Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Deep Learning for Lung Cancer Detection on Screening CT Scans: Results of a Large-Scale Public Competition and an Observer Study with 11 Radiologists

Colin Jacobs, PhD • Arnaud A. A. Setio, PhD • Ernst T. Scholten, PhD • Paul K. Gerke, MS • Haimasree Bhattacharya, PhD • Firdaus A. M. Hoesein, PhD • Monique Brink, PhD • Erik Ranschaert, PhD • Pim A. de Jong, PhD • Mario Silva, PhD • Bram Geurts, MD • Kaman Chung, PhD • Steven Schalekamp, PhD • Joke Meersschaert, MD • Anand Devaraj, PhD • Paul F. Pinsky, PhD • Stephen C. Lam, PhD • Bram van Ginneken, PhD • Keyvan Farahani, PhD

From the Department of Radiology, Nuclear Medicine and Anatomy, Radboud University Medical Center, Geert Grooteplein 10, 6525 GA, Nijmegen, the Netherlands (C.J., A.A.A.S., E.T.S., P.K.G., H.B., M.B., B.G., S.S., B.v.G.); Department of Digital Technology & Innovation, Siemens Healthineers, Erlangen, Germany (A.A.A.S.); Department of Radiology, University Medical Center Utrecht, Utrecht, the Netherlands (F.A.M.H., P.A.d.J.); ETZ (Elisabeth-TweeSteden Ziekenhuis), Tilburg, the Netherlands (E.R.); Section of Radiology, Department of Medicine and Surgery (DiMeC), University of Parma, Parma, Italy (M.S.); Department of Radiology, Meander Medical Center, Amersfoort, the Netherlands (K.C., S.S.); Department of Radiology, AZ Zeno, Knokke-Heist, Belgium (J.M.); Department of Imaging, Royal Brompton Hospital, London, England (A.D.); Division of Cancer Prevention (P.F.P.) and Center for Biomedical Informatics & Information Technology (K.F.), National Cancer Institute, National Institutes of Health, Bethesda, Md; British Columbia Cancer Agency and the University of British Columbia, Vancouver, Canada (S.C.L.); and Fraunhofer MEVIS, Bremen, Germany (B.v.G.). Received January 19, 2021; revision requested May 4; revision received October 11; accepted October 13. **Address correspondence to** C.J. (e-mail: *colin.jacobs@radboudumc.nl*).

Supported by a research grant of MeVis Medical Solutions, Bremen, Germany.

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2021; 3(6):e210027 • https://doi.org/10.1148/ryai.2021210027 • Content codes: AI CT

Purpose: To determine whether deep learning algorithms developed in a public competition could identify lung cancer on low-dose CT scans with a performance similar to that of radiologists.

Materials and Methods: In this retrospective study, a dataset consisting of 300 patient scans was used for model assessment; 150 patient scans were from the competition set and 150 were from an independent dataset. Both test datasets contained 50 cancer-positive scans and 100 cancer-negative scans. The reference standard was set by histopathologic examination for cancer-positive scans and imaging follow-up for at least 2 years for cancer-negative scans. The test datasets were applied to the three top-performing algorithms from the Kaggle Data Science Bowl 2017 public competition: grt123, Julian de Wit and Daniel Hammack (JWDH), and Aidence. Model outputs were compared with an observer study of 11 radiologists that assessed the same test datasets. Each scan was scored on a continuous scale by both the deep learning algorithms and the radiologists. Performance was measured using multireader, multicase receiver operating characteristic analysis.

Results: The area under the receiver operating characteristic curve (AUC) was 0.877 (95% CI: 0.842, 0.910) for grt123, 0.902 (95% CI: 0.871, 0.932) for JWDH, and 0.900 (95% CI: 0.870, 0.928) for Aidence. The average AUC of the radiologists was 0.917 (95% CI: 0.889, 0.945), which was significantly higher than grt123 ($P = .02$); however, no significant difference was found between the radiologists and JWDH ($P = .29$) or Aidence ($P = .26$).

Conclusion: Deep learning algorithms developed in a public competition for lung cancer detection in low-dose CT scans reached performance close to that of radiologists.

Supplemental material is available for this article.

Lung cancer is the leading cause of cancer-related death worldwide (1). Although smoking rates continue to decline in most developed countries, a substantial portion of the population remains at high risk for lung cancer (2). Two large randomized controlled trials, the National Lung Screening Trial in the United States and the Dutch-Belgian NELSON trial (NELSON is a Dutch acronym for "Nederlands-Leuvens Longkanker Screenings Onderzoek"), demonstrated that annual screening with low-dose CT of individuals at high risk led to a reduction in lung cancer mortality in the screening groups compared with the control groups (3,4). After the National Lung Screening Trial and the subsequent positive reimbursement recommendation by regulatory organizations, lung cancer screenings with low-dose CT in populations at high risk are being implemented in the United States (5,6). Following the positive results of the National Lung Screening Trial and NELSON trial, as well as the positive recommendations by societies such as the European Society of Radiology and the European Respiratory Society (7,8), many European countries are considering initiating screening programs, as well.

