Outstanding negative prediction performance of solid pulmonary nodule volume AI for ultra-LDCT baseline lung cancer screening risk stratification

Lancaster, Harriet L; Zheng, Sunyi; Aleshina, Olga O; Yu, Donghoon; Yu Chernina, Valeria; Heuvelmans, Marjolein A; de Bock, Geertruida H; Dorrius, Monique D; Willem Gratama, Jan; Morozov, Sergey P

Published in:
Lung Cancer

DOI:
10.1016/j.lungcan.2022.01.002

Publication date:
2022

Lung Cancer 165 (2022) 133–140

Contents lists available at ScienceDirect

Lung Cancer

journal homepage: www.elsevier.com/locate/lungcan

Outstanding negative prediction performance of solid pulmonary nodule volume AI for ultra-LDCT baseline lung cancer screening risk stratification

Harriet L. Lancaster [a,i,1], Sunyi Zheng [b,i,1], Olga O. Aleshina [c], Donghoon Yu [d], Valeria Yu. Chernina [c], Marjolein A. Heuvelmans [a,i], Geertruida H. de Bock [a], Monique D. Dorrius [a,e], Jan Willem Gratama [f], Sergey P. Morozov [c], Victor A. Gombolevskiy [c,g], Mario Silva [h], Jaeyoun Yi [d], Matthijs Oudkerk [i,j,*]

[a] Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[b] Department of Radiotherapy, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[c] State Budget-Funded Health Care Institution of the City of Moscow «Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department, Moscow, Russian Federation
[d] Coreline Soft, Seoul, Republic of Korea
[e] Department of Radiology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[f] Department of Radiology, Gelre Hospital, Apeldoorn, the Netherlands
[g] AIRI, Moscow, Russian Federation
[h] Scienze Radiologiche, Department of Medicine and Surgery (DiMeC), University of Parma, Parma, Italy
[i] Institute for Diagnostic Accuracy, Groningen, Netherlands
[j] Faculty of Medical Sciences, University of Groningen, Groningen, Netherlands

ARTICLE INFO

Keywords:
Artificial intelligence
Computer-aided detection
Lung cancer
Ultra LDCT
Screening
Pulmonary nodule

ABSTRACT

Objective: To evaluate performance of AI as a standalone reader in ultra-low-dose CT lung cancer baseline screening, and compare it to that of experienced radiologists.
Methods: 283 participants who underwent a baseline ultra-LDCT scan in Moscow Lung Cancer Screening, between February 2017–2018, and had at least one solid lung nodule, were included. Volumetric nodule measurements were performed by five experienced blinded radiologists, and independently assessed using an AI lung cancer screening prototype (AVIEW LCS, v1.0.34, Coreline Soft, Co. ltd, Seoul, Korea) to automatically detect, measure, and classify solid nodules. Discrepancies were stratified into two groups: positive-misclassification (PM); nodule classified by the reader as a NELSON-plus /EUPS-indeterminate/positive nodule, which at the reference consensus read was < 100 mm^3, and negative-misclassification (NM); nodule classified as a NELSON-plus /EUPS-negative nodule, which at consensus read was ≥ 100 mm^3.
Results: 1149 nodules with a solid-component were detected, of which 878 were classified as solid nodules. For the largest solid nodule per participant (n = 283); 61 [21.6 %; 53 PM, 8 NM] discrepancies were reported for AI as a standalone reader, compared to 43 [15.1 %; 22 PM, 21 NM], 36 [12.7 %; 25 PM, 11 NM], 29 [10.2 %; 25 PM, 4 NM], 28 [9.9 %; 6 PM, 22 NM], and 50 [17.7 %; 15 PM, 35 NM] discrepancies for readers 1, 2, 3, 4, and 5 respectively.
Conclusion: Our results suggest that through the use of AI as an impartial reader in baseline lung cancer screening, negative-misclassification results could exceed that of four out of five experienced radiologists, and radiologists' workload could be drastically diminished by up to 86.7%.

1. Introduction

Lung cancer is a global problem as it remains the most common cause of cancer deaths. In 2020, 1.8 million deaths were attributable to lung cancer [1]. Extensive research has now unequivocally shown that lung cancer mortality can be significantly reduced through early detection of

* Corresponding author.
E-mail address: m.oudkerk@rug.nl (M. Oudkerk).
[1] Authors contributed equally

https://doi.org/10.1016/j.lungcan.2022.01.002
Received 28 June 2021; Received in revised form 4 October 2021; Accepted 3 January 2022
Available online 6 January 2022

preclinical disease represented by lung nodules, using low-dose CT lung cancer screening in well-defined high risk populations [2,3]. Therefore, the focus has now shifted to implementation research [4].

The implementation of LDCT lung cancer screening is about to represent a massive increase in workload in radiologist practice. The capacity of clinical radiologists is already under extreme pressure due to an ever-increasing demand for radiology services and workforce shortages [5]. Thus, radiologists' challenging workloads associated with LDCT lung cancer screening is a noteworthy hurdle which needs to be addressed and overcome in the planning of lung cancer screening implementation.

