Artificial intelligence for detection and characterization of pulmonary nodules in lung cancer CT screening: ready for practice?

Anton Schreuder[1], **Ernst T. Scholten**[1], **Bram van Ginneken**[1,2], **Colin Jacobs**[1]

[1]Department of Radiology, Nuclear Medicine, and Anatomy, Radboudumc, Nijmegen, The Netherlands; [2]Fraunhofer MEVIS, Bremen, Germany

Contributions: (I) Conception and design: All authors; (II) Administrative support: None; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: None; (V) Data analysis and interpretation: None; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Colin Jacobs. Department of Radiology, Nuclear Medicine, and Anatomy, Radboudumc, Geert Grooteplein 10, 6525 GA Nijmegen, The Netherlands. Email: colin.jacobs@radboudumc.nl.

Abstract: Lung cancer computed tomography (CT) screening trials using low-dose CT have repeatedly demonstrated a reduction in the number of lung cancer deaths in the screening group compared to a control group. With various countries currently considering the implementation of lung cancer screening, recurring discussion points are, among others, the potentially high false positive rates, cost-effectiveness, and the availability of radiologists for scan interpretation. Artificial intelligence (AI) has the potential to increase the efficiency of lung cancer screening. We discuss the performance levels of AI algorithms for various tasks related to the interpretation of lung screening CT scans, how they compare to human experts, and how AI and humans may complement each other. We discuss how AI may be used in the lung cancer CT screening workflow according to the current evidence and describe the additional research that will be required before AI can take a more prominent role in the analysis of lung screening CT scans.

Keywords: Lung cancer; artificial intelligence (AI); computed tomography (CT); pulmonary nodule

Submitted Feb 26, 2020. Accepted for publication Jun 18, 2020.
doi: 10.21037/tlcr-2020-lcs-06
View this article at: http://dx.doi.org/10.21037/tlcr-2020-lcs-06

Lung cancer screening and reporting systems

Lung cancer is the leading cause of cancer-related death worldwide, for which the 5-year survival rates have yet to surpass 20% (1,2). Tobacco smoking remains the main risk factor for lung cancer. Although there is a decreasing prevalence of smokers in most countries, tobacco control is not the only measure for decreasing lung cancer mortality (3,4). In 2011, the National Lung Screening Trial (NLST) was the first multicenter randomized controlled trial (RCT) to demonstrate that three rounds of annual screening of a high-risk population using low-dose chest computed tomography (CT) lead to 20% fewer lung cancer deaths after seven years of follow-up, compared to annual screening with chest radiography (5). Over 53,000 participants were included in this landmark study. The Dutch-Belgian NELSON trial—the second largest RCT with 15,789 participants—recently published their results and showed a 24% mortality reduction in a high-risk population of men compared to no screening (6). Various other smaller RCTs have also reported evidence for the beneficial effects of screening, such as the German Lung cancer Screening Intervention (LUSI) (7) and the Multicentric Italian Lung Detection (MILD) trials (8), but were underpowered.

In the screening workflow, the main task for a radiologist is to search for pulmonary nodules and assess the malignancy risk of these nodules based on characteristics such as size, type, morphology, location, and growth (if prior scans are available). The NLST definition of a positive screen—the presence of at least one solid nodule >4 mm—led to a 24% false positive rate (5). The results of the NELSON trial showed that growth-rate assessment for indeterminate nodules is an effective way to reduce the false positive rate to approximately 2% (6). Taking these findings into consideration, various CT reporting systems were published with the aim of improving the false positive rate while maintaining a high sensitivity.

There is currently one major reporting system for the interpretation of the annual screening CT scans in the United States: Lung-RADS (9). Usage of this classification system is obligatory for screening centers in the United States to receive reimbursement. Other nodule management guidelines designated for screening settings are those from the British Thoracic Society (10), National Comprehensive Cancer Network (11), European Union Position Statement on Lung Cancer Screening (12), and International Early Lung Cancer Action Program (13); these recommendations were summarized by Kauczor *et al.* (14).

Such reporting systems require radiologists to assess the quality of a scan, to search, measure, classify and characterize pulmonary nodules, to look for other significant findings, and finally to determine the malignancy risk of the screenee and decide on the follow-up. The categorization of these scans is laborious, has a substantial reader variability (15), and thus influences the effectiveness of lung cancer screening. In this non-systematic review, we discuss the potential role of artificial intelligence (AI) and whether state-of-the-art algorithms are ready for practice. If screening will be implemented on a large scale, AI may be able to play an important role in reducing costs and improving the efficiency of screening.

Current performance of artificial intelligence algorithms

AI is a broad term that has no clear definition, but typically refers to computer systems that can interpret and learn from data to perform certain tasks and reach certain goals. Deep learning, a methodology where computers can learn high dimensional features from large amounts of data, has led to a revolution in the field of AI because its use resulted in major improvements in the performance of AI systems. Deep learning gained momentum in 2012 when Krizhevsky *et al.* (16) successfully implemented a so-called convolutional neural network (CNN) which beat the best performing algorithm in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), an annual competition where algorithms compete to correctly classify and detect objects and scenes in natural images, by a substantial margin. This methodology is currently used for many applications such as autonomous driving, natural language processing, big data analytics, and medical image interpretation. In medical image analysis, CNNs are the methodology of choice and their performance is reaching or even surpassing human performance on an increasing number of tasks (17).