Within a screening program, participants undergo annual low-dose CT scanning. Based on the interpretation of the reading radiologist, the screening panel will determine whether the screening test is positive or negative, as well as what type of follow-up is needed. Classification schemes have been adopted to standardize CT reporting

Abbreviations

AUC = area under the ROC curve, JWDH = Julian de Wit and Daniel Hammack, DSB2017 = Kaggle Data Science Bowl 2017, Lung-RADS = Lung Imaging Reporting and Data System, PanCan = Pan-Canadian Lung Screening trial, ROC = receiver operating characteristic

Summary

An observer study showed that two of the three top-performing algorithms from a public competition (Kaggle Data Science Bowl 2017) attained performances that were not significantly worse than that of 11 radiologists for estimating lung cancer risk on low-dose CT scans.

Key Points

- An enriched dataset of 300 chest CT scans (100 cancer-positive and 200 cancer-negative scans) was assessed in an observer study of radiologists; these same scans were then input into the three top-performing models (ie, grt123, Julian de Wit and Daniel Hammack [JWDH], Aidence) from the Kaggle Data Science Bowl 2017 to assess lung cancer risk.
- The average area under the receiver operating characteristic curve (AUC) was 0.917 (ranging from 0.841 to 0.944) for the radiologists, and the model AUCs were 0.876 for grt123, 0.883 for JWDH, and 0.881 for Aidence.
- The AUC was lower for the grt123 model than for the radiologists (P = .02), whereas there was no evidence of a difference in AUC between the radiologists and either JWDH (P = .29) or Aidence (P = .26).

Keywords

Lung, CT, Thorax, Screening, Oncology

and management recommendations. The Lung Imaging Reporting and Data System (Lung-RADS) classification is used for interpreting chest CT screening scans in the United States (9). Lung-RADS has six possible categories—1, 2, 3, 4A, 4B, and 4×—and these categories correspond to four different levels of follow-up. With the Lung-RADS classification, the radiologist must identify all pulmonary nodules in the CT scan and assess their size and type. Based on this information, the radiologist assigns the corresponding Lung-RADS category.

Accurate detection and risk assessment of pulmonary nodules is a crucial component of a successful screening program. However, reading of screening CT scans by radiologists is time-consuming and expensive and is associated with substantial interobserver variability (10,11). Computer algorithms, particularly artificial intelligence algorithms, may help by assisting in the detection and classification of lung nodules, by assisting in the three-dimensional measurement of lung nodules, or by providing an accurate lung cancer risk estimate for each pulmonary nodule or at the scan level. Such computer support may help optimize the interpretation of screening CT scans and may lead to better management recommendations that increase the cost-effectiveness of screening.

In 2017, the high-profile Kaggle Data Science Bowl was launched (*https://www.kaggle.com/c/data-science-bowl-2017*). This competition challenged computer and data scientists to develop software that could accurately determine when lesions in the lungs are cancerous. The top 10 teams in the final ranking of the competition received monetary prizes on the condition that they made the software they had written publicly available under an open-source license. A leaderboard with the results of this competition is available; however, at the time it was unknown whether these algorithms could have value in a clinical lung cancer screening setting. As such, there were two key questions to address: how the performance of the computer systems compares with that of radiologists, and if the algorithms perform well on independent datasets from sources that were not used in the Kaggle Data Science Bowl 2017 (DSB2017) competition.

The purpose of this study, therefore, was to assess and compare the performance of the models to that of radiologists for identifying lung cancer on low-dose chest CT scans using scans from the DSB2017 test set and scans from an independent external dataset.

Materials and Methods

The DSB2017 Competition

The DSB2017 competition was set up in two stages: (*a*) model development, training, and testing on an initial test set of 198 scans and (*b*) model testing on an additional set of 500 scans. The test performance on the 500 scans was used to determine the final ranking of the submitted algorithms. The metric used for ranking was the logarithmic loss metric (hereafter, referred to as logloss), which measures the performance of a binary classification when the prediction is a probability value between 0 and 1. In short, logloss is the logarithmic transform of the sum of the probabilities that an algorithm assigns to the samples that it misclassifies. Further details about the setup of the competition can be found in Appendix E1 (supplement).