Artificial intelligence (AI) could offer a solution. Over the last decade significant progress has been made to improve AI algorithms for use in LDCT lung cancer screening. Computer aided detection/diagnosis (CAD) systems that integrate machine learning and image processing have been developed to predominantly act as a 'second reader' for the radiologist, with the aim of improving reader accuracy [6]. However, despite multiple studies presenting a possible value to using CAD as a second reader [7,8], too many doubts remain over accuracy and reproducibility for it to be clinically accepted [9]. The greatest obstacle is the false-positive results (lung nodules classified as clinically significant, which when worked-up are benign), and false-negative results (clinically significant lung nodules which are not reported). False-positive results lead to an increased workload for clinicians and potentially unnecessary psychological stress for the patient as well as morbidity. False-negative results, on the other hand, represent clinically significant findings which could go undiagnosed. CAD systems are being constantly developed and progress has been substantial. Improvements are continuously being shown in both sensitivity and specificity. However, for AI to be of maximum benefit in lung cancer screening, the system used should be robust enough to safely rule out benign lung nodules.

Additionally, as CT-lung cancer screening involves radiation exposure, questions have also been raised over associated hazardous effects. Consequently, the use of ultra-LDCT screening is now being deliberated [10]. To the best of our knowledge, the efficacy of AI as a standalone reader in ultra-LDCT lung cancer screening has never been validated.

The aim of this study was to evaluate the performance of an AI prototype as an impartial reader in ultra-LDCT lung cancer baseline screening, and compare it to that of experienced radiologists and a consensus read reference standard, when using volumetric measurements with the 100 mm^3 NELSON-plus/EUPS protocol threshold [4,11].

2. Materials and methods

2.1. Study design and population

A dataset of CT-scans from 283 participants who underwent a baseline ultra-LDCT thorax scan and had at least one solid nodule of any size at their baseline scan, between February 2017 and 2018, as part of Moscow Lung Cancer Screening (MLCS) was used in this present study. MLCS participants' CT scans were selected for this study if they met the following inclusion criteria; 50–80 years of age; $\geq$ 30 packyears smoking history; current smoker or former smoker (ceased smoking $<$ 15 years previously); and did not develop lung cancer within two years of their baseline ultra-LDCT scan. MCLS participants were excluded based on the following criteria: history of lung cancer or lung surgery (not including lung biopsy); cancer diagnosis within $<$ 5 years of baseline screening; life expectancy of $<$ 5 years due to severe cardiovascular, immunological, respiratory, or endocrine illness; acute respiratory disease; antibiotic treatment $<$ 12 weeks prior to screening; hemoptysis; weight loss $>$ 10 kg within the year prior to screening; or no lung nodules detected during baseline screening.

MCLS was conducted within the framework of order no. 49 dated 01.02.2017 in the Moscow Department of Health. All participants signed an informed consent document and approval was granted by the Independent Local Ethics Committee of the Office of the President of the Russian Federation, Federal State Budgetary Institution "Central Clinical Hospital with Polyclinic" (Moscow) dated 20.05.2017.

2.2. Ultra-low dose thorax CT (ultra-LDCT) scan protocol

MCLS ultra-LDCT scan protocol has been published previously [12]. In short, MCLS participants underwent a baseline ultra-LDCT scan using Toshiba Aquilion 64 (Canon Medical Systems, Japan) slice CT scanners with a tube voltage of 135 kV, current from 15 mA to 25 mA, time rotation of 0.50 *sec*, pitch 1.484, slice thickness of 1 mm, and slice increment 1 mm. Radiation dose did not exceed one mSv, which is recognized by ERS as ultra-LDCT [13]. Participants were scanned from lung apices to bases during a single breath hold scan.

2.3. Data management

The dataset used in this study was retrospectively collected from the MLCS data management systems. All participant data is stored centrally in the Unified Medical Information and Analysis System (EMIAS) of Moscow and the Unified Radiological Information Service (URIS).

2.4. AI deep learning-based nodule measurement

Pulmonary nodules were identified, segmented and classified by an automated AI lung cancer screening prototype (AVIEW LCS, v1.0.34, Coreline Soft, Co. ltd, Seoul, Korea). The prototype was trained on 888 CT scans from the public LUNA16 dataset using Densenet and Resnet architectures for nodule detection. After nodules were detected, nodule segmentation was performed. A threshold of −450 HU and −200 HU was used to coarsely extract solid component regions for solid and part-solid nodules, respectively. To select ground-glass regions for sub-solid nodules a histogram-based threshold was applied [14]. Subsequently, an asymmetric deformable model that utilized a modified energy function, and intensity constrained averaging function was designed to refine the segmentation of nodule regions [15]. Following the segmentation of nodules, radiomics features, such as 3D_Texture_firstOrder_variance, 3D_Texture_GLCM_CP, were extracted and a random forest model was trained with ten most important features to classify nodules into solid, part-solid, and ground-glass types.

2.5. Volumetric lung nodule measurement

All 283 participants' ultra-LDCT scans were analyzed independently by five thoracic radiologists with more than seven years of experience in reading CT scans in lung cancer screening programs. In the case of multiple nodules per participant, the largest solid nodule was selected, as is recommended in the current CT-lung cancer screening guidelines [11,16]. Location of the largest nodules were correlated prior to analysis.

