy estimation through deep learning. We believe this research could supplement future approaches to lung cancer screening as well as support assisted- or second-read workflows. In addition, we believe the general approach employed in our work, mainly outcomes-based training, full volume techniques and directly comparable clinical performance evaluation, may lay additional groundwork toward deep learning medical applications.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41591-019-0447-x.

Received: 1 October 2018; Accepted: 5 April 2019;
Published online: 20 May 2019

References

1. American Lung Association. Lung cancer fact sheet. *American Lung Association* http://www.lung.org/lung-health-and-diseases/lung-disease-lookup/lung-cancer/resource-library/lung-cancer-fact-sheet.html (accessed 11 September 2018).
2. Jemal, A. & Fedewa, S. A. Lung cancer screening with low-dose computed tomography in the United States—2010 to 2015. *JAMA Oncol.* **3**, 1278 (2017).
3. US Preventive Services Task Force. Final update summary: lung cancer: screening (1AD). *US Preventive Services Task Force* https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/lung-cancer-screening (2018).
4. National Lung Screening Trial Research Team et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.* **365**, 395–409 (2011).
5. Black, W. C. et al. Cost-effectiveness of CT screening in the National Lung Screening Trial. *N. Engl. J. Med.* **371**, 1793–1802 (2014).
6. Lung CT screening reporting & data system. *American College of Radiology* https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads (accessed 11 September 2018).
7. van Riel, S. J. et al. Observer variability for Lung-RADS categorisation of lung cancer screening CTs: impact on patient management. *Eur. Radiol.* **29**, 924–931 (2019).
8. Singh, S. et al. Evaluation of reader variability in the interpretation of follow-up CT scans at lung cancer screening. *Radiology* **259**, 263 (2011).
9. Mehta, H. J., Mohammed, T.-L. & Jantz, M. A. The American College of Radiology lung imaging reporting and data system: potential drawbacks and need for revision. *Chest* **151**, 539–543 (2017).
10. Martin, M. D., Kanne, J. P., Broderick, L. S., Kazerooni, E. A. & Meyer, C. A. Lung-RADS: pushing the limits. *Radiographics* **37**, 1975–1993 (2017).
11. Winkler Wille, M. M. et al. Predictive accuracy of the pancan lung cancer risk prediction model—external validation based on CT from the Danish Lung Cancer Screening Trial. *Eur. Radiol.* **25**, 3093–3099 (2015).
12. De Koning, H., Van Der Aalst, K., Ten Haaf, M. & Oudkerk, H. D. K. C. PL02.05 Effects of volume CT lung cancer screening: mortality results of the NELSON randomised-controlled population based tria. *J. Thorac. Oncol.* **13**, S185 (2018).
13. Field, J. K. et al. UK Lung Cancer RCT Pilot Screening Trial: baseline findings from the screening arm provide evidence for the potential implementation of lung cancer screening. *Thorax* **71**, 161–170 (2016).
14. McMahon, P. M. et al. Cost-effectiveness of computed tomography screening for lung cancer in the United States. *J. Thorac. Oncol.* **6**, 1841–1848 (2011).
15. Goffin, J. R. et al. Cost-effectiveness of lung cancer screening in canada. *JAMA Oncol.* **1**, 807 (2015).
16. Tomiyama, N. et al. CT-guided needle biopsy of lung lesions: a survey of severe complication based on 9783 biopsies in Japan. *Eur. J. Radiol.* **59**, 60–64 (2006).
17. Wiener, R. S., Schwartz, L. M., Woloshin, S. & Welch, H. G. Population-based risk for complications after transthoracic needle lung biopsy of a pulmonary nodule: an analysis of discharge records. *Ann. Intern. Med.* **155**, 137–144 (2011).
18. Ciompi, F. et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Sci. Rep.* **7**, 46479 (2017).
19. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **278**, 563–577 (2016).
20. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
21. Bogoni, L. et al. Impact of a computer-aided detection (CAD) system integrated into a picture archiving and communication system (PACS) on reader sensitivity and efficiency for the detection of lung nodules in thoracic CT exams. *J. Digit. Imaging* **25**, 771–781 (2012).
22. Ye, Xujiong et al. Shape-based computer-aided detection of lung nodules in thoracic CT images. *IEEE Trans. Biomed. Eng.* **56**, 1810–1820 (2009).
23. Bellotti, R. et al. A CAD system for nodule detection in low-dose lung CTs based on region growing and a new active contour model. *Med. Phys.* **34**, 4901–4910 (2007).
24. Sahiner, B. et al. Effect of CAD on radiologists' detection of lung nodules on thoracic CT scans: analysis of an observer performance study by nodule size. *Acad. Radiol.* **16**, 1518–1530 (2009).
25. Firmino, M., Angelo, G., Morais, H., Dantas, M. R. & Valentim, R. Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy. *Biomed. Eng. Online* **15**, 2 (2016).
26. Armato, S. G. et al. Lung cancer: performance of automated lung nodule detection applied to cancers missed in a CT screening program. *Radiology* **225**, 685–692 (2002).
27. Valente, I. R. S. et al. Automatic 3D pulmonary nodule detection in CT images: a survey. *Comput. Methods Prog. Biomed.* **124**, 91–107 (2016).
28. Das, M. et al. Performance evaluation of a computer-aided detection algorithm for solid pulmonary nodules in low-dose and standard-dose MDCT chest examinations and its influence on radiologists. *Br. J. Radiol.* **81**, 841–847 (2008).
29. Quantitative Insights. Quantitative Insights gains industry's first FDA clearance for machine learning driven cancer diagnosis. *PRNewswire* https://www.prnewswire.com/news-releases/quantitative-insights-gains-industrys-first-fda-clearance-for-machine-learning-driven-cancer-diagnosis-300495405.html (2018).
30. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
31. Lin, T.-Y. et al. Microsoft COCO: Common Objects in Context. *Comput. Vis. ECCV* **2014**, 740–755 (2014).
32. Liao, F., Liang, M., Li, Z., Hu, X. & Song, S. Evaluate the malignancy of pulmonary nodules using the 3D deep leaky noisy-or network. Preprint at https://arxiv.org/abs/1711.08324 (2017).
33. Pinsky, P. F. et al. Performance of Lung-RADS in the national lung screening trial: a retrospective assessment. *Ann. Intern. Med.* **162**, 485 (2015).
34. Manos, D. et al. The Lung Reporting and Data System (LU-RADS): a proposal for computed tomography screening. *Can. Assoc. Radiol. J.* **65**, 121–134 (2014).
35. Sun, Y. & Sundararajan, M. Axiomatic attribution for multilinear functions. In *Proc. 12th ACM Conference on Electronic Commerce—EC '11* https://doi.org/10.1145/1993574.1993601 (2011).
36. Varela, C., Karssemeijer, N., Hendriks, J. H. C. L. & Holland, R. Use of prior mammograms in the classification of benign and malignant masses. *Eur. J. Radiol.* **56**, 248–255 (2005).
37. Trajanovski, S. et al. Towards radiologist-level cancer risk assessment in CT lung screening using deep learning. Preprint at https://arxiv.org/abs/1804.01901 (2019).
38. Pinsky, P. F., Gierada, D. S., Nath, P. H., Kazerooni, E. & Amorosa, J. National lung screening trial: variability in nodule detection rates in chest CT studies. *Radiology* **268**, 865–873 (2013).
39. Armato, S. G. 3rd et al. The Lung Image Database Consortium (LIDC): an evaluation of radiologist variability in the identification of lung nodules on CT scans. *Acad. Radiol.* **14**, 1409–1421 (2007).
40. Kazerooni, E.A. et al. ACR–STR practice parameter for the performance and reporting of lung cancer screening thoracic computed tomography (CT). *J. Thorac. Imaging* **29**, 310–316 (2014).

Acknowledgements

The authors acknowledge the NCI and the Foundation for the National Institutes of Health for their critical roles in the creation of the free publicly available LIDC/IDRI/NLST Database used in this study. All participants enrolled in NLST signed an informed consent developed and approved by the screening center's IRBs, the NCI IRB and the Westat IRB. The authors thank the NCI for access to NCI data collected by the NLST. The statements herein are solely those of the authors and do not represent or imply concurrence or endorsement by the NCI. The authors would like to thank M. Etemadi and his team at Northwestern Medicine for data collection, de-identification and research support. These team members include E. Johnson, F. Garcia-Vicente, D. Melnick, J. Heller and S. Singh. We also thank C. Christensen and his team at Northwestern Medicine IT, including M. Lombardi, C. Wilbar and R. Atanasiu. We would also like to acknowledge the work of the team working on labeling infrastructure and, specifically, J. Yoshimi, who implemented many of the features we needed for labeling of ROIs, and J. Wong for coordinating and recruiting radiologist labelers. We would also like to acknowledge the work of the team that put together the data-handling infrastructure, including G. Duggan and K. Eswaran. Lastly, we would like to acknowledge the helpful feedback on the initial drafts from Y. Liu and S. McKinney.

Author contributions

D.A., A.P.K., S.B. and B.C. developed the network architecture and data/modeling infrastructure, training and testing setup. D.A. and A.P.K. created the figures, wrote the methods and performed additional analysis requested in the review process. J.J.R., S.S., D.T., D.A., A.P.K. and S.B. wrote the manuscript. D.P.N. and J.J.R. provided clinical expertise and guidance on the study design. G.C and S.S. advised on the modeling techniques. M.E., S.S., J.J.R., B.C., W.Y. and D.A. created the datasets, interpreted the data and defined the clinical labels. D.A., B.C., A.P.K. and S.S. performed statistical analysis. S.S, L.P. and D.T. initiated the project and provided guidance on the concept and design. S.S. and D.T supervised the project.