Datasets

The scans for DSB2017 originated from the National Lung Screening Trial (*ClinicalTrials.gov*: NCT00047385), the Danish Lung Cancer Screening Trial (*ClinicalTrials.gov*: NCT00496977), and the screening program at the Lahey Hospital and Medical Center (Burlington, Mass). Institutional review board approval and informed consent were obtained during inclusion into the screening trials. Scans that were positive for cancer were selected from individuals diagnosed with lung cancer and confirmed at histopathologic examination. For the cancer-positive scans, the screening CT scan obtained before the lung cancer diagnosis was included. Only scans in patients for whom the diagnosis followed within 1 year of the CT scan were included. Noncancer scans were selected from individuals who did not have a lung cancer diagnosis during the course of the screening program and for whom the minimum follow-up period was 2 years. The noncancer scans were enriched with scans in which suspicious pulmonary nodules (Lung-RADS 3 and above) were present to ensure that the dataset would not contain a large proportion of CT scans without any nodules present. Each CT scan originated from a different participant. All images were supplied in Digital Imaging and Communications in Medicine format.

The DSB2017 attracted considerable interest, with almost 2000 teams joining the competition. All winning teams publicly released their code under a permissive open-source license.

We summarized the links to the code in Table E1 (supplement). The labels of the final test set of 500 scans were released by Kaggle several weeks after the competition ended. For this study, the raw scores of the top 10 algorithms were obtained from Kaggle, and receiver operating characteristic (ROC) analysis was performed to achieve a more clinically useful analysis of the results of the competition.

External Validation Data

Data from the Pan-Canadian Early Detection of Lung Cancer (PanCan) trial (*ClinicalTrials.gov*: NCT00751660) were included in this study as external validation data. The PanCan trial enrolled 2537 participants, and all low-dose baseline CT scans were considered for inclusion in this study. Details on the population and CT scanning protocol are provided in previously published studies (14,15). The data were acquired at five institutions across Canada between 2008 and 2010. Lung cancer status was set on the basis of histopathologic examination for cancer-positive scans and imaging follow-up for at least 2 years for noncancer scans. The PanCan dataset is an independent test dataset; the data are not publicly available and none of the top 10 teams from the DSB2017 had access to these scans.

To obtain an external validation of the algorithms, the code of the top 10 algorithms was downloaded and then reviewed by a team of experienced research software engineers. For the top three solutions—grt123, Julian de Wit and Daniel Hammack (JWDH), and Aidence—the team compiled a software package that could be used to process unseen CT scans. All systems required training deep learning networks on DSB2017 training data and, in some cases, on additional data such as the Lung Image Database Consortium (12) or the LUNA16 (13) data (see open-source code links in Table E1 [supplement]). To verify code correctness, the scores on the DSB2017 test set were recomputed and a correlation test between the scores that were submitted to DSB2017 and the recomputed scores was performed (see Appendix E2 and Table E2 [supplement]). Subsequently, the grt123, JWDH, and Aidence models were applied to all baseline CT scans from the PanCan dataset.

Observer Study

We conducted a retrospective observer study with a dataset of 300 CT scans. Because we were unsure of the effect size prior to this study, we did not perform power calculations; instead, we determined the number of scans based on a trade-off between sample size and reading effort. An enriched set was chosen to prevent compiling a dataset in which the vast majority of scans would be normal, which would have hindered a proper ROC analysis. We included 150 scans from the DSB2017 test set containing 50 cancer-positive scans, as well as 150 baseline CT scans from the PanCan dataset containing 50 cancer-positive scans. The 100 cancer-negative and 50 cancer-positive scans from the DSB2017 test set were randomly extracted from the full DSB2017 test set, which contained 349 cancer-negative and 151 cancer-positive scans. For the PanCan data, we randomly selected 50 cancer-positive scans from patients who were diagnosed with lung cancer within 1 year after the included CT scan was obtained. For the remaining 100 scans negative for cancer, we randomly selected 50 cancer-negative scans with a nodule larger than 8 mm and 50 random cancer-negative scans without a nodule larger than 8 mm, as annotated by the PanCan radiologists during the screening trial. Patients with multiple lung cancers were not included in this study. Note that this dataset is enriched in a similar fashion to the enrichment of the DSB2017 data. This dataset allowed us to compare the performance of the algorithms with that of radiologists on the DSB2017 test data and on an external dataset. Details on the nodule characteristics for the included scans can be found in Table E3 (supplement).

A web-accessible workstation was developed on which radiologists could review chest CT scans. The web workstation includes common tools found in a professional medical viewing workstation, such as the ability to scroll through the scan in any orthogonal direction, change the window and level settings, pan and zoom, change section thickness, and select sliding maximum and minimum projections. The reading time per scan was measured by the software. No time limit was imposed.