Readers 1, 2, and 3, read the ultra-LDCT for visual detection of nodules and used CLS- semi-automated 3D to segment the nodule volume (AVIEW LCS, v1.0.34, Coreline Soft, Co. ltd, Seoul, Korea). Reader 4 used AGFA-semi-automated volume measurement software (AGFA Enterprise 8.0 Imaging software - Agfa HealthCare, Belgium), and reader 5 used Syngo.via MM Oncology VB20 semi-automated volume measurement software. Our AI lung cancer screening prototype then independently analyzed all ultra-LDCT scans to automatically detect, measure, and classify nodules. Categorization of lung nodules was based on the NELSON-plus/EUPS protocol [4]. Nodules $<$ 100 mm^3 were classified as negative, and $\geq$ 100 mm^3 as indeterminate/positive, as 100 mm^3 is the upper volume threshold of benign nodule growth. At this threshold lung cancer risk probability increases in comparison to patients without nodules as reflected in the inclusion criteria of a lung cancer screening program [17].

An independent consensus read was performed by a panel of three radiologists with $>$ 10 years' experience and an experienced IT

Table 1
Distribution of results per reader.

	Reader 1 [CLS]	Reader 2 [CLS]	Reader 3 [CLS]	Reader 4 [AGFA]	Reader 5 [Syngo.via]	AI [CLS]
Positive-misclassification	22 (7.8)	25 (8.8)	25 (8.8)	6 (2.1)	15 (5.3)	53 (18.7)
Negative-misclassification	21 (7.4)	11 (3.9)	4 (1.4)	22 (7.8)	35 (12.4)	8 (2.8)
Total Discrepancies	*43 (15.1)*	*36 (12.7)*	*29 (10.2)*	*28 (9.9)*	*50 (17.7)*	*61 (21.6)*
NPV (95 % CI)	0.89 (0.85–0.94)	0.94 (0.91–0.97)	0.98 (0.96–1.00)	0.90 (0.86–0.94)	0.84 (0.79–0.89)	0.95 (0.91–0.98)

[semi-automated volume measurement software package]; (percentage % of nodules n = 283); NPV negative predictive value; CI confidence interval.

technologist, of all the 283 largest nodules. Two of the consensus panel were not involved in the first individual read. This consensus read served as the reference standard and was used to determine the number of positive-misclassification (PM) and negative-misclassification (NM) results. PM's were nodules classified as ≥ 100 mm^3 by readers/AI, which at consensus read measured < 100 mm^3. NM's were nodules classified as < 100 mm^3, which at consensus read measured ≥ 100 mm^3. Correct-positive (CP) and correct-negative (CN) results were those in agreement with consensus read.

As various segmentation software packages were used (CLS, AGFA, and Synco.via), we compared the coherence of the radiologists interpretation, by looking at the variation in discrepant results.

2.6. Statistical analysis

A case by case analysis was performed to determine the volume of the largest nodule per participant. The results of each reader and AI were compared to the consensus read to determine the number of PM and NM results, and the total number of discrepancies. We report these results as absolute frequencies and percentages, and include a negative predictive value (NPV) with 95 % confidence interval (CI) where $\alpha = 0.05$. An upper and lower limit for workload reduction when using AI was calculated based on nodule presence in a general population. During the NELSON-trial, in approximately 50 % of the population no nodules were detected [4]. The prevalence of participants with no-nodules/nodules $<$ 30 mm^3 reported in a general Dutch population who underwent a LDCT-scan was 62 %, although this was not a lung cancer screening trial [18]. In a sub-study of the NLST-trial and in the Korean Lung Cancer Screening project (K-LUCAS), 40 % and 54 % of participants respectively were in the Lung-RADS 1 category; no nodules or definitely benign [19,20]. As our dataset contained only participant scans (n = 283) where nodules were detected, we extrapolated the number of participants to the average nodule distribution in the general lung cancer screening population (50 %) to a total number of 2×283 (n = 566). As an average of 50 % have no nodules, there can be no NM results in this group, and PM results will not exceed the false positive-rate of the population with nodules. Hence, the percentage of PM results would be lower in a general lung cancer screening population. Therefore, the upper limit was calculated based on there being no PM results, and the lower limit based on the PM findings reported when using AI in the nodule group, which is probably a significant overestimation. Workload reduction limits were calculated as follows; upper limit = (n + CN + NM)/566 and lower limit = ((n + CN + NM) – PM)/566, where n is the number of participants with a nodule (n = 283), CN is the number of negative nodules reported at consensus read, and NM and PM are the number of negative-misclassification and positive-misclassification findings reported by AI, discrepant with consensus read.

3. Results

3.1. Population characteristics

A total of 283 participants with one or more lung nodule(s) were included. Participants were 50–80 years of age (mean ± SD; 64.6 ± 5.3), and 161 (56.9 %) were male.

3.2. Volumetric lung nodule measurements

For the included participants, a total of 1149 lung nodules with a solid component were detected, of which 878 were classified as pure solid nodules. The largest solid nodule per participant was used: the consensus read reported 83/283 (29 %) nodules ≥ 100 mm^3 (NELSON-plus indeterminate/positive nodules), and 200/283 (71 %) nodules $<$ 100 mm^3 (NELSON-plus negative nodules).