Competing interests

D.P.N. and J.J.R. are paid consultants of Google Inc. D.P.N. is on the Medical Advisory Board of VIDA Diagnostics, Inc. and Exact Sciences. M.E.'s lab received funding from Google Inc. to support the research collaboration. This study was funded by Google Inc. The remaining authors are employees of Google Inc. and own stock as part of the standard compensation package. The authors have no other competing interests to disclose.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-019-0447-x.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-019-0447-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to D.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Development and validation datasets. We used data from the NLST study, consisting of 42,290 CT cases from 14,851 patients, 638 of whom developed biopsy-confirmed cancer within 1 year of a LDCT screening (see Extended Data Fig. 1 for more details on NLST dataset selection)[41]. Patients were randomly assigned to a training set (70%), a tuning set (15%) or a test set (15%). Because not all negative cases from NLST have been made publicly available, the training, tuning and test sets had cancer percentages of 3.9, 4.5 and 3.7, respectively (slightly higher than the 1–2% range reported for NLST in general and in real-world practice). Supplementary Tables 1, 2 and 3 describe demographics, scanner information and nodule and cancer characteristics for relevant subsets of this dataset. All participants enrolling in NLST signed an informed consent developed and approved by the screening centers' 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. Briefly, LDCTs were collected from multiple institutions, with slice spacing varying from 1.25 to 5 mm and scanner vendors varying by site. We filtered out the 5-mm scans to better represent the slice spacing of a typical modern screening protocol[42], and the largest remaining slice spacing was 2.5 mm. A diagnosis of lung cancer established by biopsy at any time during the same year as a screening case counted as a ground truth true-positive case. This included cases identified as incidental cancers diagnosed during the same screening year as an initially negative screening exam. An exam was considered negative if the patient proved cancer-free on 1-year follow-up; patients in the trial had multi-year follow-up. Patients had up to 3 years of screening, all via LDCT and, in nearly all cases, only one visit occurred per year with exceptions made for patients with inadequate imaging or interval development of symptoms concerning for cancer. In cases where prior imaging was used for testing and development purposes, the screening exam from the preceding year was selected. As screening read data from NLST were gathered once per year for each patient, it was important also to evaluate the model once per year for each patient for our tuning and testing sets. We chose the latest case per screening year, since this was the most likely case to have generated the screening read because patients typically were asked to return only if imaging was inadequate. Within each case we used the best available reconstruction kernel (See Supplementary Information, Kernel selection) with the highest number of slices.

An independent dataset from an academic medical center was used to further validate the model's performance. This dataset consisted of 1,139 cases from 907 patients collected as part of a screening program (see Extended Data Fig. 9 for exclusion criteria and further details); 209 of the patients and 232 of the cases had priors available. These data were not used in the training or tuning of the model. The data were a fully de-identified lung cancer screening CT dataset. The ground truth for cancer on this dataset was defined based on lung cancer International Classification of Disease codes with biopsy or surgical confirmation of cancer via manual review of the pathology note. For cancer-negatives, patients had a cancer-free follow-up examination at least 1 year after the initial screening exam. Slice spacing for CTs in this dataset varied from 1.25 to 3.0 mm, with the majority (84%) being 3.0 mm. Notably, our training set in NLST had a maximum spacing of 2.5 mm, suggesting that our model generalized to different scanning parameters.

Model development and training. Overall, the model is trained to take the entire CT volume and automatically produce a score predicting whether the patient received a cancer diagnosis in the same study year. First, for clarity it is important to define the following terms.

Volume always refers to the full CT volume (that is, the entire set of axial images comprising the volume)—whether in original resolution or resampled. When we describe that the 'volume' is labeled as malignant or non-malignant, we intend to communicate that the label is at a case level (that is, 'there is cancer in the CT scan somewhere').

Bounding box is a rectangular 3D sub-volume containing a malignancy. Our radiologist labelers were instructed to draw boxes that tightly encapsulate the malignancy. We call these resulting sub-volumes bounding boxes for this reason. Our detection model aims to predict these bounding boxes.

ROI is a fixed-size, 3D sub-volume containing a malignancy and some surrounding context. Once we have bounding boxes from our detection model, we take a fixed 90-mm^3 region around each bounding box. We call this larger 3D sub-volume an ROI.

Since the use of only a single label for an entire volume can be a challenging learning task, part of the model used a two-stage approach leveraging bounding box labels. First, two candidate ROIs were detected using a detection component trained on radiologist-annotated bounding boxes (see Methods, Localization analysis for annotation details and Extended Data Fig. 6 for details on how candidate regions were cropped from detected bounding boxes). We tried using up to seven candidates and arrived at two based on tune set performance.

Next, we combined the scores $p1$ and $p2$ from both candidates using the 'noisy-or' equation$1 - (1 - p1)(1 - p2)$ to produce a final score which was then trained against the case-level cancer diagnosis labels (see Extended Data Fig. 10). To summarize, the use of 'noisy-or' lets us train against the case-level ground truth in NLST even though internally the model is making predictions about two ROIs.

When classifying each candidate region, a purely two-stage approach would have access only to features within the candidate region and not from the full volume. It was not technically feasible to train a model on the full volume at the original resolution. To provide this global context for every candidate region, we trained a model on the full volume at a reduced resolution to predict cancer diagnosis and then combined features extracted from this model to those extracted from each candidate region. The input volume for each case was the entire 3D CT volume for the case, including the lung, mediastinum, heart, chest wall and so on, just as a radiologist would be given in practice. No manual image segmentation was performed. A total of 29,541 cases were used for training, including all volumes with slice thickness less than or equal to 2.5 mm. The model consists of lung segmentation, cancer ROI detection, a full-volume model and a final cancer risk prediction model based on the outputs of the full-volume model and the cancer ROI detection model. For each of these components, we chose a more general computer vision task (that is, instance segmentation, object detection and video classification) that was similar to the task performed by the component. Then, for each task we chose an approach that was state of the art at the time of our modeling experiments. For a schematic overview of the model see Fig. 1, and for a more detailed overview see Extended Data Fig. 10.

The approach consists of four components, all trained using the TensorFlow platform (Google Inc.)[43]:

(1) Lung segmentation. We trained a lung segmentation Mask-RCNN[44] approach, trained on the LUNA[45] dataset using the TensorFlow Object Detection API[46], which produced the lung segmentation mask. This mask was used to compute the center of its bounding box for step (c) and to determine an alignment with the prior volume. Since only the bounding box center is the key result of interest, the precise segmentation boundaries are not a factor in our modeling approach. It is likely that other lung segmentation approaches could substitute this component. Finding the lung center allows us to focus further processing on the lungs.

(2) Cancer ROI detection model. This was trained on 1.4 × 0.7 mm^2 (spacing, pixel size) voxel size volumes. The cancer ROI detection architecture was a RetinaNet[47] modified to be in 3D and to remove the feature pyramid network[48]. Extended Data Fig. 6a demonstrates how a large ROI was cropped around each bounding box detected. The detection model was initialized by first training on LIDC[39] and then trained on radiologist-annotated lesion bounding boxes collected on the NLST dataset. The cancer ROI detection component outputs ROIs from all input volumes, even if no nodules are present. In this case, the most nodule-like regions are proposed as ROIs.

(3) Full-volume model. An end-to-end convolutional model, 3D inflated Inception V1 (ref. [49,50]), was trained on the 1.5-mm^3 voxel size volumes to predict cancer within 1 year, fine-tuning from a checkpoint trained on ImageNet[51]. Each of these volumes was a large region cropped around the center of the bounding box as determined by lung segmentation. This cancer prediction model was trained with focal loss[47] to try to mitigate the sparsity of positive examples. We trained the model to predict cancer probability and then used the last layer before the final probability, which contains 1,024 units. We take these 1,024 numbers as the output for this model, and use them as features later on.

(4) Cancer risk prediction model. A final cancer classification model was used to consider the output of the previous two models. In all cases, 3D Inception is used to extract features. Throughout the model components, our approach to classifying and extracting features from 3D volumes is heavily based on this 3D Inception model[51]. First, features were extracted from the detected ROIs (Extended Data Fig. 6a). Features from the full-volume model were appended to the final layers of each detected ROI in the second-stage model, so that all predictions relied on both nodule-level local information and global context from the entire CT volume. Extended Data Fig. 10 illustrates the unified end-to-end approach, after the top two candidate ROIs were passed to the second-stage malignancy classification model. It was trained as a single convolutional neural network with shared parameters across all detected ROIs. Each ROI was passed through this network to predict its individual malignancy score. The final prediction was generated by combining the two probability scores as shown in Extended Data Fig. 10 (ref. [32]). This model was also trained with focal loss[47] to try to mitigate the sparsity of positive examples.

The final cancer prediction model was developed to allow as input either a single CT scan (without prior) or both the current and prior year scan (with prior). The prior and current volumes were aligned based on the lung bounding box centers of two volumes and then by aligning nearby center candidate ROIs from the prior scan when available (Extended Data Fig. 6b). In each case a 3D shift of prior volume is performed to align the two centers. Higher-level spatial feature maps from the current and prior scans were combined and passed through additional convolutional layers with batch normalization. Since features of the current and prior scan considered at these higher levels represent the entire 90-mm^3 (the 64 × 128-mm^2 cropped sub-volume with voxel size 1.4 × 0.7 mm^2) sub-volume at a low spatial resolution, precise alignment of the nodules is not required.

In the case of a malignant prediction, nodule localization was performed by selecting the ROI with the highest malignancy score. For a benign prediction, the detection model is still forced to produce two ROIs which are then later rejected by the cancer risk prediction model. The final model is an ensemble of ten models trained with different random initializations. Additional detail can be found in Supplementary Information, Additional modeling details.

Clinical validation. The NLST-based test set comprised 6,716 cases, 86 of which had a biopsy-confirmed cancer within 1 year of screening. The model's output is a probability between 0 and 1, which was bucketed using three thresholds. We used a previously developed approach to estimate the positive predictive value (PPV) of Lung-RADS 3, Lung-RADS 4A and Lung-RADS 4B/X[33]. We then chose three operating points that matched these PPV values on our tuning set, to have comparable probability of malignancy with the four existing Lung-RADS risk buckets. Since Lung-RADS 1 and 2 have the same management recommendations (return to routine annual screening) and risk of malignancy, we grouped these in the same bucket for this experiment. These operating points define LUMAS by establishing cutoffs for 1/2 versus 3 and 4A/B/X, 1/2/3 versus 4A/B/X and 1/2/3/4A versus 4B/X: the same cutoffs within Lung-RADS at which the likelihood of malignancy increases and management is changed. When readers gave S- (other non-lung cancer findings) or C- (prior lung cancer diagnosis) modified ratings, these were treated in the same way as those without modifications (for example, 3C was treated the same as 3), and cases with ratings of 0 were considered not gradable and dropped from the analysis. Both test sets were run only once to avoid influencing model development. Additionally, all individuals who worked on modeling and image analysis were blinded to the diagnoses in the test set.