Radiologists and radiology residents with experience in chest reading were approached for this study. Participation was on a voluntary basis. After agreeing to participate, the radiologists received written instructions and filled out a questionnaire about their professional experience and their experience with reading lung cancer screening CT scans. In total, 11 readers participated (M.B., K.C., A.D., B.G., F.A.M.H., P.A.d.J., J.M., E.R., E.T.S., M.S., and S.S.), with varying levels of experience. They consisted of seven radiologists with 7–30 years of experience in chest radiology; two general radiologists with 7–25 years of experience in radiology; and two radiology residents, each with more than 2 years of experience in chest radiology. For each scan, the radiologist had to assign a score on a continuous scale between 0 (very low likelihood) and 100 (very high likelihood) for being indicative of cancer. The radiologists were encouraged to use the full scale. The radiologists were informed that about one-third of the scans were cancer positive, but they were blinded to clinical information and the results of the algorithms. They were instructed to start with a training batch of 30 scans, taken from the remaining scans from the PanCan set and the DSB2017 test set, to become acquainted with the web viewing environment and the task. After each scan, the observer could decide either to continue or to stop and continue at a later time.

Statistical Analysis

The primary outcome measure in this study was the area under the ROC curve (AUC) for presence of malignancy. We compared the performance of each of the top three algorithms independently with the radiologists using multireader, multicase ROC analysis. We used the publicly available iMRMC software (version 4.0.3; U.S. Food and Drug Administration) (*https://github.com/DIDSR/iMRMC/releases/tag/iMRMC-v4.0.3*). The average ROC curve of the radiologists was computed using the diagonal average, which is area preserving, meaning that the AUC of the average ROC curve equals the average of all the separate AUC values of the individual ROC curves. The ROC

curve from each of the top three algorithms was compared with the average radiologist ROC curve on the full set of 300 scans from the observer study using the methods by Gallas et al (16) implemented in iMRMC. This method uses *U* statistics to provide unbiased estimates of the variance components, and we used the method that decomposes the total variance into eight moments from first principles. In the multireader, multicase analysis, the first modality was set to radiologists analyzing the CT images, and the deep learning algorithm was set as the independent second modality, representing an algorithm analyzing the CT images. The 95% CIs for the individual ROC curves were computed using bootstrapping with 1000 iterations using a Python (version 3.6; Python Software Foundation; *https://www.python.org/*) (open source) script developed in house.

Table 1: Top 10 Teams from the DSB2017 and Model Performance Statistics

Rank	Team Name	Logloss Score	AUC
1	grt123	0.39975	0.876
2	Julian de Wit & Daniel Hammack	0.40117	0.883
3	Aidence	0.40127	0.881
4	qfpxfd	0.40183	0.877
5	Pierre Fillard (Therapixel)	0.40409	0.879
6	MDai	0.41629	0.871
7	DL Munich	0.42751	0.855
8	Alex Andre Gilberto Shize	0.43019	0.858
9	Deep Breath	0.43872	0.849
10	Owkin Team	0.44068	0.876

Note.—Results are shown from the 500 test scans from the Kaggle Data Science Bowl 2017 (DSB2017). AUC = area under the receiver operating characteristic curve, logloss = logarithmic loss metric.

Results

DSB2017 Competition Results

The results of the competition are shown in Table 1, where the performance of the top 10 algorithms on the DSB2017 test set of 500 scans was ranked using the logloss metric. The AUCs are also included in Table 1. The ROC curves corresponding to these submissions on the 500 test scans from the competition test set are depicted in Figure 1. The AUC values of the top 10 algorithms were high and ranged between 0.849 and 0.883.

Observer Experiment

For the top three solutions (grt123, JWDH, and Aidence), software packages that can process unseen CT scans were compiled, and the correlation scores between the recomputed and the submitted scores of the algorithms were all above 0.99 (Table E2 [supplement]). The grt123, JWDH, and Aidence models were all successfully applied to all 300 CT scans from the observer experiment.

All readers completed the full set of 300 scans of the observer experiment, and the average reading time per scan ranged from 96 seconds to 275 seconds. The AUC values were 0.877 (95% CI: 0.842, 0.910) for grt123, 0.900 (95% CI: 0.870, 0.928) for Aidence, and 0.902 (95% CI: 0.871, 0.932) for JWDH when the models were assessed on all 300 scans (Table 2). For the radiologists, the AUCs ranged from 0.841 (95% CI: 0.800, 0.882) to 0.944 (95% CI: 0.923, 0.963), with an average AUC of 0.917 (95% CI: 0.889, 0.945). The ROC curves from the top three algorithms on the 150 scans from the DSB2017 dataset, the 150 scans from the PanCan dataset, and the full dataset from the observer study are shown in Figure 2. In these figures, the ROC curve of the average reader performance is also plotted. Individual ROC curves for all observers are depicted in Figure E1 (supplement). The top three algorithms showed good performance, and no performance drop was seen on the independent validation data (Fig 2C compared with Fig 2B). Figure 2A shows the performance of the top three algorithms on the full set of 300 scans that were included in the observer experiment. The statistical analysis showed that the average AUC among the 11 radiologists was higher than that of the grt123 algorithm (P = .02); whereas the AUCs from the other two models were not significantly worse compared with those of the radiologists (JWDH, P = .29; and Aidence, P = .26) (Table 2).