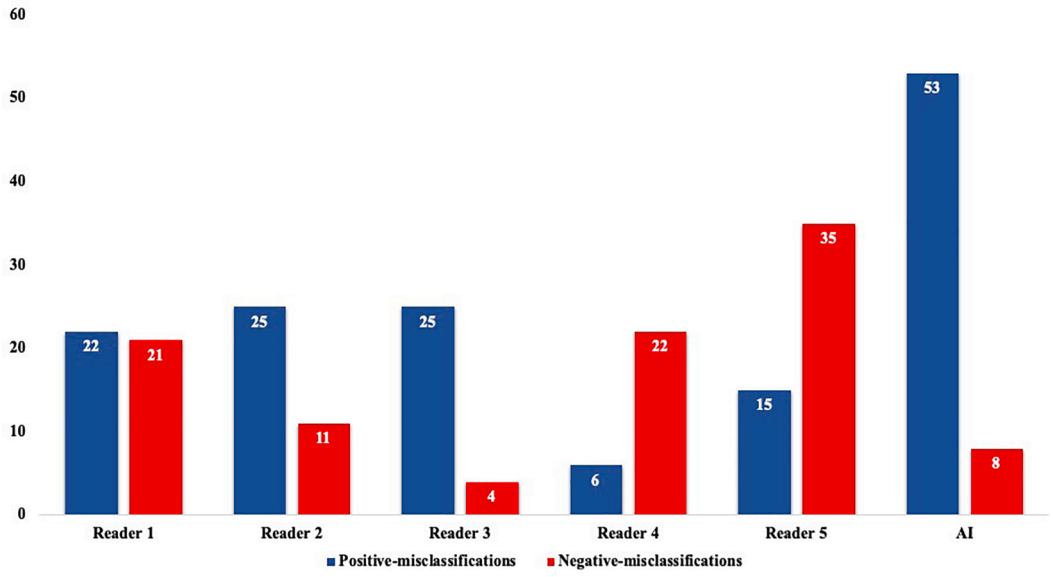

Fig. 1. Graphical representation of the number of positive-misclassification and negative-misclassification findings per reader.

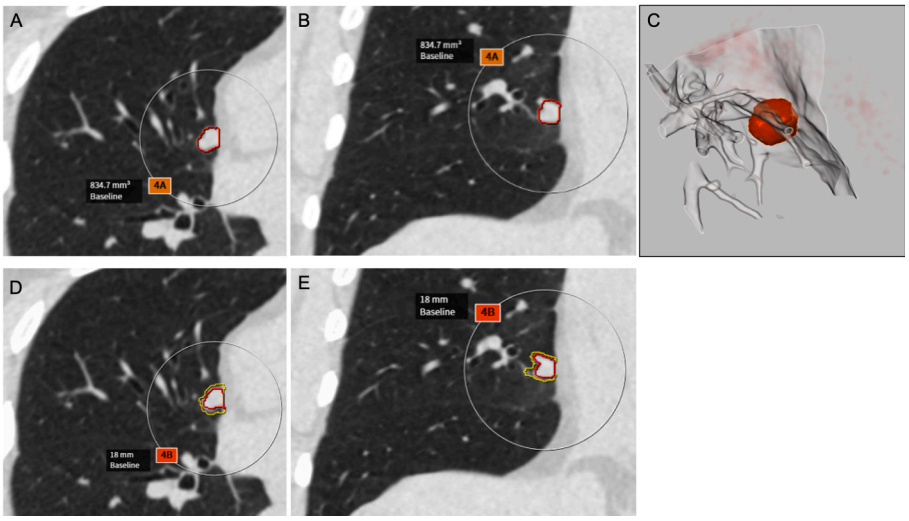

Fig. 2. Automatic classification of nodule density: correct volume measurement of part-solid component by AI. Axial (A) and coronal (B) CT reformation showing a large solid nodule (noted by red outline) attached to pleura and abutting the aspect of mediastinal pleura: the nodule was reported by 4 out of 5 radiologists and classified as solid, one radiologist missed the nodule and AI classified the nodule as part-solid. Multiplanar reconstructions capture minimal ground-glass opacity surrounding the paracardiac space (potentially motion artifact from cardiac cycle), which was allegedly conditioned the AI classification into part-solid nodule (D, E). The volume rendering reconstruction (C) captures the complexity of geometrical layout between the large solid nodule (noted by red surface) and the solid vessel structure that is partially notched into the solid nodule. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

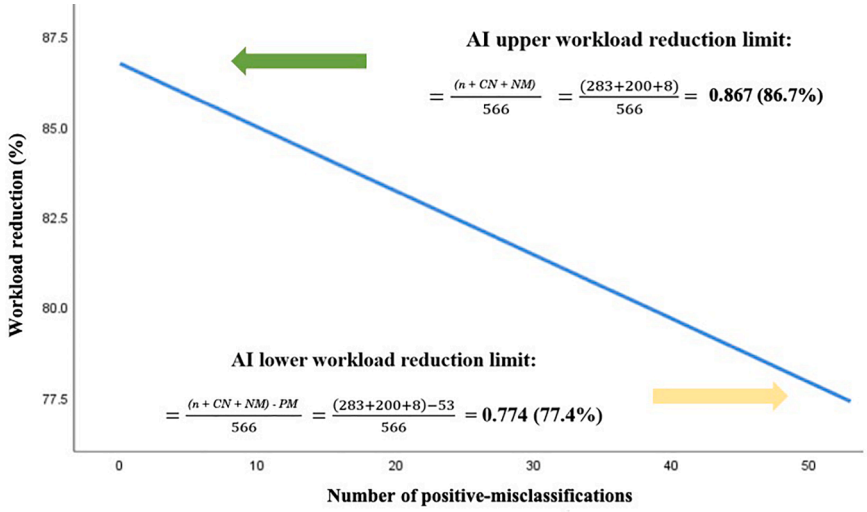

Fig. 3. AI upper and lower workload reduction limit calculations. n is the number of participants with a nodule (n = 283), CN is the number of negative nodules reported at consensus read, NM and PM are the number of negative-misclassification and positive-misclassification findings reported by AI discrepant with consensus read, and 566 is the extrapolated number of participants to the average nodule distribution in the general lung cancer screening population, a total number of 2 × 283.