The first papers describing algorithms for automatically detecting and characterizing pulmonary nodules on CT were published over two decades ago. The number of studies on this topic has increased markedly in the last 10 years. This can mainly be attributed to the rise of deep learning, the organization of several challenges, the availability of public datasets, and the imminent implementation of lung cancer screening.

In the next paragraphs, the current status of algorithms for various subtasks of the interpretation of a lung cancer screening CT scan are discussed.

Scan quality

For humans or AI to be able to diagnostically assess a CT scan, a minimum quality level is required. In the screening setting, it is especially important to keep the radiation dose as low as reasonably possible. Due to the high contrast between air and lung parenchyma, high quality scans were already obtainable using an average effective dose of 1.5 mSv. These low-dose CTs were used for most lung cancer screening trials, including the NLST (5). Since 2009, technological advancements enabled the introduction of iterative reconstruction algorithms to clinical practice. As opposed to filtered back projection, this technique revises each reconstructed image for multiple iterations in order to remove artefacts and improve overall image quality (18). This development also led to the introduction of ultra-low-dose CTs, boasting a radiation dose approaching that of X-rays for scanning the chest (approximately 0.5 mSv on average).

In the last years, deep learning techniques have also been incorporated to optimize both radiation dose and reconstruction time (18). A pilot study found that all nodules >2 mm which were visible in standard low-dose CT scans were also found in the ultra-low-dose images (19). Another study reported that two independent observers had a higher sensitivity on ultra-low-dose CT with iterative reconstruction than low-dose CT with filtered back projection (20).

Nodule detection

The typical visual manifestation of lung cancer on CT is in the form of opacities in the lung parenchyma which are not considered part of the normal anatomy, more commonly referred to as pulmonary nodules. The first step in the workflow towards lung cancer diagnosis is the detection of all pulmonary nodules. It is known that radiologists do not find all nodules and that there is considerable disagreement to what constitutes a pulmonary nodule (21,22). Searching for something specific in an image cluttered with vessels and airways is a difficult task for humans, especially when under time pressure and when the number of nodules present is unknown.

Numerous papers have been published on AI algorithms for detecting lung nodules (23,24). Among scientific publications, it is difficult to compare algorithms' performances because each study may use a different dataset, reference standard, and evaluation metric. This is why challenges are important in the field of AI: challenges are competitions open to the public for developing an algorithm for a specified task. They allow researchers to compare different methodological approaches for a certain task on the same dataset and using the same evaluation metric.

The first web-based framework for comparing nodule detection algorithms from lung cancer screening CT scans was the Automated Nodule Detection 2009 (ANODE09) study (25). All submitted algorithms would be tested on the same 50 anonymized scans and evaluated using the same procedure; the reference values of the 207 nodules were kept secret. This study also proposed a method for combining the output from various AI algorithms to achieve an improved combined performance level. The main limitation of this study was the dataset size and uniformity; all were obtained from one center using the same scanner and protocol.

To account for these limitations, the Lung Nodule Analysis 2016 (LUNA16) challenge was set up using 888 scans with 1,186 nodule annotations from the Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI) database for training and testing (22,26). To ensure robustness, the reference values for each scan were based on annotations from four radiologists. At the time of publication of the challenge, the best algorithm reached a sensitivity of 97.2% at the expense of 1 false positive per scan on average. The LUNA16 challenge was officially closed in January 2018, but the organizers open-sourced the evaluation scripts and all data; this challenge therefore continues to be used as a benchmark for more recent AI algorithms.

Most comparative studies between AI algorithms and humans as individual readers for this specific task were performed over a decade ago. Algorithms showed slightly inferior or equivalent sensitivities compared to radiologists at the expense of a noticeable increase in the false positive rate (27-30).

Nodule classification and measurement

After nodules are found, guidelines stratify them into malignancy risk groups based on two main criteria: size and type (9-12). Automatic measurement or classification of nodules was not included as a task in the ANODE09 or LUNA16 challenge (22,25).

With the aim of automatically classifying clinically relevant nodule types, Ciompi *et al.* (31) developed an AI algorithm for differentiating between six nodule types: solid, part-solid, non-solid, perifissural, calcified, and spiculated. The algorithm was validated in an external dataset which was also assessed by four experienced human readers. The authors found that the performance of the AI algorithm was within the inter-observer variability of the four experienced readers, thus performing equivalently to an independent human expert. It was concluded that the algorithm could be reliably used to automatically categorize pulmonary nodules in lung cancer screening.