Operating point selection. We define three LUMAS operating points as a way to compare the model to the readers. We computed Lung-RADS 3+ performance on our tune set using the nodule annotations from the original NLST readers, to arrive at a PPV of 0.11. We then adjusted the threshold of our model on the tune set to match this PPV of 0.11 and used the resulting model score threshold as our LUMAS 3+ threshold. We estimated 4A+ and 4B/X PPVs using a previous analysis of NLST[33], which gave a PPV of 0.15 for 4A+ and 0.25 for 4B/X from which we computed LUMAS thresholds for 4A and 4B/X, respectively. We present an alternative way of making model-to-reader comparisons in Supplementary Table 4.

Retrospective application of model to NLST. The model used current and prior CT volumes when available. We followed the methodology in prior work[33] to estimate the performance of Lung-RADS 3 across the entire held-out test set, using nodule growth annotations to take into account priors when possible. For brevity, we call this performance estimate retrospective-Lung-RADS. As can be seen from Figs. 2b and 3b, retrospective-Lung-RADS seems to overestimate the specificity and underestimate the sensitivity compared to the readers for Lung-RADS 3+. Reasons for these differences may include the fact that the NLST dataset nodule annotations are insufficient for accurate computation of Lung-RADS retrospectively. For example, endobronchial nodules were not noted in the NLST data and were therefore ignored, and the exact amount of nodule growth was not noted. As retrospective-Lung-RADS operates in such a different part of the receiver operating curve compared to Lung-RADS, we chose a different, non-LUMAS, operating point to compare our model's performance to the readers in NLST (see Fig. 4a). We found two operating points, one which matched the sensitivity and one which matched the specificity of retrospective-Lung-RADS on our tune set. We then chose a final operating point midway between these two to improve both sensitivity and specificity in a balanced manner.

Reader studies. A two-part reader study was conducted comparing the model to six radiologists on a subset of the test set. All radiologists were US board-certified with an average of 8 years' clinical experience (range 4–20 years). Each reader independently reviewed the same set of cases and applied the Lung-RADS 2014 v.1 criteria to determine a Lung-RADS score. A fully featured, web-based DICOM viewer (eUnity, Client Outlook Inc.) with FDA 510(k) clearance was used to evaluate cases. While the first reader study did not use prior imaging, the second used a single prior CT scan for comparison. In each case, readers were given information about the patient: race, gender, ethnicity, smoking history and cancer history. The model does not make use of this clinical information, as initial experiments with this data did not improve performance.

Performance comparisons were made for malignancy risk evaluation between the model and the average results of the six radiologists. For the model (using LUMAS) and the average reader (using Lung-RADS), we computed sensitivity and specificity at each of the three risk bucket thresholds—3+, 4A+ and 4B/X. The average reader sensitivity and specificity were computed by taking the average of the six individual reader sensitivities and specificities, respectively. The without priors reader study subset consisted of 507 patients, 83 of which were cancer-positives. There was a single volume per patient and the subset was enriched for biopsied cases (see Extended Data Fig. 1 for details on exclusion and enrichment). The cancer-negative biopsy cases were down-weighted in the subsequent analysis such that the final metrics on negatives were representative of a random sampling of negatives from the NLST test set. The with priors reader study was conducted on all cases from the without priors reader study that had an available prior.

Reader study—lung cancer screening on a single CT volume. A total of 507 cases (83 cancer-positives) were each independently interpreted by six US board-certified radiologists. In this study, neither the model nor the readers were given access to prior cases. Only axial CT slices were available for the first 250 cases; for the remaining cases, sagittal and coronal reformations and maximum-intensity projection images were available. Lung-RADS scores, slice number and anatomic lung location were recorded, and readers saved an ROI for each lesion with a Lung-RADS score of 2 or greater.

Reader study—lung cancer screening using current and prior CT volume. After completing part one, 308 patients (40 cancer-positives) that were known to have prior CTs available were re-presented to the same readers, now with a scan from the prior year available and with readers following the guidelines for Lung-RADS with baseline comparisons. They were then allowed to modify their Lung-RADS scores from part one. The model was also given access to the same CT scan from the prior year.

Localization analysis. Each NLST cancer-positive volume was labeled by two radiologists from a pool of nine (including five fellows and four practicing radiologists with a range of experience of 4–21 years (average, 9 years)). The radiologists were given the coarse locations of pathology-confirmed malignancies noted in NLST, and were then asked to label all malignancies with bounding boxes. We used all boxes labeled by either radiologist. Overlapping boxes referencing the same malignancy were combined into a single box by averaging the coordinates comprising the box. The Hit@N metric was defined as the fraction of true-positive cases in which the top N candidate lesions from the detection model made any overlap with an annotated malignancy. The recall metric involving all cancer-positive cases is presented in Supplementary Table 8.

Statistical analysis. All confidence intervals were computed based on the percentiles of 1,000 random resamplings (bootstraps) of the data. Confidence intervals for differences were derived by computing the metric of interest and then computing a reader–model difference on each bootstrap. *P* values for sensitivity and specificity comparisons were computed using a standard permutation test[52] using 10,000 random resamplings of the data. Briefly, for each resampling we randomly swapped the reader and model results for each case[35]. We then performed a two-sided hypothesis test comparing the model–reader difference with the distribution of 10,000 model–reader differences across the resampled data to obtain an empirical *P* value.

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

Data availability

This study used three datasets that are publicly available: 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 Medicine was used under license for the current study, and is not publicly available.

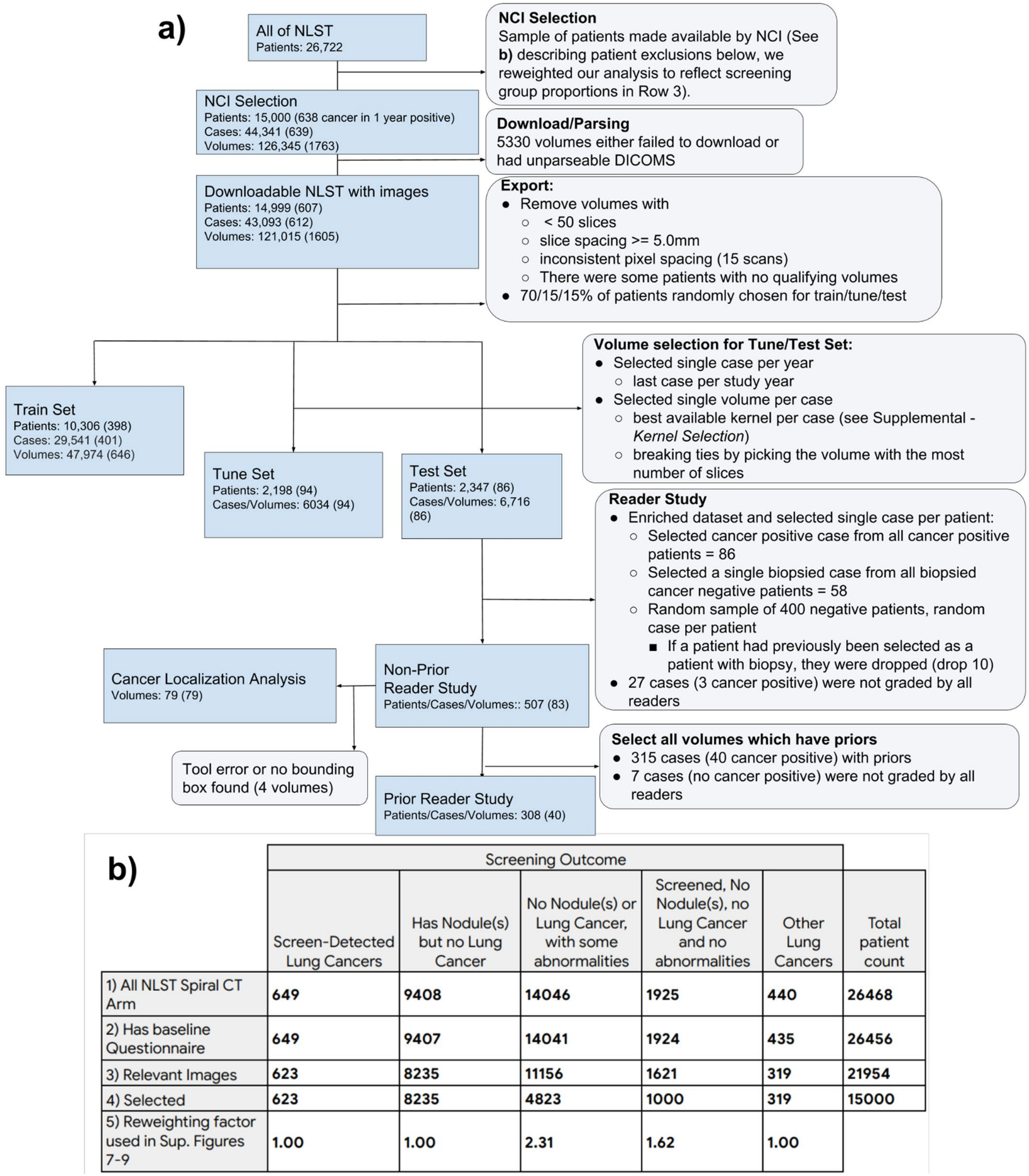
Code availability

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 Methods section to allow independent replication with non-proprietary libraries. Several major components of our work are available in open source repositories: Tensorflow: https://www.tensorflow.org; Tensorflow Estimator API: https://www.tensorflow.org/guide/estimators; Tensorflow Object Detection API: https://github.com/tensorflow/models/tree/master/research/object_detection—the lung segmentation model and cancer ROI detection model were trained using this framework; Inflated Inception: https://github.com/deepmind/kinetics-i3d—the full-volume model and the second-stage model were trained using this feature extractor.