AI as an impartial reader had 61 [21.6 %; 53 PM, 8 NM] discrepancies reported, compared to 43 [15.1 %; 22 PM, 21 NM], 36 [12.7 %; 25 PM, 11 NM], 29 [10.2 %; 25 PM, 4 NM], 28 [9.9 %; 6 PM, 22 NM], and 50 [17.7 %; 15 PM, 35 NM] discrepancies for readers 1, 2, 3, 4, and 5 respectively. An overview of results per reader can be seen in Table 1 and graphically in Fig. 1.

The eight negative-misclassification findings reported by AI were analyzed further. Four of the eight NM findings were recognized by AI but classified as being part-solid, although the AI correctly measured the solid component (≥100 mm^3), see Fig. 2. Twelve nodules detected by AI were between 90 and 100 mm^3 and eleven were between 100 and 110 mm^3. Of these 23 nodules, 7 were PMs and there were no NM findings.

When looking specifically at the performance of the manual software packages (CLS, AGFA and Syngo.via), we see no notable variation in performance of volume segmentation. We do however see variation in the individual performance of the radiologists, the inter-reader performance.

When using AI in a general lung cancer screening population, based on the findings in this study, we could expect a workload reduction lower limit of 77.4 % and upper limit of 86.7 %. An overview of the AI workload reduction calculations can be found in Fig. 3.

4. Discussion

Our aim was to shed light on the performance of AI as an impartial reader in ultra-low-dose CT lung cancer baseline screening, and compare it to that of experienced radiologists.

We demonstrated that AI as a standalone reader outperforms all except one experienced radiologist, when looking specifically at negative misclassifications. AI had 8 (2.8 %) NM results compared to 21 (7.4 %), 11 (3.9 %), 4 (1.4 %), 22 (7.8 %), and 35 (12.4 %) for readers 1 to 5 respectively. Also, only 53 (18.7 %) positive-misclassification results were reported by AI for all lung nodules. False-negative results, in our study represented by negative-misclassifications, are particularly undesirable in screening programs as they can lead to a potential delay in the detection of cancer, and public confidence in screening could be reduced [21].

Previous research has demonstrated the value of AI as a 'second reader'. Christe et al., investigated the best pairing of first and second reader, human and CAD, when using an anthropomorphic lung phantom and artificial lung nodules. They found the highest sensitivity (between 97 % and 99 %) of lung nodule detection when combining a human reader with CAD, independent of the CT-examination dose. When comparing any two CAD systems, lower sensitivity was found (between 85 % and 88 %), which was significantly less than the combination of