Larger nodule size and nodule growth are by far the best CT predictors of malignancy (32-34). Size is traditionally determined by manually measuring the longest and perpendicular diameters in the transverse plane. This is prone to inter- and intra-radiologist variability (35), which can influence the diagnostic workup recommendation (15,36). Volumetric segmentation methods have been around for more than a decade; they have the advantage of being more reproducible and less subject to intra- and interobserver variability, but are not commonly available and were not used in most lung cancer screening trials (37).

In studies where two scans of the same patient are made within the same day, the volume differences have been found to be in the order of ±25% (38,39). Additionally, there is a large variation among different algorithms (37,40). Therefore, for the purpose of reliably measuring growth over time, the same segmentation algorithm and version should be used. The reported variability would likely be reduced if these studies would be repeated using novel AI-based volumetric segmentation algorithms and more modern CT data, but no studies have been reported to date. Nevertheless, its key role in several screening trials has led to lung nodule management guidelines advocating for the use of (semi-)automatic volumetric segmentation (9,10,12,13). Note that the diameter can also be automatically measured from the resulting three-dimensional segmentation. A recent study found that mean diameter derived from a CAD is just as predictive for malignancy as the CAD-derived volume when used in a multivariable logistic regression model (41).

Malignancy prediction

Ultimately, the goal of lung cancer CT screening is to predict whether a participant has lung cancer. In the screening workflow, malignancy risk is estimated after detecting and characterizing nodules on CT. The most renown statistical risk model for estimating nodule malignancy risk is the Brock model (also known as the PanCan model) (34), currently incorporated into the British Thoracic Society nodule management guidelines (10) and recommended in Lung-RADS version 1.1 (9). This model was developed using data from the PanCan screening trial and has shown good performance on other independent screening and clinical datasets (32,42). Though the Brock model incorporates several predictors based on patient demographics, nodule size, type, and morphology, previous studies have shown that radiologists can more accurately assess the malignancy risk of a nodule (43,44). However, when radiologists were asked to characterize the signs of malignancy, no consensus is found (43).

We describe two challenges on the topic of malignancy prediction in chest CT scans. The first is the LUNGx Challenge (45) which provided scans from The University of Chicago containing 37 benign and 36 malignant size-matched nodules for testing algorithms. Without having to search for the nodules, the objective was to classify each nodule as either benign or malignant. Of the 11 participating algorithms, only three achieved an area under the receiver operating characteristic curve (AUC) statistically superior to random guessing (range, 0.50–0.68). In comparison, six participating radiologists obtained AUCs between 0.70 and 0.85, three of which were statistically better than the best performing algorithm.

A subsequent major challenge was the 2017 Kaggle Data Science Bowl which focused on the detection of lung cancer on CT and included a total prize money purse of one million dollars (46). Rather than estimating each nodule's malignancy risk, the primary aim was to develop an AI algorithm which can predict whether a person would get a lung cancer diagnosis within 1 year based on a CT scan. In total, over 2000 teams worked on this challenge. The winning team published a paper about their winning solution (47). The top 10 prize winners were required to make their code publicly available online, such that these algorithms could be used by future efforts to integrate them into screening practice.

Considering the number of participants, it can be assumed that the winning algorithms were among the best available worldwide at that moment. An observer study including 11 radiologists (of which seven were specialized in the chest) found that, on average, human expert readers still performed only slightly superior to the top three algorithms [AUC =0.90 (95% confidence intervals from 1,000 bootstrap iterations: 0.85–0.94) *vs.* 0.86 (0.81–0.91), respectively] (48).

In 2019, Ardila *et al.* (49) published a study claiming a superior performance of their deep learning network compared to six radiologists when assessing lung cancer risk from one CT scan (absolute false positive reduction =11%; absolute false negative reduction =5%). When multiple scans were available, the model performance was on par with that of radiologists. The authors concluded that these algorithms may already be able to work independently on certain tasks. Despite the promising results, the conclusion was argued to have been too strong (50): validation was performed on a subset of the cohort that was used for training (NLST) and a small independent cohort, radiologists' performances were based on Lung-RADS (a nodule management guideline, not a 1-year lung cancer risk model), and the radiologists were not thoracic radiologists. The resulting code was not made publicly available and cannot be independently assessed.

Growth of a nodule on CT is the most important predictor of cancer, and growth cannot be assessed from a single scan. The publication by Ardila *et al.* (49) is a good example where the analysis of multiple scans led to a performance on par with or better than radiologists. Another recent publication that designed a neural network to assess the lung cancer risk of follow-up CT scans showed good performance on an independent dataset (51).

Next to the binary prediction of malignancy, it is important to differentiate between different types of tumor. Subsolid nodules have a higher risk of malignancy than solid nodules, but when malignant tend to present an indolent behavior, showing a slower growth rate and a lower metastatic potential (52,53). The previously mentioned medical imaging challenges did not perform subgroup-analysis to investigate the performance of AI for malignancy risk prediction of solid and subsolid nodules separately.

How can artificial intelligence be used?