References

41. National Cancer Institute. National Lung Screening Trial https://www.cancer.gov/types/lung/research/nlst (2018)
42. The American College of Radiology. Adult lung cancer screening specifications. https://www.acr.org/Clinical-Resources/Lung-Cancer-Screening-Resources (2014).
43. Abadi, M. et al. Tensorflow: a system for large-scale machine learning. *OSDI* **16**, 265–283 (2016).
44. He, K., Gkioxari, G., Dollar, P. & Girshick, R. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)* https://doi.org/10.1109/iccv.2017.322 (IEEE, 2017).
45. Setio, A. A. A. et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. *Med. Image Anal.* **42**, 1–13 (2017).
46. Huang, J. et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* https://doi.org/10.1109/cvpr.2017.351 (IEEE, 2017).

47. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollar, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* Preprint at https://doi.org/10.1109/TPAMI.2018.2858826 (2018).
48. Lin, T.-Y. et al. Feature pyramid networks for object detection. in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* https://doi.org/10.1109/cvpr.2017.106 (IEEE, 2017).
49. Carreira, J. & Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* https://doi.org/10.1109/cvpr.2017.502 (IEEE, 2017).
50. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* https://doi.org/10.1109/cvpr.2016.308 (IEEE, 2017).
51. J. Deng et al. ImageNet: a large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition* https://doi.org/10.1109/cvprw.2009.5206848 (IEEE, 2009).
52. Chihara, L. M. & Hesterberg, T. C. *Mathematical Statistics with Resampling and R* (John Wiley & Sons, 2014).



All of NLST
Patients: 26,722

NCI Selection
Sample of patients made available by NCI (See **b)** describing patient exclusions below, we reweighted our analysis to reflect screening group proportions in Row 3).

NCI Selection
Patients: 15,000 (638 cancer in 1 year positive)
Cases: 44,341 (639)
Volumes: 126,345 (1763)

Download/Parsing
5330 volumes either failed to download or had unparseable DICOMS

Downloadable NLST with images
Patients: 14,999 (607)
Cases: 43,093 (612)
Volumes: 121,015 (1605)

Export:
- Remove volumes with
  - < 50 slices
  - slice spacing >= 5.0mm
  - inconsistent pixel spacing (15 scans)
  - There were some patients with no qualifying volumes
- 70/15/15% of patients randomly chosen for train/tune/test

Volume selection for Tune/Test Set:
- Selected single case per year
  - last case per study year
- Selected single volume per case
  - best available kernel per case (see Supplemental - *Kernel Selection*)
  - breaking ties by picking the volume with the most number of slices

Train Set
Patients: 10,306 (398)
Cases: 29,541 (401)
Volumes: 47,974 (646)

Tune Set
Patients: 2,198 (94)
Cases/Volumes: 6034 (94)

Test Set
Patients: 2,347 (86)
Cases/Volumes: 6,716 (86)

Reader Study
- Enriched dataset and selected single case per patient:
  - Selected cancer positive case from all cancer positive patients = 86
  - Selected a single biopsied case from all biopsied cancer negative patients = 58
  - Random sample of 400 negative patients, random case per patient
    - If a patient had previously been selected as a patient with biopsy, they were dropped (drop 10)
- 27 cases (3 cancer positive) were not graded by all readers

Cancer Localization Analysis
Volumes: 79 (79)

Non-Prior Reader Study
Patients/Cases/Volumes:: 507 (83)

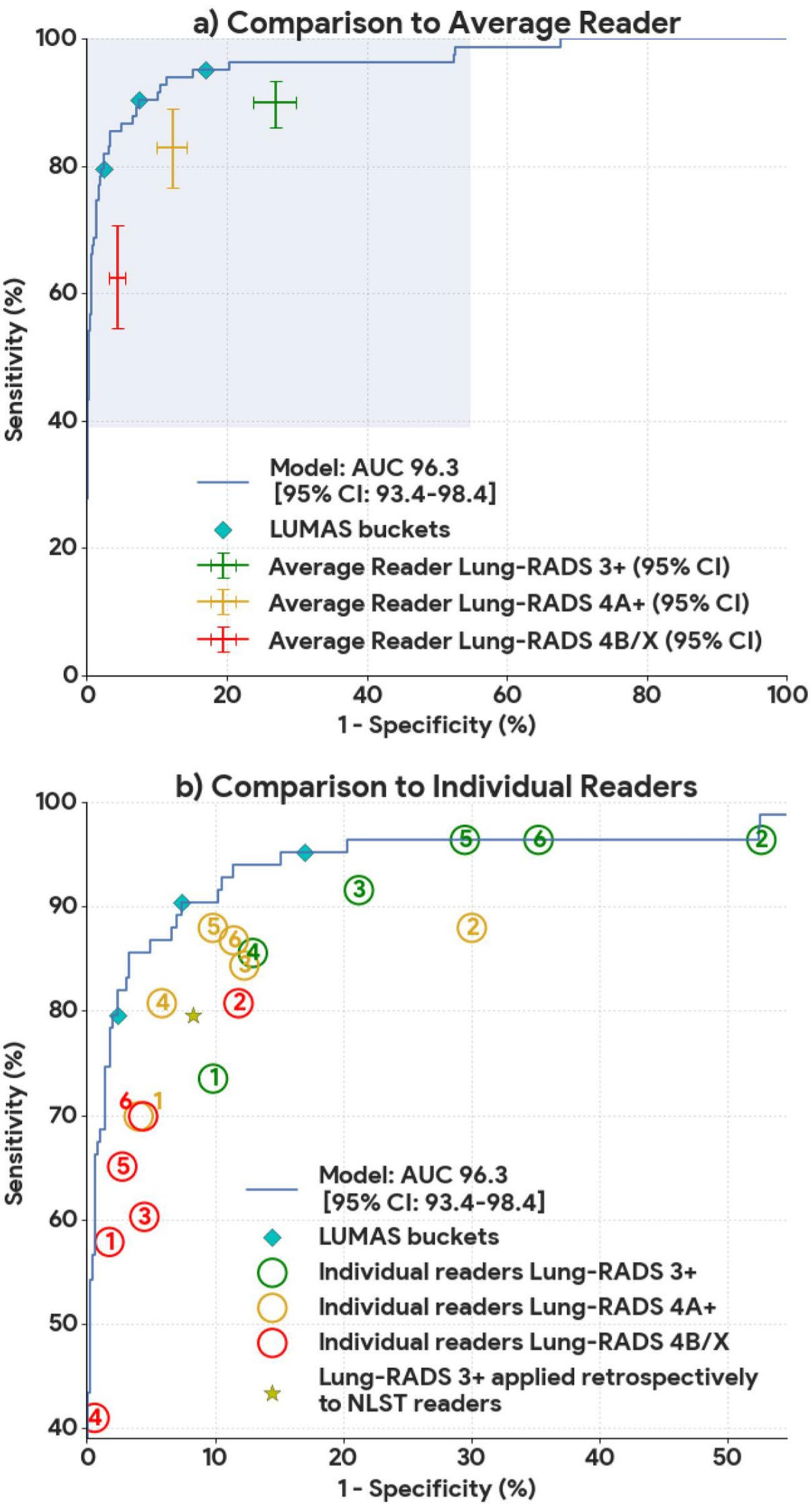
Tool error or no bounding box found (4 volumes)

Select all volumes which have priors
- 315 cases (40 cancer positive) with priors
- 7 cases (no cancer positive) were not graded by all readers

Prior Reader Study
Patients/Cases/Volumes: 308 (40)

	Screening Outcome					
	Screen-Detected Lung Cancers	Has Nodule(s) but no Lung Cancer	No Nodule(s) or Lung Cancer, with some abnormalities	Screened, No Nodule(s), no Lung Cancer and no abnormalities	Other Lung Cancers	Total patient count
1) All NLST Spiral CT Arm	**649**	**9408**	**14046**	**1925**	**440**	**26468**
2) Has baseline Questionnaire	**649**	**9407**	**14041**	**1924**	**435**	**26456**
3) Relevant Images	**623**	**8235**	**11156**	**1621**	**319**	**21954**
4) Selected	**623**	**8235**	**4823**	**1000**	**319**	**15000**
5) Reweighting factor used in Sup. Figures 7-9	**1.00**	**1.00**	**2.31**	**1.62**	**1.00**	

Extended Data Fig. 1 | NLST STARD diagram. a, Diagram describing exclusions made in our analysis. **b**, Table describing exclusions made by the NCI when selecting images to release from NLST. Note that there were 623 screen-detected cancers but a total of 638 cancer-positive patients. The additional 15 patients were diagnosed during the screening window, but not due to a positive screening result. In this case Row 3 'Relevant Images' meant that, for cancer-positive patients, there were images from the year of the cancer diagnosis, and for cancer-negative patients it meant that all 3 years of screening images were available. Note that the publicly available version of NLST downsampled the screening groups 3 (no nodule, some abnormality) and 4 (no nodule, no abnormalities). In Extended Data Figs. 2, 3, 4 and Supplementary Table 4 we present another version of the main analysis that compensates for this downsampling by upweighting patients within these groups.



c) Sensitivity Comparison

Risk Buckets			Delta
1,2 vs. 3+	Average Reader	90.0 [86.1, 93.4]	+5.2* [.4, 9.8] p=0.0386
	Model	95.2 [89.9, 98.9]	
1,2,3 vs. 4A+	Average Reader	82.9 [76.6, 89.0]	+7.4* [1.7, 12.9] p=0.0114
	Model	90.4 [83.3, 96.3]	
1,2,3,4A vs. 4B/X	Average Reader	62.5 [54.4, 70.7]	+17.1* [9.0, 24.5] p<10e-4
	Model	79.5 [70.8, 88.2]	