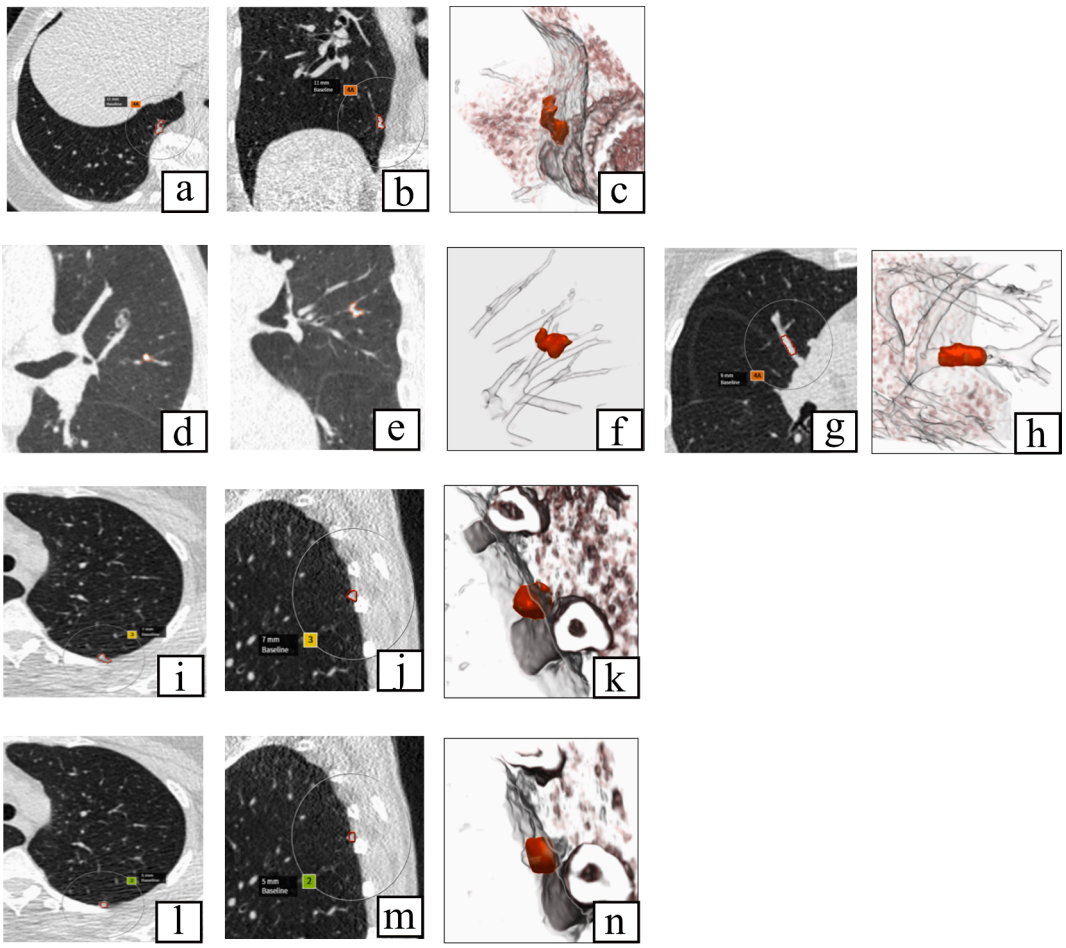

Fig. 4. Positive-misclassification findings of AI due to segmentation of right paravertebral scar, and over-segmentation of solid nodules attached to vessels or pleura. Axial (A) and coronal (B) CT reformation showing a right paravertebral focal opacity with elongated shape (noted by red outline) into the expected pattern of paravertebral scarring, the automatic detection and characterization classified this finding as solid nodule with volume segmentation in AI reading reported volume above 100 mm3 (376 mm^3). The volume rendering reconstruction (C) captures the segmentation of the elongated volumetric structure (noted by red surface) and its proximity to the paraspinal pleura. Axial (D) and coronal (E) CT reformation showing a solid nodule attached to vessel, the automatic volume segmentation in AI reading reported volume above 100 mm3 (149 mm^3), which however resulted from inclusion of vessel structure (noted by red outline). The volume rendering reconstruction (F) captures the exaggerated segmentation of nodule volume and vessel structure (noted by red surface). A second axial (G) CT reformation showing a vessel (vein) interpreted as solid nodule by AI and classified as $>$ 100 mm^3 (233 mm^3, noted by red outline). The volume rendering reconstruction (H) captures the segmentation of the cylindric vascular structure (noted by red surface). Axial (I) and sagittal (J) CT reformation showing a solid subpleural nodule (noted by red outline), the automatic segmentation in AI reading reported volume above 100 mm3 (137 mm^3). The volume rendering reconstruction (K) captures the segmentation of the solid nodule with exaggerated segmentation across the pleural surface into the extrapleural fat space (noted by red surface). Reader 3 measurement in axial (L) and sagittal (M) showing the edited segmentation (semi-automated tools based on predefined size and shape tuning) to minimize the segmentation error in solid nodule abutting the pleura (N, volume rendering with nodule segmentation noted by red surface). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

radiologist and CAD, $p < 0.003$ [7]. Similar research has since taken place in 'real-world' LDCT lung cancer screening. Liang et al., investigated the value of CAD systems in reducing false-negative results. They reported CAD would be valuable as a second reader. Up to 70 % of lung cancers which were undetected by radiologists during LDCT screening were subsequently detected by the CAD system. However, CAD failed to detect 20 % of lung cancers previously detected by radiologists [8]. Liu et al., has also confirmed the usefulness of AI in nodule identification and management. Their deep learning model was robust and showed good sensitivity when compared to manual human review. Additionally, their model was not dependent on radiation dose, CT-scanner manufacturer, or patient characteristics [22]. Heuvelmans et al., has recently published additional supporting evidence for value of AI. They used US National Lung Cancer Screening Trial (NLST) data to train a Lung Cancer Prediction Convolutional Neural Network (LCP-CNN) to generate a malignancy score for each lung nodule detected. When independently evaluated on a European multicenter trial dataset, the LCP-CNN performed excellently in the identification of benign nodules [23].

While a minority of FP findings, in our study represented by positive-misclassifications, is related to interpretation of focal findings, we suspect that FP findings of AI could largely be due to nodule attachment (for example vessel or pleural attachment, due to overestimation of nodule size (see Fig. 4). Multiple studies have shown that through the use of AI-based vessel-suppression, the detection and classification of lung nodules can be improved [24–26]. Hence, further refinement of the AI system to exclude nodule attachment could add further value. During a nuanced analysis of NM findings, we found four out of eight NM results reported by AI were due to misclassification of the nodule (part-solid in place of solid), despite the AI being able to correctly identify the size of the solid component ($\geq$100 mm^3). Therefore, through fine-tuning of this AI to correctly categorize these four nodules, we could yield NM results equivalent to that of the best performing reader. A comparison between AI nodule findings and radiologists findings is shown in Fig. 5.