Though the claims require more extensive validation, the previous section indicates that state-of-the-art AI algorithms for detecting lung cancer CT detection may have achieved radiologist-level performance. However, these studies only compared individual performances and did not consider a collaboration between man and machine. Certain tasks which are considered more difficult for the radiologist may be easier for the algorithm, and vice versa. For example, it is known that subsolid nodules are more often missed by radiologists since there is less contrast with the lung parenchyma (54,55); alternatively, very irregular nodules may not be recognized by the AI due to their rarity in the training data.

Three paradigms have been described for a human reader receiving assistance from an AI system: second reader, concurrent reader, and first reader. As a second reader, the AI system is only enabled after the radiologist has finished reading the scan. This allows the radiologist to perform an initial unbiased assessment, subsequently going through the AI's findings to check whether nodules were missed or misinterpreted. In the concurrent reading paradigm, the radiologist has immediate access to the results of the AI system and uses this while interpreting the image. Finally, a first reader AI system would only have the radiologist assess the nodules already detected by the AI. The latter strategy restricts the radiologist interpretation to areas of interest and hence enables the shortest reading times, but nodules missed by the AI system will go undetected. For the first reader paradigm, a high AI sensitivity is crucial. Note that commercial systems for nodule detection to date have only been approved for use as a concurrent or second reader.

When an AI system is used as a second reader to the radiologist, the goal is that the nodule detection sensitivity is increased. From before the rise of deep learning, Roos *et al.* (56) confirmed this hypothesis by reporting an algorithm which detected 74% (141/190) of the nodules on CT of which 18% (25/141) were not detected by any of the three independent radiologists; on the other hand, 14% (27/190) of the nodules detected by at least one radiologist were missed by the software. Liang *et al.* (57) sought out lung cancer nodules from the NLST which had been visible in a prior scan but had been missed by the radiologists. They ran four nodule detection systems on the prior scans and found nodule detection rates between 56% and 70%. However, in the subsequent scan which ultimately led to the lung cancer diagnoses, the detection rate ranged from 74% to 82%. Both studies concluded that the algorithms could function as a second reader, but the proof that humans and computers would likely complement each other had originated much earlier.

In 2004, Wormanns *et al.* (29) was the first to publish the pulmonary nodule detection performance on CT of a commercial AI and its added value as a second reader. Individually, the AI had a similar sensitivity to three radiologists (0.55 *vs.* 0.51 to 0.55, respectively). When double reading, the sensitivity of two radiologists was between 0.67 and 0.68 while that between a radiologist and AI was between 0.77 and 0.81. This was at the expense of a 7% greater false positive rate compared to radiologists. Since then, various other publications have mirrored these results (27,28,30,58-60).

A good example of a concurrent reader paradigm is a commercial AI system which creates a second CT image with suppressed vessels and detected lung nodules that can be simultaneously viewed as the original CT scan (61). The study reported a significantly increased sensitivity for actionable lung nodules at a somewhat reduced specificity and a significantly reduced interpretation time.

Besides the standard reading paradigm where one or multiple radiologists sign off all CT scans, other screening workflow strategies have been considered. Ritchie *et al.* (62) tested pulmonary nodule detection performance by a trained technician supported by an AI algorithm. For identifying "abnormal" CT scans with at least one nodule (≥1 mm), the technician plus AI had a sensitivity of 0.98 and a specificity of 0.98. Of the malignant nodules, the technician plus AI found 93% (104/112) compared to 85% (95/112) having been detected by PanCan radiologists without AI. With an average prescreen time of 208 seconds per scan, the authors concluded that technicians supported by an AI was a viable option for triaging scans for radiologists.

The replacement of radiologists with technicians can make screening more cost-effective and feasible in countries where there is a shortage of radiologists. This is similar to the workflow in cervical cancer screening using Papanicolaou (Pap) smears, where normal findings are signed-off by trained technologists and only the abnormal tests are forwarded to cyto-pathologists.

What studies are needed next?

The current state and next steps needed for detection and

Table 1 Summary of the current state and next steps needed for detection and characterization of pulmonary nodules in lung cancer CT screening

Task	Current state	Next steps
Detection, segmentation and classification	• Numerous publications presenting good performance; • Commercial systems are available for clinical use as second or concurrent reader	• Evaluate the performance of AI for pathologically proven cancers in solid nodules instead of suspicious nodules defined by a consensus of radiologists; • Continue evaluation studies with novel deep learning-based AI systems in multi-center studies; • Investigate workflows in which AI + trained technicians can triage screening CT scans to be sent for review by radiologists
Malignancy prediction	• Recent publications show performance better than or on par with radiologists; • Results from Kaggle DSB 2017 demonstrate the potential of AI for malignancy prediction; • No commercial systems available that provide a malignancy risk score	• Evaluate the effect of an AI risk score on the performance of radiologists; initiate multi-center evaluation studies; • Evaluate whether and how an AI risk score can be integrated into nodule follow-up guidelines

characterization of pulmonary nodules in lung cancer CT screening are summarized in *Table 1*. The current literature suggests that state-of-the-art AI systems for lung nodule detection and characterization come close to experienced radiologists' performance levels. Many AI studies describe novel architectures for detecting lung nodules, where the reference standard is set by the consensus of radiologists. However, the ultimate goal is not to find all nodules but to find all lung cancers. Future studies should therefore focus on a reference standard where the detection of cancer is measured, determined by histopathological proof or follow-up imaging for at least 2 years (depending on morphology) to show stability of lesions. Unfortunately, there are no public datasets with a substantial number of malignant nodules on CT. The largest database that is publicly available is the NLST database, but the metadata lacks information which nodules were biopsied. Even with the information about the pathological proof and all screening scans available, it is not always obvious which lesions on CT were found to be malignant.