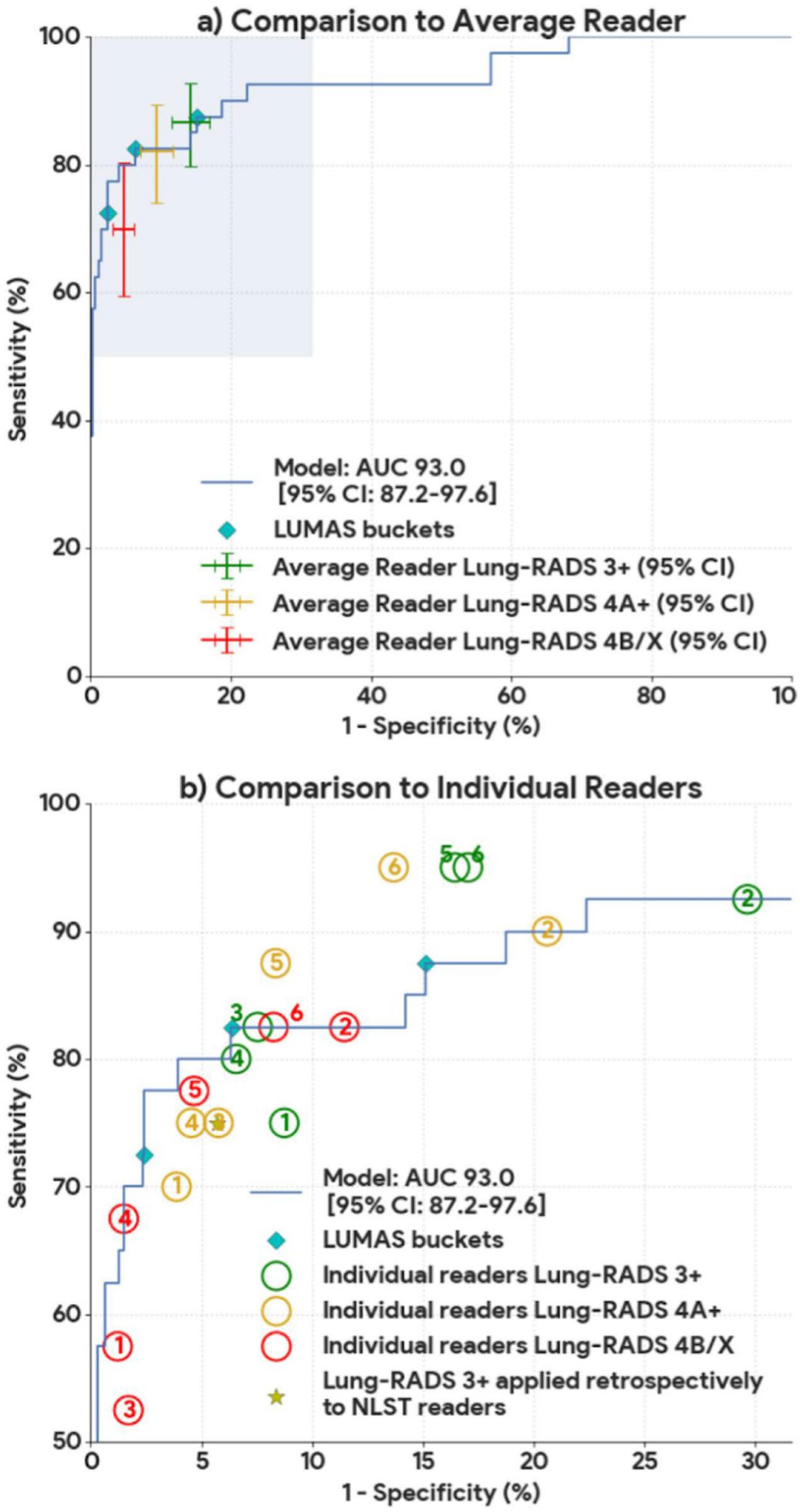
d) Specificity Comparison

Risk Buckets			Delta
1,2 vs. 3+	Average Reader	73.1 [69.9, 76.0]	+10.0* [6.0, 13.7] p<10e-4
	Model	83.0 [85.4, 89.8]	
1,2,3 vs. 4A+	Average Reader	87.7 [85.4, 89.8]	+4.9* [2.1, 7.8] p=.0003
	Model	93.2 [90.6, 95.6]	
1,2,3,4A vs. 4B/X	Average Reader	95.7 [94.6, 96.8]	+1.9* [0.7, 3.1] p=0.0017
	Model	97.6 [96.3, 98.8]	

e) Localization: Hit Rate

Risk Buckets		
1,2 vs. 3+	Hit@1	73/74
	Hit@2	74/74
1,2,3 vs. 4A+	Hit@1	72/73
	Hit@2	73/73
1,2,3,4A vs. 4B/X	Hit@1	62/63
	Hit@2	63/63

Extended Data Fig. 2 | Results from the reader study—lung cancer screening on a single CT volume: reweighted. a–e, Identical to Fig. 2, except that we took into account the biased sampling done in the selection of the NLST data released. This meant that examples in screening groups 3 (no nodule, some abnormality) and 4 (no nodule, no abnormality) were upweighted by the same factor by which they were downsampled (see Extended Data Fig. 1 for further details on the groups). Model performance shown in the AUC curve and summary tables is based on case-level malignancy score. LUMAS buckets refers to operating points selected to match the predicted probability of cancer for Lung-RADS 3+, 4A+ and 4B/X. **a**, Performance of model (blue line) versus average radiologist for various Lung-RADS categories (crosses) using a single CT volume. The lengths of the crosses represent the confidence intervals. The area highlighted in blue is magnified in **b** to show the performance of each of the six radiologists at various Lung-RADS risk buckets. **c**, Sensitivity comparison between model and average radiologist. **d**, Specificity comparison between model and average radiologist. Both sensitivity and specificity analysis were conducted with *n* = 507 volumes from 507 patients, with *P* values computed using a two-sided permutation test with 10,000 random resamplings of the data. **e**, Hit rate localization analysis used to measure how often the model correctly localized a cancerous lesion.



c) Sensitivity Comparison

Risk Buckets			Delta
1,2 vs. 3+	Average Reader	86.7 [79.7, 92.9]	+0.8 [-9.8, 11.8] p = 0.9007
	Model	87.5 [76.5, 97.2]	
1,2,3 vs. 4A+	Average Reader	82.1 [74.1, 89.4]	+0.4 [-10.3, 10.3] p=0.975
	Model	82.5 [69.0, 93.9]	
1,2,3,4A vs. 4B/X	Average Reader	70.0 [59.4, 80.3]	+2.5 [-6.8, 12.3] p=0.674
	Model	72.5 [58.8, 85.7]	

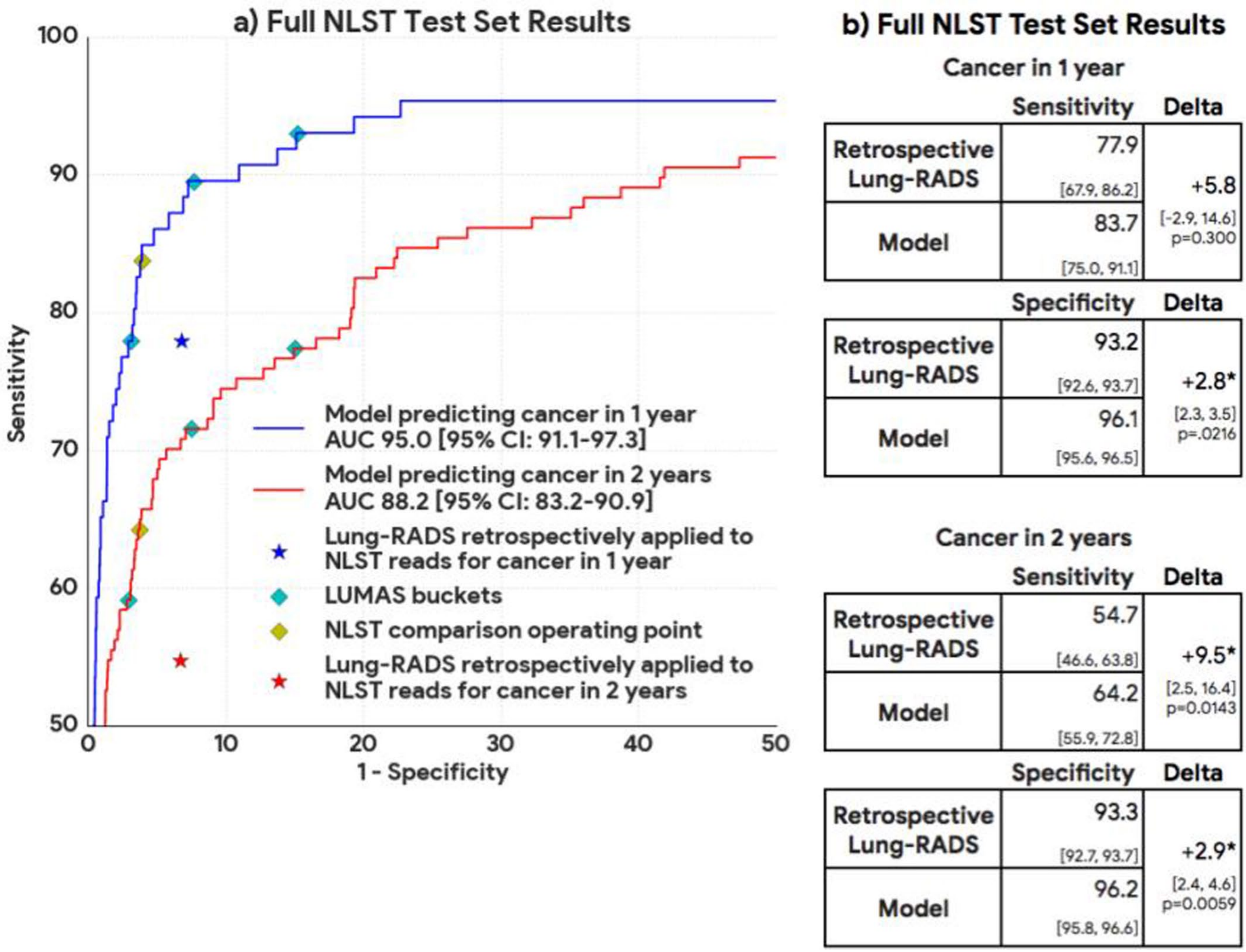
d) Specificity Comparison

Risk Buckets			Delta
1,2 vs. 3+	Average Reader	85.7 [82.9, 88.4]	-0.8 [-9.8, 11.8] p=0.730
	Model	84.9 [79.9, 89.4]	
1,2,3 vs. 4A+	Average Reader	90.5 [88.2, 92.9]	+3.1* [0.1, 5.9] p = 0.0425
	Model	93.6 [90.4 96.5]	
1,2,3,4A vs. 4B/X	Average Reader	95.2 [93.6, 96.7]	+2.4* [.6, 4.1] p = 0.0062
	Model	97.6 [96.0, 99.0]	

e) Localization: Hit Rate

Risk Buckets		
1,2 vs. 3+	Hit@1	31/31
	Hit@2	31/31
1,2,3 vs. 4A+	Hit@1	31/31
	Hit@2	31/31
1,2,3,4A vs. 4B/X	Hit@1	27/27
	Hit@2	27/27