To the best of our knowledge, we show for the first time that AI

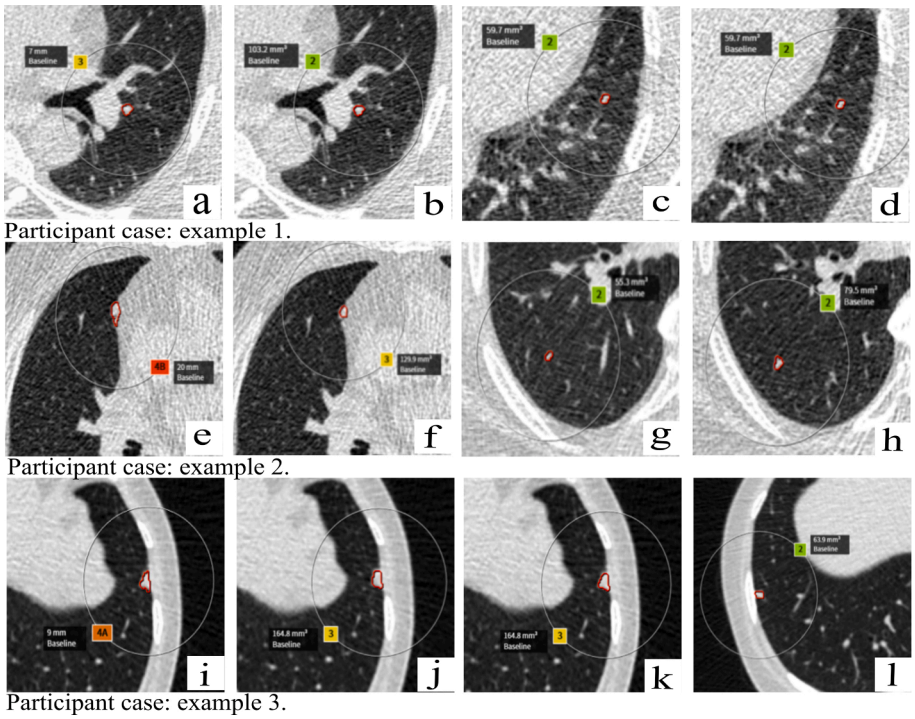

Fig. 5. Detection negative-misclassification's of radiologists: example participant cases of radiologists reporting inconsistently the dominant central solid nodule (above 100 mm^3), and heterogeneous annotation of dominant solid nodules. Example 1. Axial CT reformation showing a solid nodule (noted by red outline) classified above 100 by AI (A, 117 mm^3 [maximum diameter 7 mm]) and reader 3 (B, 103.2 mm^3), whereas reader 1 and 2 classified below 100 mm^3 (C and D, 59.7 mm^3), thus representing a "detection negative-misclassification" for two radiologists as well as underscoring the heterogeneity of visual reading. Example 2. Axial CT reformation showing a solid nodule (noted by red outline) classified above 100 mm^3 by AI (E, 533 mm^3 [maximum diameter 20 mm]) and reader 3 (F, 129.9 mm^3). Of note, the volume for such solid nodule with large surface abutting the mediastinal pleura resulted in over-segmentation by fully automatic AI segmentation (E), which could be edited by human reader with dedicated editing tools for size and shape carving towards optimal tailoring of nodule contour (F). The further human readers annotated different dominant solid nodule, which was below 100 mm^3 both for reader 1 (G, 55.3 mm^3) and reader 2 (H, 79.5 mm^3), representing a "detection negative-misclassification" for two radiologists. This case underscores the complementary performance of AI and radiologists for detection and measurement of solid nodules, with optimal classification performance from combination of the high sensitivity of AI and the human refinement of granular segmentation of nodules abutting solid structures. Example 3. Axial CT reformation showing a solid nodule (noted by red outline) classified above 100 mm^3 by AI (I, 170 mm^3 [maximum diameter 9 mm]), reader 2 and 3 (J and K, 168.4 mm^3), whereas reader 1 annotated a different dominant solid nodule below 100 mm^3 (L, 63.9 mm^3), thus representing a "detection negative-misclassification" for one radiologist. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

acting as an impartial reader in baseline screening can significantly reduce a radiologist's workload whilst not compromising on false-negative results of lung cancer screening with volume-based management of nodules and ultra-LDCT. Radiologists would then only need to read scans where nodules $\geq$ 100 mm^3 are present in order to determine the follow-up strategy, instead of reading all scans. To confirm this outcome, this AI diagnostic algorithm should be implemented as a first read filter in an independent non-selected LDCT-lung cancer screening dataset including participants without nodules, to rule out lung nodules $<$ 100 mm^3 (benign nodules). The negative predictive value of AI will likely improve due to the inclusion of participants with no nodules. Should the results be confirmed, AI could be used as a first reader in lung cancer screening, which will be a major step in the standardization and implementation of lung cancer screening worldwide.

Our study nevertheless has limitations. Our dataset only contained ultra-LDCT-scans where lung nodules were known to be present, which is not representative of lung cancer screening in the general population. We know from previous LDCT lung cancer screening trials, depending on the detection limit, roughly 50 % of participants have no reported lung nodules [4]. Therefore, our study is likely to considerably overestimate the rate of positive-misclassification results. If the same AI was used in a general LDCT-lung cancer screening population, depending on the FP rate in nodule negative participants, we could expect a workload reduction of 77.4 %–86.7 %. Second, the use of AI is currently limited to lung nodules, which sets apart from other incidental findings episodically reported by CT in lung cancer screening such as mediastinal tumors, extrathoracic tumors, and non-neoplastic disease. The reporting of incidental findings is a much-debated topic. A consensus statement from the British Society of Cardiovascular Imaging/British Society of Cardiac Computed Tomography (BSCI/BSCCT) and the British Society of Thoracic Imaging (BSTI) recommends that where the heart can be visualized on a CT-scan, it is reviewed [27]. However, in a study which used CT-scans from the NELSON trial, the impact of such incidental extra-pulmonary findings was demonstrated trivial compared to the specific purpose of lung cancer screening, notably without perceivable advantage for the small number of incidental detections [28]. AI has nevertheless proved to be valuable in coronary artery calcium (CAC) scoring of LDCT thorax scans. This has been investigated on low-dose electrocardiography-triggered cardiac CT scans from the ROBINSCA trial, and AI showed high agreement with manual CAC scoring (k = 0.87; 95 % CI: 0.85–0.89) [29]. Third, non-overlapping reconstructions were used, which are known to be suboptimal for the use of volume segmentation by software [13]. Segmentation performance is expected to improve by overlapping slice reconstruction.