The AI subtask of attributing a malignancy risk score to a nodule is directly associated with the diagnostic follow-up recommendation. In Lung-RADS (9), the 4X category is a special category for lesions that show additional signs of malignancy. Radiologist can assign this category to upgrade nodules scored 4A or lower to undergo the most urgent follow-up management. Chung *et al.* (43) showed that radiologist were able to pick out malignant nodules from lower Lung-RADS categories and appropriately upgrade them to 4X. If AI systems would be able to recognize certain malignancies typically missed by radiologists, their input may improve radiologists' accuracy in upgrading lesions. However, to the best of our knowledge, no study has yet demonstrated the effect on decision making when an AI's estimated malignancy risk is revealed to a radiologist. More specifically, are the radiologists' decisions affected by this additional information? If yes, when do radiologists choose to deviate from the AI's recommendation, and how often were they right to do so? This is an important area where more research is needed.

Another challenge of lung cancer screening will be to avoid unnecessary interventions. As with every screening program, overdiagnosis is a side-effect that needs to be carefully monitored. Although the 5-year death rate from lung cancer is very high, not all malignancies lead to morbidity or death. In an extended follow-up study of the NLST, the authors reported the same lung cancer incidence in the CT and control groups after a period of 10 years, indicating that there was no overdiagnosis in the NLST study (63). However, other studies warn for overdiagnosis and that the consequences must be considered (64,65). It is difficult to predict which patients would not benefit from treatment. At present, there are no AI algorithms that focus on this by for example predicting the histological subtype of screening-detected pulmonary nodules, the growth rate, or the metastatic potential of pulmonary nodules. Though there are algorithms which were designed to predict the time of death from a scan (66-69), there is a lack of appropriate data to perform studies which attempt to predict the risk that lung cancer progression will be the cause of death.

Triaging screening CT scans using trained technicians aided by AI algorithm is a promising direction to substantially reduce the costs and radiologists' workload of

CT interpretation in lung cancer screening programs, but this reading paradigm requires more validation, preferably in a prospective setting. At present, every screening CT scan in the United States needs to be signed off by an American College of Radiology board certified radiologist. If demonstrated that trained technicians can take over a sizable portion of the responsibilities by triaging a large portion of the normal scans without reducing the quality of care, policy changes are needed.

Another potential strategy to reduce costs would be if the trained technicians would be replaced by fully autonomous AI algorithms that are able to perform triage and optimize the selection of screening CT scans that are sent to the screening radiologists. The first fully autonomous AI algorithm that is able to perform diagnostic assessment without supervision of an expert clinician—an AI system called IDx-DR which analyzes fundus photographs in a primary care setting to detect diabetic retinopathy—has recently been approved by the FDA (70). The 2017 Kaggle DSB challenge showed that fully automatic algorithms—incorporating both nodule detection and malignancy risk estimation—reached a promising performance for one-year lung cancer predictions, but performed slightly inferior to expert radiologists (46). Post-challenge algorithms have reported superior performances, with Google's lung cancer AI claiming superiority to or on par with radiologists (49). However, these fully autonomous AI algorithms need to be extended to be explainable and highlight all areas of interest. Even at a radiologist's performance level, an AI black-box which overrules clinical guidelines established by experts will not be readily accepted (50). In addition, these systems should include additional components which help to guarantee the robustness of the AI output. For example, the IDx-DR system has a component which measures the quality of the scan and returns to the operator if deemed insufficient for AI analysis (67).

Various commercial products are on the market which are cleared for use as second reader or concurrent reader (71). These are ready to be adopted in screening centers to assist the reading of screening CTs. In the coming years, evaluation studies which test these AI algorithms in adequately sized datasets from multiple centers will give more insight into their effect on sensitivity, false positive rate, and interpretation time. Indications for ideal statistical and sample size considerations when testing such algorithms have been described (72).