Extended Data Fig. 3 | Results from the reader study—lung cancer screening using current and prior CT volume: reweighted. a–**e**, Identical to Fig. 3, except that we took into account the sampling done in the selection of the 15,000 patient NLST data released. This meant that for screening groups 3 (no nodule, some abnormality) and 4 (no nodule, no abnormality) we upweighted each example by the same factor by which they were downsampled. Model performance in the AUC curve and summary tables is based on case-level malignancy score. The term LUMAS buckets refers to operating points selected to represent sensitivity/specificity at the 3+, 4A+ and 4B/X thresholds. **a**, Performance of model (blue line) versus average radiologist at various Lung-RADS categories (crosses) using a CT volume and a prior CT volume for a patient. The length of the crosses represents the 95% confidence interval. The area highlighted in blue is magnified in **b** to show the performance of each of the six radiologists at various Lung-RADS categories in this reader study. **c**, Sensitivity comparison between model and average radiologist. **d**, Specificity comparison between model and average radiologist. Both sensitivity and specificity analysis were conducted with $n = 308$ volumes from 308 patients with P values computed using a two-sided permutation test with 10,000 random resamplings of the data. **e**, Hit rate localization analysis used to measure how often the model correctly localized a cancerous lesion.



Cancer in 1 year

	Sensitivity	Delta
Retrospective Lung-RADS	77.9 [67.9, 86.2]	+5.8 [-2.9, 14.6] p=0.300
Model	83.7 [75.0, 91.1]	

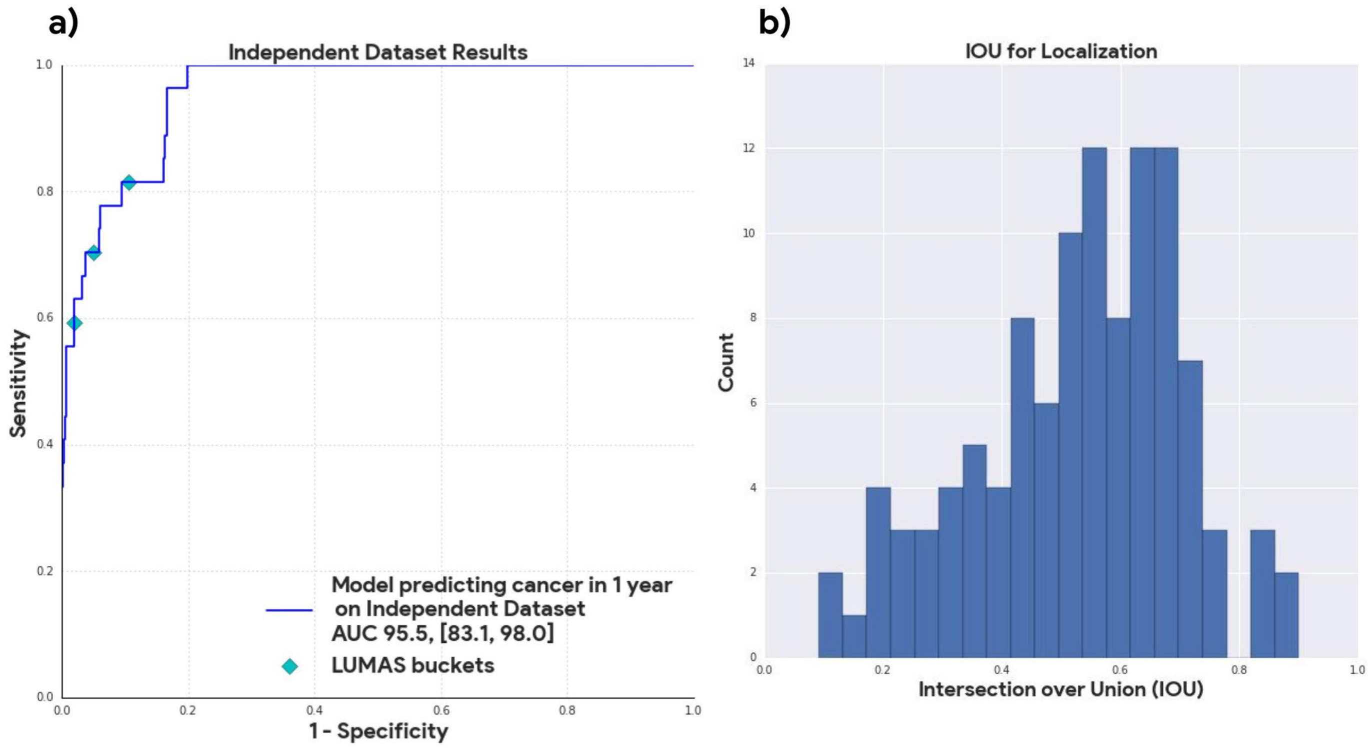
	Specificity	Delta
Retrospective Lung-RADS	93.2 [92.6, 93.7]	+2.8* [2.3, 3.5] p=.0216
Model	96.1 [95.6, 96.5]	

Cancer in 2 years

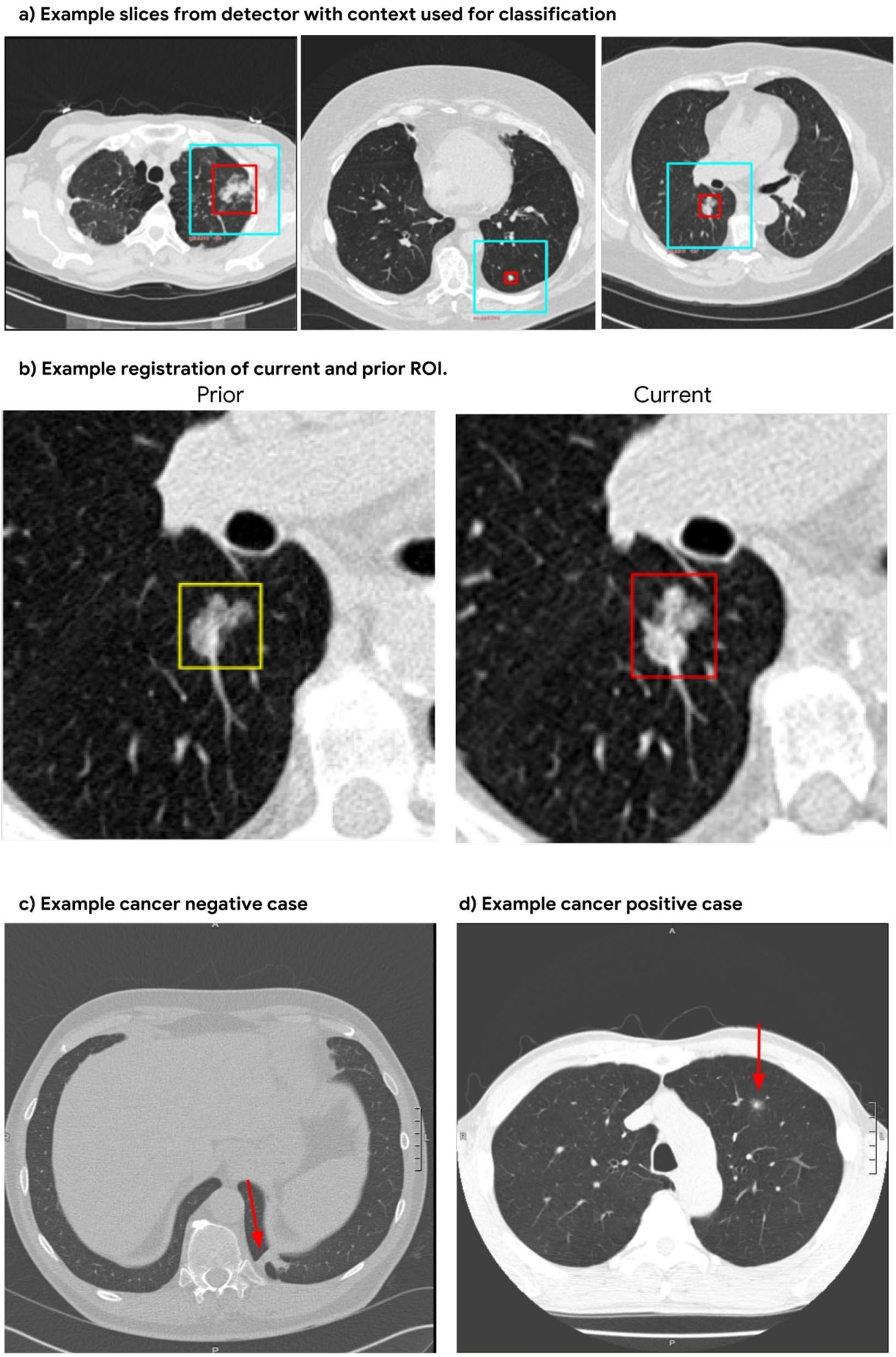
	Sensitivity	Delta
Retrospective Lung-RADS	54.7 [46.6, 63.8]	+9.5* [2.5, 16.4] p=0.0143
Model	64.2 [55.9, 72.8]	

	Specificity	Delta
Retrospective Lung-RADS	93.3 [92.7, 93.7]	+2.9* [2.4, 4.6] p=0.0059
Model	96.2 [95.8, 96.6]	