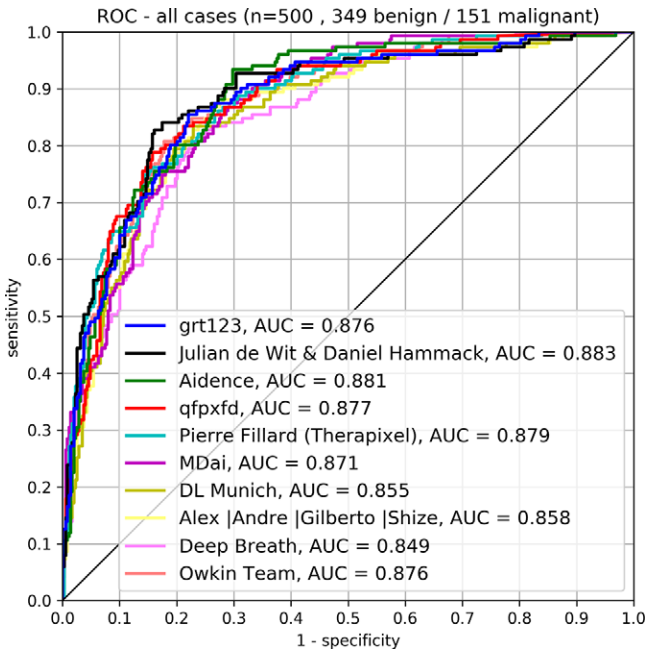

Figure 1: Receiver operating characteristic (ROC) curves for the top 10 algorithms on the 500 scans from the Kaggle Data Science Bowl 2017 test set. AUC = area under the ROC curve.

Discussion

The reading of screening CT scans by radiologists is an important component of a lung cancer screening program. Integrating support by artificial intelligence tools into this reading

Table 2: Performance of Readers and Top Three Algorithms

Parameter	DSB2017	PanCan	All Scans	*P* Value*
Reader				
R1 (Che)	0.909 (0.872, 0.945)	0.928 (0.888, 0.963)	0.918 (0.891, 0.944)	NA
R2 (Che)	0.899 (0.856, 0.943)	0.970 (0.950, 0.987)	0.938 (0.915, 0.959)	NA
R3 (Che)	0.894 (0.852, 0.935)	0.938 (0.903, 0.969)	0.919 (0.891, 0.944)	NA
R4 (Rad)	0.837 (0.782, 0.893)	0.850 (0.789, 0.905)	0.841 (0.800, 0.882)	NA
R5 (Che)	0.920 (0.881, 0.957)	0.966 (0.945, 0.985)	0.944 (0.923, 0.963)	NA
R6 (Res)	0.871 (0.823, 0.915)	0.951 (0.922, 0.974)	0.911 (0.883, 0.937)	NA
R7 (Res)	0.918 (0.877, 0.958)	0.953 (0.924, 0.976)	0.935 (0.910, 0.958)	NA
R8 (Rad)	0.899 (0.858, 0.936)	0.927 (0.888, 0.961)	0.913 (0.884, 0.937)	NA
R9 (Che)	0.930 (0.895, 0.963)	0.945 (0.910, 0.974)	0.939 (0.913, 0.960)	NA
R10 (Che)	0.927 (0.886, 0.965)	0.938 (0.901, 0.970)	0.932 (0.908, 0.958)	NA
R11 (Che)	0.856 (0.801, 0.908)	0.935 (0.900, 0.966)	0.897 (0.864, 0.928)	NA
Average reader	0.896 (0.853, 0.939)	0.937 (0.907, 0.966)	0.917 (0.889, 0.945)	NA
Algorithm				
grt123	0.845 (0.790, 0.894)	0.905 (0.862, 0.944)	0.877 (0.842, 0.910)	.02
JWDH	0.880 (0.831, 0.924)	0.920 (0.879, 0.954)	0.902 (0.871, 0.932)	.29
Aidence	0.885 (0.842, 0.927)	0.917 (0.877, 0.950)	0.900 (0.870, 0.928)	.26