In medicine, new drugs are allowed to the market after one or multiple prospective multicenter RCTs (phase III studies) have shown benefits in the target population. If we translate this to AI algorithms, prospective multicenter RCTs would be needed to build up the necessary evidence. Though various papers have highlighted the importance of extensive testing before the implementation of AI into practice (73-76), RCTs for AI software are not commonly performed and are not mandatory for regulatory approval. Proving the effectiveness of an AI system is complex because integration of AI systems into health systems depends on many factors which are difficult to be investigated simultaneously (e.g., integration into workflow, extent of information display, training of physicians). In addition, the design of RCTs for AI is complicated by the constant rate of improvement of AI algorithms as they are fed with increasing amounts of training data. Although the RCT is an important tool to prove causality, there is no consensus on its role for guiding the deployment of AI in health care.

Conclusions

Recent studies have shown that AI performance is approaching or already on par with radiologists for various tasks that are needed for the current reporting schemes used in lung screening. In its current state, AI algorithms can be used in a supportive role for radiologists when interpreting lung cancer screening CT scans. Future studies should focus on large-scale validation of novel deep learning-based algorithms and need to address novel reading paradigms. If trained readers aided by AI algorithms can be used for triaging normal scans, this may have a substantial effect on the cost-effectiveness of screening. This effect would be larger if fully autonomous algorithms would be allowed to perform triage by selecting potentially abnormal CT scans to be sent for review by radiologists. However, to guarantee that their use is safe and responsible, the requirements for implementing autonomous algorithms should be more extensive than when there is still a trained reader in the loop.

Acknowledgments

Funding: CJ and BVG receive funding and royalties from MeVis Medical Solutions AG, (Bremen, Germany) for the development of software related to lung cancer screening. BVG reports grants and stock/royalties from Thirona, and grants and royalties from Delft Imaging Systems, outside

the submitted work.

Footnote

Provenance and Peer Review: This article was commissioned by the Guest Editors (Paul Van Schil and Annemiek Snoeckx) for the series "Lung cancer screening" published in *Translational Lung Cancer Research*. The article has undergone external peer review.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/tlcr-2020-lcs-06). The series "Lung cancer screening" was commissioned by the editorial office without any funding or sponsorship. CJ and BVG receive funding and royalties from MeVis Medical Solutions AG, (Bremen, Germany) for the development of software related to lung cancer screening. BVG reports grants and stock/royalties from Thirona, and grants and royalties from Delft Imaging Systems, outside the submitted work. The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. World Health Organization. The top 10 causes of death [Internet]. 2018. Available online: https://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death
2. World Health Organization. Global Health Observatory (GHO) data. 2016. Available online: https://www.who.int/gho/tobacco/use/en/
3. Morabia A. Enigmas of Health and Disease: How Epidemiology Helps Unravel Scientific Mysteries. Columbia University Press; 2014.
4. Ritchie H, Roser M. Smoking. Our World in Data. 2019. Available online: https://ourworldindata.org/smoking
5. 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;365:395-409.
6. 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;382:503-13.
7. Becker N, Motsch E, Trotter A, et al. Lung cancer mortality reduction by LDCT screening-Results from the randomized German LUSI trial. Int J Cancer 2020;146:1503-13.
8. Pastorino U, Silva M, Sestini S, et al. Prolonged lung cancer screening reduced 10-year mortality in the MILD trial: new confirmation of lung cancer screening efficacy. Ann Oncol 2019;30:1162-9.
9. American College of Radiology. Lung CT Screening Reporting & Data System v1.1. 2019. Available online: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads
10. Callister ME, Baldwin DR, Akram AR, et al. British Thoracic Society guidelines for the investigation and management of pulmonary nodules. Thorax 2015;70 Suppl 2:ii1-54.
11. Wood DE, Kazerooni EA, Baum SL, et al. Lung Cancer Screening, Version 3.2018, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw 2018;16:412-41.
12. Oudkerk M, Devaraj A, Vliegenthart R, et al. European position statement on lung cancer screening. Lancet Oncol 2017;18:e754-66.
13. International Early Lung Cancer Action Program: Screening Protocol. Available online: www.ielcap.org/protocols. Accessed February 2020.
14. Kauczor HU, Baird AM, Blum TG, et al. ESR/ERS statement paper on lung cancer screening. Eur Respir J 2020;55:1900506.
15. 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;29:924-31.
16. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60:84-90.
17. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal