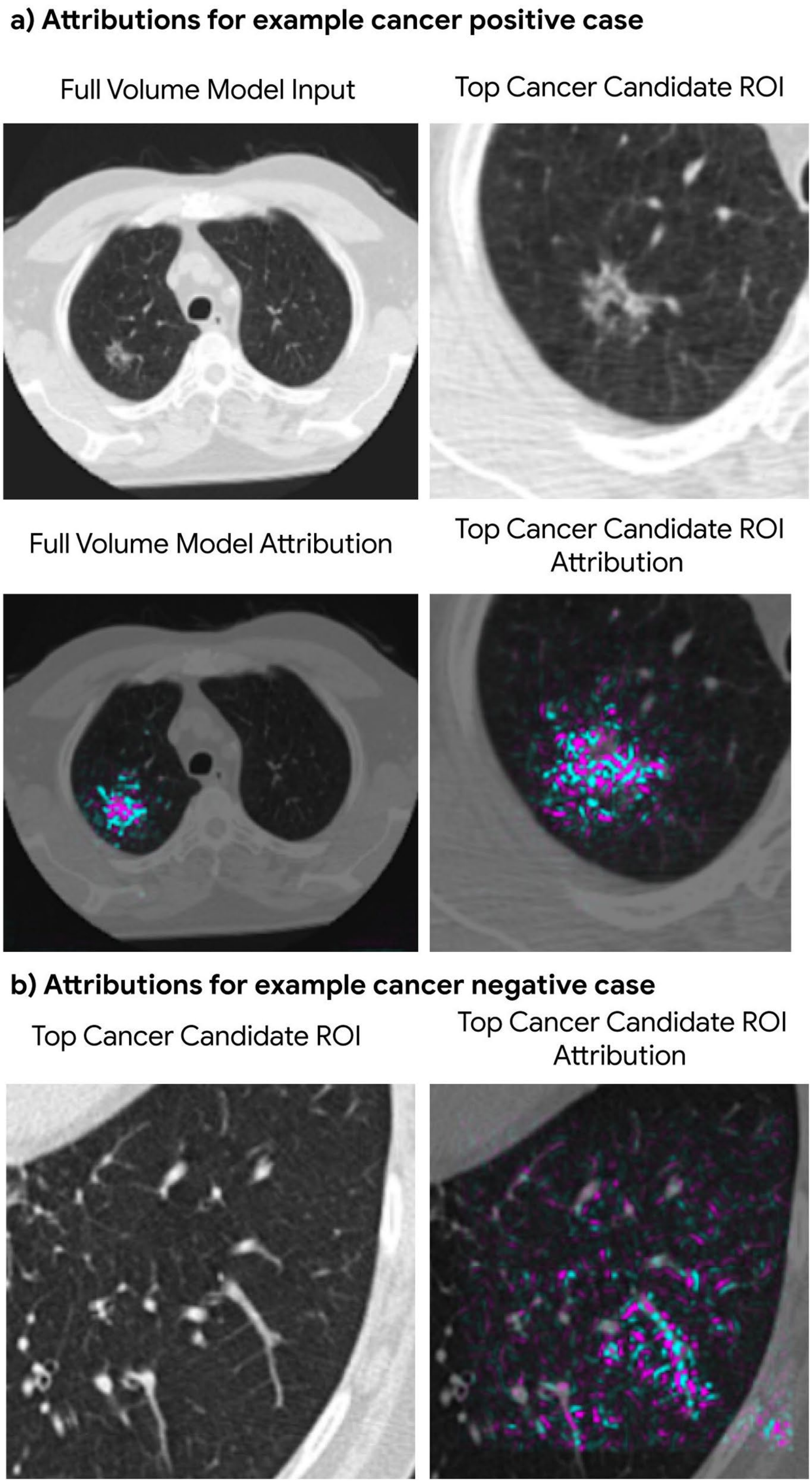
Extended Data Fig. 4 | Results from the full NLST test set and independent test set: reweighted. a,**b**, Identical to Fig. 4 except that we took into account the biased sampling done in the selection of the NLST data released. This meant that for screening groups 3 (no nodule, some abnormality) and 4 (no nodule, no abnormality) we upweighted each example by the same factor by which they were downsampled. The comparison was performed on $n = 6{,}716$ cases, using a two-sided permutation test with 10,000 random resamplings of the data. **a**, Comparison of model performance to NLST reader performance on the full NLST test set. NLST reader performance was estimated by retrospectively applying Lung-RADS 3 criteria to the NLST reads. **b**, Sensitivity and specificity comparisons between the model and Lung-RADS retrospectively applied to NLST reads.



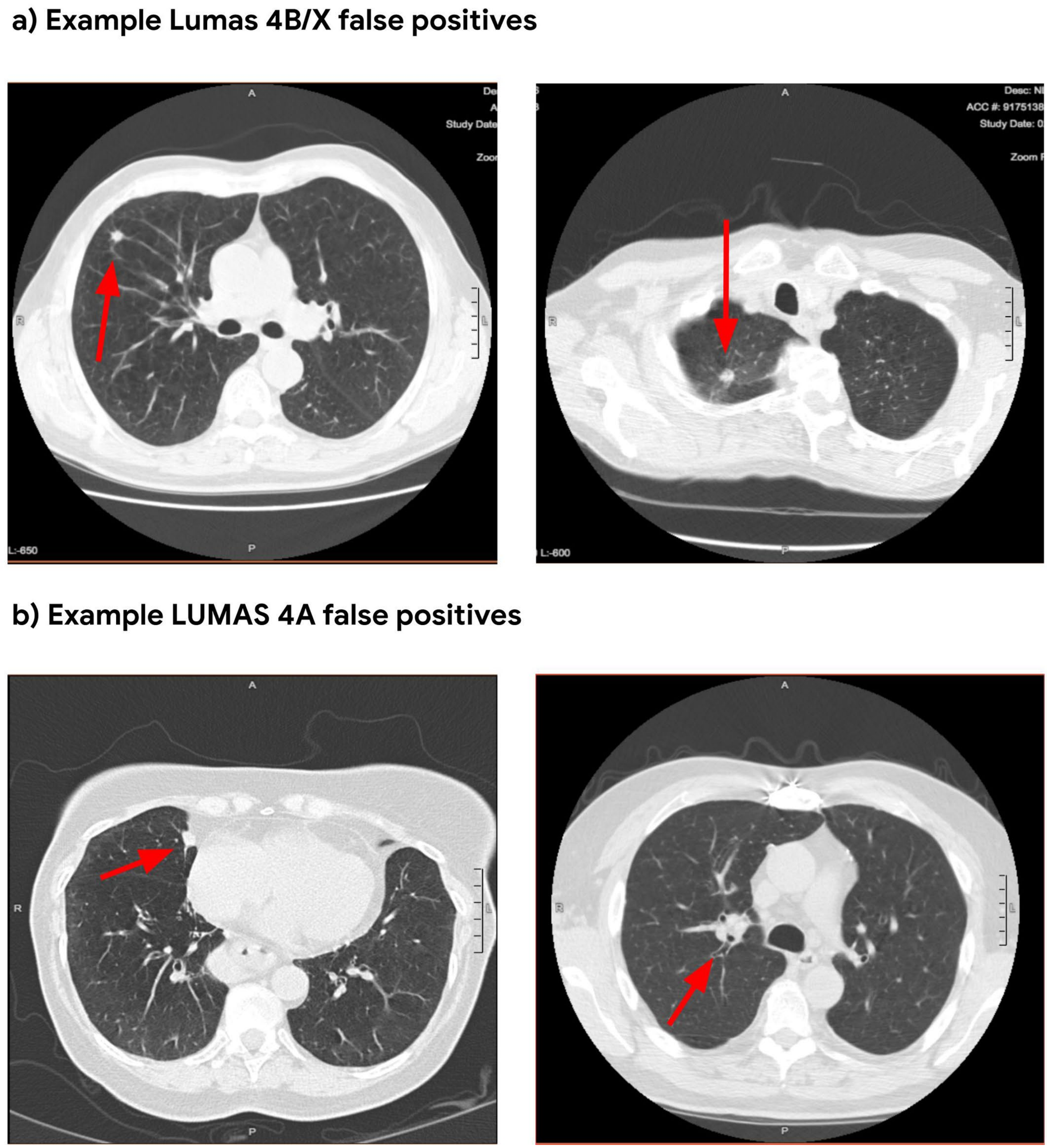
Extended Data Fig. 5 | Independent dataset ROC curve and intersection over union for localization. a, AUC curve for the independent data test set with $n=1{,}139$ cases using a two-sided permutation test with 10,000 random resamplings of the data. **b**, For each detection that was a 'hit' (overlapped with a labeled malignancy), this plot shows the volume of the intersection between the detection and the ground truth divided by the volume of the union of the ground truth and the detection. In 3D, intersection over union (IOU) drops much faster than in two dimensions (2D). For example, given a 1-mm^3 nodule and a correctly centered 2-mm^3 bounding box, the resulting IOU will be 0.125. In 2D, a similar situation would result in an IOU of 0.25.


a) Example slices from detector with context used for classification
b) Example registration of current and prior ROI.
Prior
Current
c) Example cancer negative case
d) Example cancer positive case
A
P