Note.—There were 150 DSB2017 scans and 150 PanCan scans. Unless otherwise indicated, data are areas under the receiver operating characteristic curve; data in parentheses are 95% CIs. AUC = area under the receiver operating characteristic curve, Che = chest radiologist, JWDH = Julian de Wit and Daniel Hammack, DSB2017 = Kaggle Data Science Bowl 2017, NA = not applicable, PanCan = Pan-Canadian Lung Screening trial, Rad = radiologist, R = reader, Res = resident.
**P* value is shown for the comparison of the algorithm AUC with the average reader AUC.

process may improve the efficiency and accuracy of screening. The Kaggle DSB2017 competition resulted in 10 open-source algorithms that were capable of detecting lung cancer on a CT scan. At the time of the competition, however, the algorithms were not compared with readings by radiologists, and thus it was difficult to assess the clinical effect of these algorithms. To address this gap, we performed a study to compare the top-performing algorithms to radiologists. We found that two of the top three algorithms from the DSB2017 competition were able to provide lung cancer risk predictions based on a CT scan with a performance close to that of radiologists. These results, which were found by using a test dataset from the original competition and an independent test dataset, offer several opportunities to optimize the reading of lung cancer screening CT scans.

The algorithms used in this study produce a score between 0 and 1 for each CT scan that indicates the likelihood that the participant will have a lung cancer diagnosis within 1 year. No location of a possible cancer or explanation for the score of the deep learning models is provided, however. Therefore, currently this score can only be used as a sign that a radiologist needs to carefully check the CT scan for abnormalities. Potentially, direct estimation of the malignancy risk may one day be an effective way to optimize current guidelines. Alternatively, these algorithms could be used to triage normal scans with the result that only possibly abnormal scans are sent for radiologist review. Such triaging may have a substantial effect on the cost effectiveness of screening; however, this has not yet been investigated and thus, there is no scientific evidence supporting this idea. If future validation studies show that this approach is feasible, policy changes will be needed because to qualify for reimbursement in the United States, every screening CT scan must be categorized according to Lung-RADS by a board-certified radiologist.

Future development of the deep learning models should focus on providing more information to the user (eg, the location of the suspicious pulmonary nodules that have been found by the model). This should be feasible to accomplish because winning solutions used an approach in which they first detected lung nodule candidate locations and subsequently used the detected locations to produce a malignancy risk score at the scan level. Subsequently, studies are needed that focus on evaluating how the use of these algorithms can be integrated with the work of radiologists to positively change the follow-up recommendations in a screening program.

In this study, we tested only the discriminative power of the algorithms. The performance of the algorithms was evaluated on an enriched set of scans, which does not reflect disease prevalence in real-world lung cancer screening practice. Therefore, to guarantee accurate and reliable predictions in real-world situations, the predictions of the algorithms will need to be calibrated and future studies should investigate the most optimal cutoff points for decision-making.

The performance of the algorithms and the observers was highest on the subset of 150 scans from the PanCan trial. Because

