LUNG CANCER

Google's lung cancer AI: a promising tool that needs further validation

Colin Jacobs and Bram van Ginneken

Researchers from Google AI have presented results obtained using a deep learning model for the detection of lung cancer in screening CT images. The authors report a level of performance similar to, or better than, that of radiologists. However, these claims are currently too strong. The model is promising but needs further validation and could only be implemented if screening guidelines were adjusted to accept recommendations from black-box proprietary AI systems.

Refers to Ardila, D. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).

Lung cancer is the most deadly cancer worldwide. Survival outcomes are often poor because most patients are diagnosed when the disease is already at an advanced stage. Early detection is, therefore, a promising strategy to fight lung cancer. Data from the pivotal National Lung Screening Trial (NLST) showed that regularly screening individuals with a high risk of lung cancer using low-dose CT imaging enables more early stage cancers to be detected and reduces lung cancer-related mortality[1]. CT-based lung cancer screening has been widely implemented in the USA and is under consideration in many other countries[2].

Early stage lung cancer is visible on CT as pulmonary nodules, which are small lesions measuring between 3 mm and 30 mm in diameter. Since the 1990s, researchers have been developing computer algorithms designed to enable the automatic detection of such nodules; however, these computer-aided systems had suboptimal levels of sensitivity and resulted in too many false positives. The advent of deep learning has addressed these failures and recent research shows that convolutional neural networks can enable the detection of nodules with a high level of accuracy, as demonstrated in the LUNA16 challenge[3]. Several products are now on the market that enable the automatic detection of lung nodules. These systems, including Veolity LungCAD, Riverian ClearRead CT and others, have regulatory clearance for use as a second reader or a concurrent reader. Future algorithms could potentially also be used as a first reader, or even for fully automated analysis. In a study using lung cancer screening data from Canada, Ritchie et al.[4] showed that a technician aided by an automatic detection system could accurately categorize scans as 'normal' or 'abnormal' for review by a radiologist. Such 'pre-reading' might be an effective way to reduce the costs of screening.

In 2017, the Kaggle Data Science Bowl awarded a total of US$1 million in prize money for the ten best algorithms that could predict lung cancer from a single screening CT scan[5]. The task was to predict whether an individual would be diagnosed with lung cancer within 1 year of the scan. No localization of the lung cancer was needed. A training data set of approximately 1,500 CT scans, mostly from the NLST, was provided and the final ranking was based on performance in a test set of 500 CT scans from different screening programmes. The challenge required the developers of the winning solutions to make the underlying code available under a permissive open-source licence. Almost 2,000 teams participated and the ten algorithms that showed the most promising performance are now publicly available.

In May 2019, Google AI and collaborators published results obtained using a deep learning model for the prediction of lung cancer from CT scans[6]. Similar to the Kaggle algorithms, this model also aims to predict whether or not the individual will be diagnosed with lung cancer within 1 year. Google also used NLST data for training purposes. In addition to a 'scan score', the model also outputs two locations where the lung cancer might be located and has the option of incorporating a scan taken 1 year before the screening scan into the analysis. These features are important additional components that were not included in the algorithms that emerged from the Kaggle competition. The Google model was internally validated on a 'held-out' set (comprising 15% of the NLST data set) from NLST and externally validated using images from a small independent data set of 1,739 images, including 27 with known cancers. The study compared the developed model to scores from six board-certified US-based radiologists. The authors claimed that their model outperformed radiologists when a single CT scan was analysed and performed similarly to radiologists when a prior scan was also available.

> "a black-box AI-based system that overrules well-established clinical guidelines that prompt radiologists to base management decisions on the size and growth rate of nodules is unlikely to be quickly accepted"

These claims are too strong for several reasons. First, both the model training and validation were conducted using data from NLST. Second, the radiologists used Lung-RADS scores, a scoring system developed by the American College of Radiology, which is a mandatory requirement of lung cancer screening protocols in the USA[7]. However, Lung-RADS scores do not directly correspond to the probability that a lung cancer

diagnosis is made within 1 year but instead provide a management recommendation. Lung-RADS uses cut-offs based on the size, type and growth of the nodules visible on CT, with suspicious nodules assigned to a special category called Lung-RADS 4X. Chung et al.[8] showed that when radiologists are asked to upgrade a patient to Lung-RADS 4X if they think a nodule is indeed cancer, they can correctly identify many cancers from nodules with lower Lung-RADS scores. Radiologists participating in the Google reader study did assess lesions using the 4X category of Lung-RADS but also combined this category with 4B, which does not include the additional suspicious features that a radiologist might detect. Third, the radiologists participating in the reader study were not thoracic radiologists, and whether they had experience in assessing the type of screening CT scans assessed in the study is not stated.

How could Google's model be implemented in clinical practice? The authors propose to use predefined cut-offs on their model output, dubbed lung malignancy scores (LUMAS). Screening protocols in the USA mandate the use of Lung-RADS. However, a high LUMAS could be used to select patients for upgrading to the 4X category. The Lung-RADS guidelines already include a statement that the nodule malignancy tool developed by McWilliams et al.[9] can be used as a decision tool for upgrading, although Lung-RADS currently does not offer the possibility to downgrade lesions. Moreover, using LUMAS could violate the Lung-RADS guidelines by assigning a high score to scans without a large nodule or vice versa. We believe that a black-box AI-based system that overrules well-established clinical guidelines that prompt radiologists to base management decisions on the size and growth rate of nodules is unlikely to be quickly accepted.

Thus, if and how Google intends to translate this tool, one of several it has developed in recent years in the field of medical image analysis, into a product currently remains to be seen. All of the solutions developed by Google AI so far are proprietary. In the *Nature Medicine* letter, the authors state that their code has "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 […] to allow independent replication"[6]. This description, however, lacks many technical details that might be crucial for obtaining a good level of performance. Therefore, we consider Google's code availability statements as only paying lip service to the principle of reproducible science.

Colin Jacobs and Bram van Ginneken *

Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, Netherlands.

**e-mail:* bram.vanginneken@radboudumc.nl

https://doi.org/10.1038/s41571-019-0248-7

1. National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.* **365**, 395–409 (2011).
2. Pinsky, P. F. Lung cancer screening with low-dose CT: a world-wide view. *Transl Lung Cancer Res.* **7**, 234–242 (2018).
3. 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).
4. Ritchie, A. J. et al. Computer vision tool and technician as first reader of lung cancer screening CT scans. *J. Thorac. Oncol.* **11**, 709–717 (2016).
5. Kaggle Inc. Data science bowl 2017. Can you improve lung cancer detection? *Kaggle* https://www.kaggle.com/c/data-science-bowl-2017/ (2017).
6. Ardila, D. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).
7. American College of Radiology. Lung CT screening reporting & data system. *ACR* https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads (2019).
8. Chung, K. et al. Lung-RADS category 4X: does it improve prediction of malignancy in subsolid nodules? *Radiology* **284**, 264–271 (2017).
9. McWilliams, A. et al. Probability of cancer in pulmonary nodules detected on first screening CT. *N. Engl. J. Med.* **369**, 910–919 (2013).

Competing interests
C.J. and B.v.G. receive funding and royalties from MeVis Medical Solutions for the development of software related to lung cancer screening.