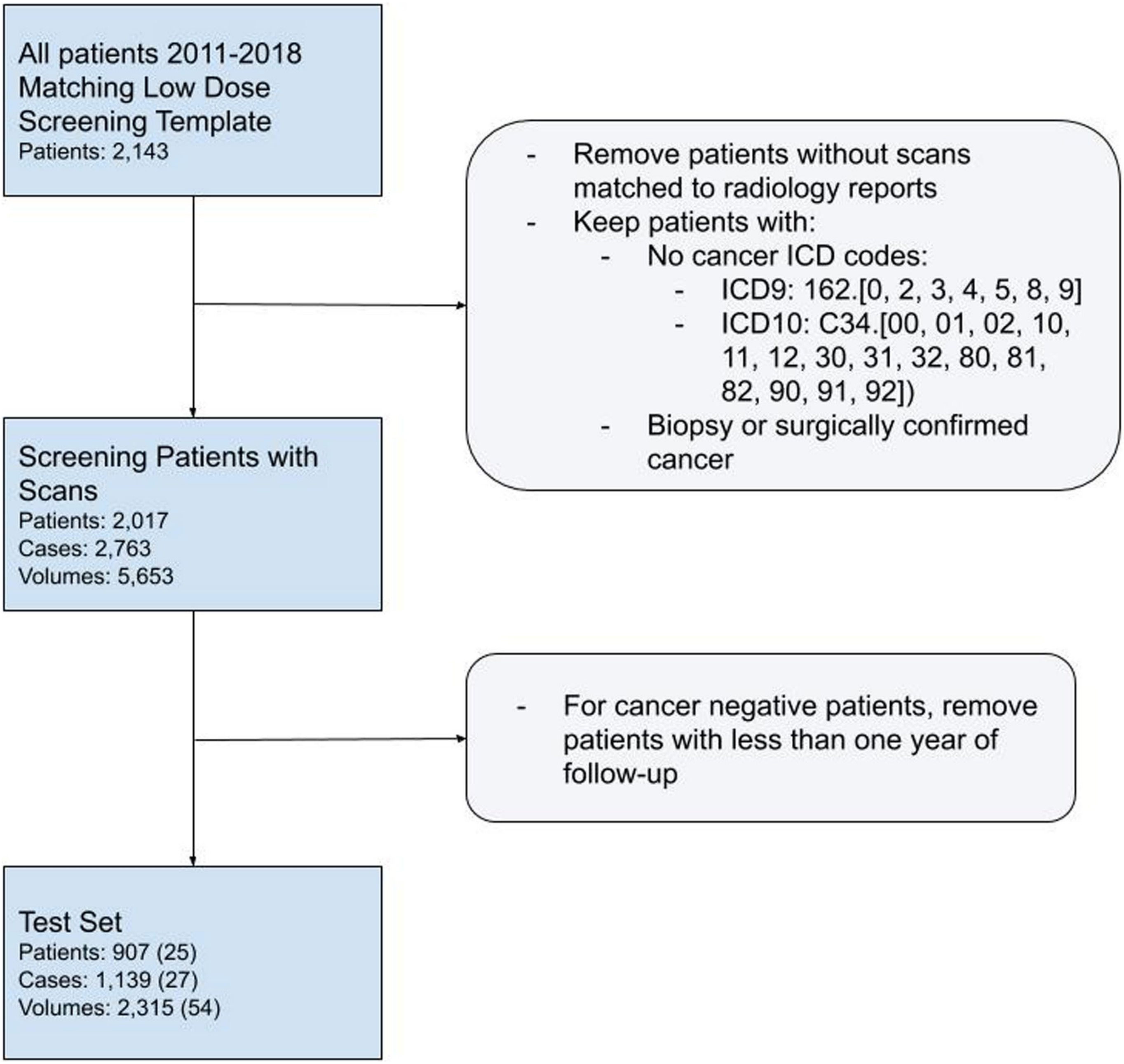
Extended Data Fig. 6 | Examples of ROIs from the detection model and examples of cases where the model prediction differs from the consensus grade. **a**, Example slices from cancer ROIs (cyan) determined by bounding boxes (red) detected by the cancer ROI detection model. The final classification model uses the larger additional context as input illustrated by the cyan ROI. **b**, Sample alignment of prior CT with current CT based on the detected cancer bounding box, which is performed by centering both sub-volumes at the center of their respective detected bounding boxes. When a prior detection is not available, the lung center is used for an approximate alignment. Note that features derived from this large, 90-mm^3 context are compared for classification at a late stage in the model after several max-pooling layers that can discard spatial information. Therefore, a precise voxel-to-voxel alignment is not necessary. **c**, Example cancer-negative case with scarring that was correctly downgraded from a consensus grade of Lung-RADS 4B to LUMAS 1/2 by the model. **d**, Example cancer-positive case with a nodule (size graded as 7–12 mm, depending on the radiologist) correctly upgraded from grades of Lung-RADS 3 and 4A (depending on the radiologist) to LUMAS 4B/X by the model.


a) Attributions for example cancer positive case
Full Volume Model Input
Top Cancer Candidate ROI
Full Volume Model Attribution
Top Cancer Candidate ROI Attribution
b) Attributions for example cancer negative case
Top Cancer Candidate ROI
Top Cancer Candidate ROI Attribution

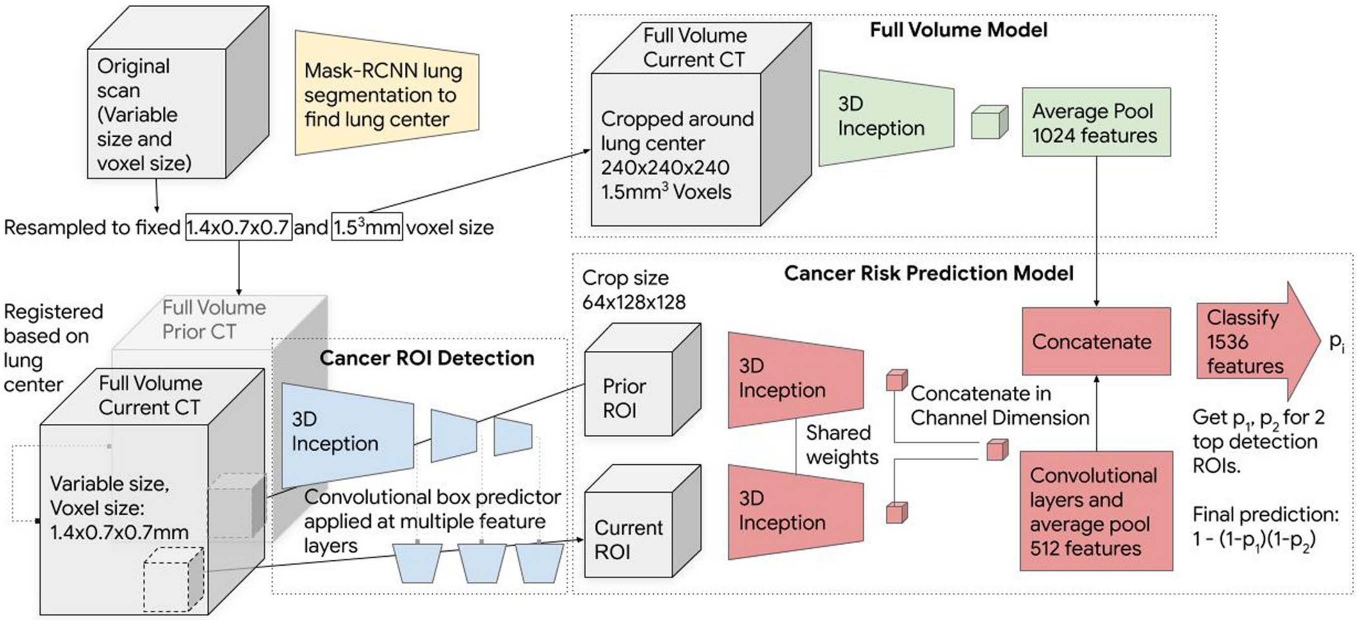
Extended Data Fig. 7 | Attribution maps generated using integrated gradients. a, Example of model attributions for a cancer-positive case. The top row shows the input volume for the full-volume and cancer risk prediction models, respectively. The lower row shows the attribution overlay with positive (magenta) and negative (blue) region contributions to the classifications. In all cancer cases under the attributions study, the readers strongly agreed that the model focused on the nodule. Also, in 86% of these cases, the global and second-stage models focused on the same region. **b**, Example of model attributions for a cancer-negative case. The left-hand image shows a slice from the input subset volume. The right-hand image image shows positive (magenta) and negative (blue) attributions overlayed. The readers found that, in 40% of the negative cases examined, the model focused on vascular regions in the parenchyma.



Extended Data Fig. 8 | Example LUMAS false positive cases. a, 4B/X false positives. **b**, 4A+ false positives.


All patients 2011-2018
Matching Low Dose
Screening Template
Patients: 2,143
- Remove patients without scans matched to radiology reports
- Keep patients with:
- No cancer ICD codes:
- ICD9: 162.[0, 2, 3, 4, 5, 8, 9]
- ICD10: C34.[00, 01, 02, 10, 11, 12, 30, 31, 32, 80, 81, 82, 90, 91, 92])
- Biopsy or surgically confirmed cancer
Screening Patients with
Scans
Patients: 2,017
Cases: 2,763
Volumes: 5,653
- For cancer negative patients, remove patients with less than one year of follow-up
Test Set
Patients: 907 (25)
Cases: 1,139 (27)
Volumes: 2,315 (54)

Extended Data Fig. 9 | STARD diagram of low-dose-screening CT patients from an academic medical center used for the independent validation test set. We require a minimum of 1 year of follow-up for cancer-negative cases. This resulted in a median follow-up time of 625 d across all patients once all exclusion criteria were taken into account. To clarify, this means that the median amount of time from the first screening CT to either a cancer diagnosis or the last follow-up event was 625 d. There were 209 patients (232 cases) with priors in this set of 1,139.


Original scan (Variable size and voxel size)
Mask-RCNN lung segmentation to find lung center
Resampled to fixed 1.4x0.7x0.7 and 1.5³mm voxel size
Full Volume Model
Full Volume Current CT
Cropped around lung center 240x240x240 1.5mm³ Voxels
3D Inception
Average Pool 1024 features
Registered based on lung center
Full Volume Prior CT
Full Volume Current CT
Variable size, Voxel size: 1.4x0.7x0.7mm
Cancer ROI Detection
3D Inception
Convolutional box predictor applied at multiple feature layers
Crop size 64x128x128
Prior ROI
Current ROI
Cancer Risk Prediction Model
3D Inception
Shared weights
3D Inception
Concatenate in Channel Dimension
Concatenate
Convolutional layers and average pool 512 features
Classify 1536 features
pi
Get p1, p2 for 2 top detection ROIs.
Final prediction: 1 - (1-p1)(1-p2)

Extended Data Fig. 10 | Illustration of the architecture of the end-to-end cancer risk prediction model. The model is trained to encompass the entire CT volume and automatically produce a score predicting the cancer diagnosis. In all cases, the input volume is first resampled into two different fixed voxel sizes as shown. Two ROI detections are used per input volume, from which features are extracted to arrive at per-ROI prediction scores via a fully connected neural network. The prior ROI is padded to all zeros when a prior is not available.

natureresearch

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