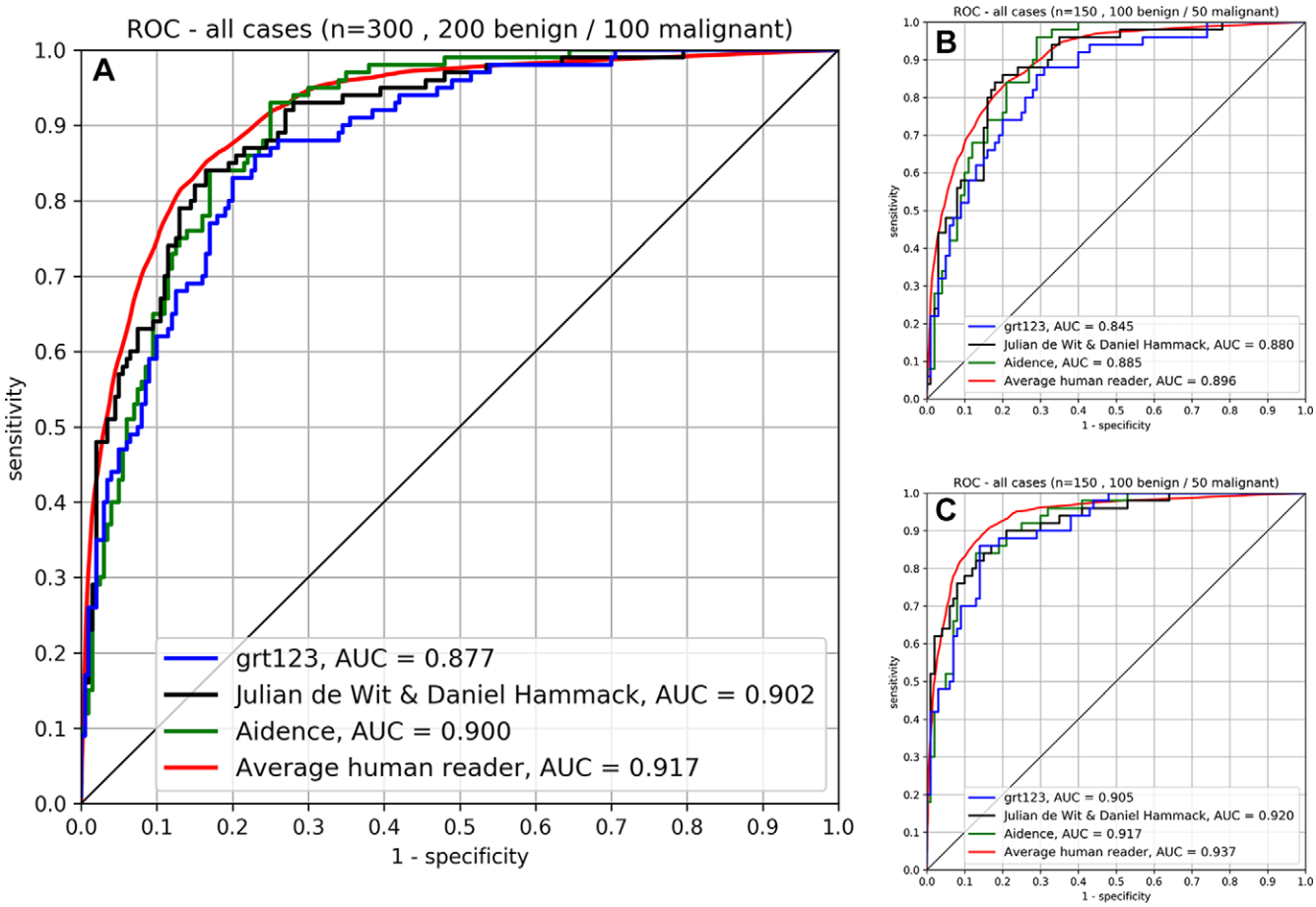

Figure 2: Receiver operating characteristic (ROC) curves for the performance of the top three algorithms and the average radiologist on **(A)** the full set of 300 scans, **(B)** the subset of 150 scans from the Kaggle Data Science Bowl 2017 dataset, and **(C)** the subset of 150 scans from the Pan-Canadian Lung Screening Trial dataset from the observer study. AUC = area under the ROC curve.

this set of scans was enriched in a way that was similar to that of the DSB2017 data, we do not expect that the difference can be explained by case selection. Potentially, the standardized CT imaging protocol of the PanCan trial (1.25-mm section thickness) played a role, but this is speculation.

The deep learning models developed in the competition were created in a relatively short amount of time (3 months) and with a predefined, relatively small training dataset consisting of approximately 400 cancer scans and 1000 benign scans (*https://www.kaggle.com/c/data-science-bowl-2017*). What the performance of the deep learning models will be when more data and more time are available to train and develop these models is unknown.

A limitation of the developed deep learning models is that they use only one CT scan per patient. In a lung cancer screening setting, multiple prior scans are often available in addition to the current scan. The availability of these different scans is important as growth of a nodule on CT is the most important predictor of cancer, and growth cannot be assessed from a single scan. Future research should focus on developing computer solutions for the scenario in which both current and prior scans are available. The 2019 study by Ardila et al (17) is a good first example and showed a performance similar to or higher than that of radiologists, but it remains unclear how this proprietary algorithm would be used in screening (18). The authors of another recent study designed a neural network with clinical features and imaging features annotated by radiologists as input to assess the lung cancer risk of follow-up CT scans and showed good performance on an independent dataset (19).

This study showed that two of the top three algorithms from the DSB2017 competition achieved a performance not significantly worse than the average of the performance of 11 radiologists for estimating the cancer risk from a CT scan on a dataset consisting of scans from the test set of the competition and from an independent validation set. Further research should focus on how these artificial intelligence models can be used most effectively to assist in the interpretation of lung cancer screening CT scans in a screening setting.

Acknowledgments: The authors thank Kaggle and Booz Allen Hamilton for their collaboration in this project. We thank all the sponsors of DSB2017 for their support for the competition.

Author contributions: Guarantor of integrity of entire study, C.J.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, C.J., A.A.A.S., P.K.G., M.S., S.C.L., B.v.G., K.F.; clinical studies, F.A.M.H., E.R., M.S., B.G., K.C., S.S., A.D., S.C.L., B.v.G., ; experimental studies, C.J., A.A.A.S., P.K.G., M.B., M.S., S.S., B.v.G., K.F.; statistical analysis, C.J., P.K.G.;

and manuscript editing, C.J., A.A.A.S., E.T.S., F.A.M.H., M.B., M.S., K.C., S.S., P.F.P., S.C.L., B.v.G., K.F.