2017;42:60-88.
18. Willemink MJ, Noël PB. The evolution of image reconstruction for CT-from filtered back projection to artificial intelligence. Eur Radiol 2019;29:2185-95.
19. Paks M, Leong P, Einsiedel P, et al. Ultralow dose CT for follow-up of solid pulmonary nodules: A pilot single-center study using Bland-Altman analysis. Medicine (Baltimore) 2018;97:e12019.
20. Sui X, Meinel FG, Song W, et al. Detection and size measurements of pulmonary nodules in ultra-low-dose CT with iterative reconstruction compared to low dose CT. Eur J Radiol 2016;85:564-70.
21. Pinsky PF, Gierada DS, Nath PH, et al. National lung screening trial: variability in nodule detection rates in chest CT studies. Radiology 2013;268:865-73.
22. 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.
23. Liu B, Chi W, Li X, et al. Evolving the pulmonary nodules diagnosis from classical approaches to deep learning-aided decision support: three decades' development course and future prospect. J Cancer Res Clin Oncol 2020;146:153-85.
24. Li D, Mikela Vilmun B, Frederik Carlsen J, et al. The Performance of Deep Learning Algorithms on Automatic Pulmonary Nodule Detection and Classification Tested on Different Datasets That Are Not Derived from LIDC-IDRI: A Systematic Review. Diagnostics (Basel) 2019;9:207.
25. van Ginneken B, Armato SG 3rd, de Hoop B, et al. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: The ANODE09 study. Med Image Anal 2010;14:707-22.
26. 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;38:915-31.
27. Christe A, Leidolt L, Huber A, et al. Lung cancer screening with CT: evaluation of radiologists and different computer assisted detection software (CAD) as first and second readers for lung nodule detection at different dose levels. Eur J Radiol 2013;82:e873-8.
28. Rubin GD, Lyo JK, Paik DS, et al. Pulmonary nodules on multi-detector row CT scans: performance comparison of radiologists and computer-aided detection. Radiology 2005;234:274-83.
29. Wormanns D, Beyer F, Diederich S, et al. Diagnostic performance of a commercially available computer-aided diagnosis system for automatic detection of pulmonary nodules: comparison with single and double reading. Rofo 2004;176:953-8.
30. Brown MS, Goldin JG, Rogers S, et al. Computer-aided lung nodule detection in CT: results of large-scale observer test. Acad Radiol 2005;12:681-6.
31. Ciompi F, Chung K, van Riel SJ, et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. Sci Rep 2017;7:46479.
32. van Riel SJ, Ciompi F, Jacobs C, et al. Malignancy risk estimation of screen-detected nodules at baseline CT: comparison of the PanCan model, Lung-RADS and NCCN guidelines. Eur Radiol 2017;27:4019-29.
33. Winkler Wille MM, van Riel SJ, Saghir Z, 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 2015;25:3093-9.
34. McWilliams A, Tammemagi MC, Mayo JR, et al. Probability of cancer in pulmonary nodules detected on first screening CT. N Engl J Med 2013;369:910-9.
35. Revel MP, Bissery A, Bienvenu M, et al. Are two-dimensional CT measurements of small noncalcified pulmonary nodules reliable? Radiology 2004;231:453-8.
36. Heuvelmans MA, Walter JE, Vliegenthart R, et al. Disagreement of diameter and volume measurements for pulmonary nodule size estimation in CT lung cancer screening. Thorax 2018;73:779-81.
37. Devaraj A, van Ginneken B, Nair A, et al. Use of Volumetry for Lung Nodule Management: Theory and Practice. Radiology 2017;284:630-44.
38. Gietema HA, Schaefer-Prokop CM, Mali WP, et al. Pulmonary nodules: Interscan variability of semiautomated volume measurements with multisection CT-- influence of inspiration level, nodule size, and segmentation performance. Radiology 2007;245:888-94.
39. Wormanns D, Kohl G, Klotz E, et al. Volumetric measurements of pulmonary nodules at multi-row detector CT: in vivo reproducibility. Eur Radiol 2004;14:86-92.
40. Han D, Heuvelmans MA, Oudkerk M. Volume versus diameter assessment of small pulmonary nodules in CT lung cancer screening. Transl Lung Cancer Res 2017;6:52-61.
41. Tammemagi M, Ritchie AJ, Atkar-Khattra S, et al. Predicting Malignancy Risk of Screen-Detected Lung Nodules-Mean Diameter or Volume. J Thorac Oncol

2019;14:203-11.