Disclosures of conflicts of interest: **C.J.** Royalties from Veolity from MeVis Medical Solutions. **A.A.A.S.** Patents planned, issued, or pending from Siemens Healthineers; stock or stock options from Siemens Healthineers. **E.T.S.** No relevant relationships. **P.K.G.** No relevant relationships. **H.B.** No relevant relationships. **F.A.M.H.** No relevant relationships. **M.B.** Payment for speaker bureau/speaker CT webinars from Canon Medical Systems Europe. **E.R.** No relevant relationships. **P.A.d.J.** Departmental grants from Philips Healthcare, Sanifit. **M.S.** No relevant relationships. **B.G.** Stock or stock options from CRISPR Technologies, C2.ai, Teladoc, Nanostring technologies, Bionano genomics, Infinity Pharmaceuticals. **K.C.** No relevant relationships. **S.S.** No relevant relationships. **J.M.** No relevant relationships. **A.D.** Consulting fees from Brainomix. **P.F.P.** No relevant relationships. **S.C.L.** Grants from Terry Fox Research Institute, British Columbia Cancer Foundation, and VGH-UBC Hospital Foundation; expert advisor for Canadian Partnership Against Cancer; Chair, Pan-Canadian Lung Screening Network. **B.v.G.** Grants from the Dutch Science Foundation; royalties or licenses from Thirona, MeVis Medical Solutions, Delft Imaging; stock or stock options in Thirona; cofounder, Thirona. **K.F.** No relevant relationships.

References

1. World Health Organization. The top 10 causes of death. https://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death. Published December 2020. Accessed November 15, 2021.
2. World Health Organization. Global Health Observatory (GHO) data. https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/gho-tobacco-control-monitor. Published 2016. Accessed November 15, 2021.
3. National Lung Screening Trial Research Team; Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365(5):395–409.
4. de Koning HJ, van der Aalst CM, de Jong PA, et al. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. N Engl J Med 2020;382(6):503–513.
5. Centers for Disease Control and Prevention. Screening for Lung Cancer with Low Dose Computed Tomography (LDCT) (CAG-00439N). https://www.cms.gov/medicare-coverage-database/details/nca-decision-memo.aspx?NCAId=274. Published February 5, 2015. Accessed May 20, 2020.
6. Moyer VA; U.S. Preventive Services Task Force. Screening for lung cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2014;160(5):330–338. doi:10.7326/M13-2771.
7. Kauczor HU, Bonomo L, Gaga M, et al. ESR/ERS white paper on lung cancer screening. Eur Respir J 2015;46(1):28–39.
8. Kauczor HU, Baird AM, Blum TG, et al. ESR/ERS statement paper on lung cancer screening. Eur Radiol 2020;30(6):3277–3294.
9. Lung-RADS Assessment Categories. Version 1.1. American College of Radiology. Lung CT Screening Reporting & Data System (Lung-RADS). https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads. Published 2019. Accessed May 20, 2020.
10. van Riel SJ, Jacobs C, Scholten ET, et al. Observer variability for Lung-RADS categorisation of lung cancer screening CTs: impact on patient management. Eur Radiol 2019;29(2):924–931.
11. Pinsky PF, Gierada DS, Nath PH, Kazerooni E, Amorosa J. National lung screening trial: variability in nodule detection rates in chest CT studies. Radiology 2013;268(3):865–873.
12. Armato SG 3rd, McLennan G, Bidaut L, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 2011;38(2):915–931.
13. Setio AAA, Traverso A, de Bel T, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. Med Image Anal 2017;42:1–13.
14. Tammemagi MC, Schmidt H, Martel S, et al. Participant selection for lung cancer screening by risk modelling (the Pan-Canadian Early Detection of Lung Cancer [PanCan] study): a single-arm, prospective study. Lancet Oncol 2017;18(11):1523–1531.
15. McWilliams A, Tammemagi MC, Mayo JR, et al. Probability of cancer in pulmonary nodules detected on first screening CT. N Engl J Med 2013;369(10):910–919.
16. Gallas BD, Bandos A, Samuelson F, Wagner RF. A Framework for Random-Effects ROC Analysis: Biases with the Bootstrap and Other Variance Estimators. Commun Stat Theory Methods 2009;38(15):2586–2603.
17. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019;25(6):954–961 [Published correction appears in Nat Med 2019;25(8):1319.].
18. Jacobs C, van Ginneken B. Google's lung cancer AI: a promising tool that needs further validation. Nat Rev Clin Oncol 2019;16(9):532–533.
19. Huang P, Lin CT, Li Y, et al. Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method. Lancet Digit Health 2019;1(7):e353–e362.