42. Chung K, Mets OM, Gerke PK, et al. Brock malignancy risk calculator for pulmonary nodules: validation outside a lung cancer screening population. Thorax 2018;73:857-63.
43. Chung K, Jacobs C, Scholten ET, et al. Lung-RADS Category 4X: Does It Improve Prediction of Malignancy in Subsolid Nodules? Radiology 2017;284:264-71.
44. van Riel SJ, Ciompi F, Winkler Wille MM, et al. Malignancy risk estimation of pulmonary nodules in screening CTs: Comparison between a computer model and human observers. PLoS One 2017;12:e0185032.
45. Armato SG 3rd, Drukker K, Li F, et al. LUNGx Challenge for computerized lung nodule classification. J Med Imaging (Bellingham) 2016;3:044506.
46. Liao F, Liang M, Li Z, et al. Evaluate the Malignancy of Pulmonary Nodules Using the 3-D Deep Leaky Noisy-OR Network. IEEE Trans Neural Netw Learn Syst 2019;30:3484-95.
47. Kaggle Inc. Data science bowl 2017. Can you improve lung cancer detection? Available online: https://www.kaggle.com/c/data- science-bowl-2017/. Accessed February 2020.
48. Jacobs C, Scholten E, Schreuder A, et al. An observer study comparing radiologists with the prize-winning lung cancer detection algorithms from the 2017 Kaggle Data Science Bowl. Annual Meeting of the Radiological Society of North America, 2019.
49. 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;25:954-61.
50. Jacobs C, van Ginneken B. Google's lung cancer AI: a promising tool that needs further validation. Nat Rev Clin Oncol 2019;16:532-3.
51. 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:e353-62.
52. Henschke CI, Yip R, Smith JP, et al. CT Screening for Lung Cancer: Part-Solid Nodules in Baseline and Annual Repeat Rounds. AJR Am J Roentgenol 2016;207:1176-84.
53. Obayashi K, Shimizu K, Nakazawa S, et al. The impact of histology and ground-glass opacity component on volume doubling time in primary lung cancer. J Thorac Dis 2018;10:5428-34.
54. Jacobs C, van Rikxoort EM, Murphy K, et al. Computer-aided detection of pulmonary nodules: a comparative study using the public LIDC/IDRI database. Eur Radiol 2016;26:2139-47.
55. Silva M, Schaefer-Prokop CM, Jacobs C, et al. Detection of Subsolid Nodules in Lung Cancer Screening: Complementary Sensitivity of Visual Reading and Computer-Aided Diagnosis. Invest Radiol 2018;53:441-9.
56. Roos JE, Paik D, Olsen D, et al. Computer-aided detection (CAD) of lung nodules in CT scans: radiologist performance and reading time with incremental CAD assistance. Eur Radiol 2010;20:549-57.
57. Liang M, Tang W, Xu DM, et al. Low-Dose CT Screening for Lung Cancer: Computer-aided Detection of Missed Lung Cancers. Radiology 2016;281:279-88.
58. Teague SD, Trilikis G, Dharaiya E. Lung nodule computer-aided detection as a second reader: influence on radiology residents. J Comput Assist Tomogr 2010;34:35-9.
59. White CS, Pugatch R, Koonce T, et al. Lung nodule CAD software as a second reader: a multicenter study. Acad Radiol 2008;15:326-33.
60. Zhao Y, de Bock GH, Vliegenthart R, et al. Performance of computer-aided detection of pulmonary nodules in low-dose CT: comparison with double reading by nodule volume. Eur Radiol 2012;22:2076-84.
61. Lo SB, Freedman MT, Gillis LB, et al. JOURNAL CLUB: Computer-Aided Detection of Lung Nodules on CT With a Computerized Pulmonary Vessel Suppressed Function. AJR Am J Roentgenol 2018;210:480-8.
62. Ritchie AJ, Sanghera C, Jacobs C, et al. Computer Vision Tool and Technician as First Reader of Lung Cancer Screening CT Scans. J Thorac Oncol 2016;11:709-17.
63. National Lung Screening Trial Research Team. Lung Cancer Incidence and Mortality with Extended Follow-up in the National Lung Screening Trial. J Thorac Oncol 2019;14:1732-42.
64. Reich JM, Kim JS. Current Controversies in Cardiothoracic Imaging: Low-dose Computerized Tomographic Overdiagnosis of Lung Cancer is Substantial; Its Consequences are Underappreciated-Point. J Thorac Imaging 2019;34:154-6.
65. Hutchinson BD, Moreira AL, Ko JP. Spectrum of Subsolid Pulmonary Nodules and Overdiagnosis. Semin Roentgenol 2017;52:143-55.
66. Hosny A, Parmar C, Coroller TP, et al. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. PLoS Med 2018;15:e1002711.
67. Guo H, Kruger U, Wang G, et al. Knowledge-Based Analysis for Mortality Prediction From CT Images. IEEE J Biomed Health Inform 2020;24:457-64.
68. Oakden-Rayner L, Carneiro G, et al. Precision Radiology:

Predicting longevity using feature engineering and deep learning methods in a radiomics framework. Sci Rep 2017;7:1648.

69. Carneiro G, Oakden-Rayner L, Bradley AP, et al. Automated 5-year mortality prediction using deep learning and radiomics features from chest computed tomography. 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). Melbourne, VIC, 2017:130-4.
70. Abràmoff MD, Lavin PT, Birch M, et al. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 2018;1:39.
71. Obuchowski NA, Bullen JA. Statistical considerations for testing an AI algorithm used for prescreening lung CT images. Contemp Clin Trials Commun 2019;16:100434.
72. Challen R, Denny J, Pitt M, et al. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019;28:231-7.
73. AI for Radiology. An implementation guide. Available online: www.aiforradiology.com. Accessed February 2020.
74. Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. JAMA 2019;322:2377-8.
75. Angus DC. Randomized Clinical Trials of Artificial Intelligence. JAMA 2020;323:1043-5.
76. Maddox TM, Rumsfeld JS, Payne PRO. Questions for Artificial Intelligence in Health Care. JAMA 2019;321:31-2.

Cite this article as: Schreuder A, Scholten ET, van Ginneken B, Jacobs C. Artificial intelligence for detection and characterization of pulmonary nodules in lung cancer CT screening: ready for practice? Transl Lung Cancer Res 2021;10(5):2378-2388. doi: 10.21037/tlcr-2020-lcs-06