To conclude, we have shown that AI can achieve a lower negative-misclassification result, surpassing that of four experienced radiologists, and when using AI in ultra-LDCT lung cancer screening, radiologists' workload could be diminished by up to 86.7 %. This AI diagnostic algorithm should now be implemented as a first read filter in an independent non-selected lung cancer screening dataset to rule out nodules $<$ 100 mm^3.

Author contributions

Harriet L. Lancaster: Conceptualization, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Sunyi Zheng:** Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing – review & editing. **Olga O. Aleshina:** Data Curation, Formal analysis, Writing – review & editing. **Donghoon Yu:** Resources, Software, Writing – review & editing. **Valeria Yu. Chernina:** Data Curation, Formal analysis, Writing – review & editing. **Marjolein A. Heuvelmans:** Conceptualization, Formal

analysis, Methodology, Visualization, Writing – review & editing, Supervision. **Geertruida H. de Bock:** Methodology, Writing – review & editing. **Monique D. Dorrius:** Data Curation, Formal analysis, Writing – review & editing. **Jan Willem Gratama:** Data Curation, Formal analysis, Writing – review & editing. **Sergey P. Morozov:** Data Curation, Formal analysis, Writing – review & editing. **Victor A. Gombolevskiy:** Resources, Data curation, Formal analysis, Writing – review & editing. **Mario Silva:** Conceptualization, Formal analysis, Methodology, Visualization, Writing – review & editing. **Jaeyoun Yi:** Resources, Software, Writing – review & editing. **Matthijs Oudkerk:** Conceptualization, Data curation, Resources, Formal analysis, Methodology, Visualization, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: M. Oudkerk is the scientific director and shareholder at the iDNA, and J. Yi is the scientific director and shareholder at Coreline Soft.

Acknowledgements

The authors express their gratitude to all doctors of the Moscow Health Department who implemented the lung cancer screening. Especially MD Ivan A. Blokhin for the administrative part of the project.

Funding

Olga O. Aleshina, Valeria Yu. Chernina, Sergey P. Morozov, and Victor A. Gombolevskiy received support as part of research (No. in the Unified State Information System for Accounting of Research, Development, and Technological Works (EGISU): AAAA-A20-120071090058-7) under the Program of the Moscow Healthcare Department "Scientific Support of the Capital's Healthcare" for 2020–2022.

Data statement

The dataset used for this study is registered «MosMedData: results of ultralow-dose computed tomography studies with lung nodules in the Moscow Lung Cancer Screening» №2020622727 from 21.12.2020.

References

[1] The Global Cancer Observatory, All cancers, 2020. https://gco.iarc.fr/today (accessed April 27, 2021).
[2] H.J. de Koning, C.M. van der Aalst, P.A. de Jong, E.T. Scholten, K. Nackaerts, M.A. Heuvelmans, J.-W.J. Lammers, C. Weenink, U. Yousaf-Khan, N. Horeweg, S. van 't Westeinde, M. Prokop, W.P. Mali, F.A.A. Mohamed Hoesein, P.M.A. van Ooijen, J. G.J.V. Aerts, M.A. den Bakker, E. Thunnissen, J. Verschakelen, R. Vliegenthart, J.E. Walter, K. ten Haaf, H.J.M. Groen, M. Oudkerk, Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial, N. Engl. J. Med. (2020). 10.1056/nejmoa1911793.
[3] S.W. Duffy, J.K. Field, Mortality Reduction with Low-Dose CT Screening for Lung Cancer, N. Engl. J. Med. 382 (6) (2020) 572–573, https://doi.org/10.1056/NEJMe1916361.
[4] M. Oudkerk, S. Liu, M.A. Heuvelmans, J.E. Walter, J.K. Field, Lung cancer LDCT screening and mortality reduction — evidence, pitfalls and future perspectives, Nat. Rev. Clin. Oncol. (2020) 1–17, https://doi.org/10.1038/s41571-020-00432-6.
[5] The Royal College of Radiologists, Clinical radiology UK workforce census 2018 report, 2019. https://www.rcr.ac.uk/system/files/publication/field_publication_files/clinical-radiology-uk-workforce-census-report-2018.pdf (accessed May 5, 2021).
[6] S. Ather, T. Kadir, F. Gleeson, Artificial intelligence and radiomics in pulmonary nodule management: current status and future applications, Clin. Radiol. 75 (2020), https://doi.org/10.1016/j.crad.2019.04.017.
[7] A. Christe, L. Leidolt, A. Huber, P. Steiger, Z. Szucs-Farkas, J.E. Roos, J. T. Heverhagen, L. Ebner, 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. 82 (2013) e873–8, https://doi.org/10.1016/j.ejrad.2013.08.026.
[8] M. Liang, W. Tang, D.M. Xu, A.C. Jirapatnakul, A.P. Reeves, C.I. Henschke, D. Yankelevitz, Low-Dose CT Screening for Lung Cancer: Computer-aided Detection of Missed Lung Cancers, Radiology 281 (2016) 279–288, https://doi.org/10.1148/radiol.2016150063.
[9] C.C. Lachance, M. Walter, Artificial Intelligence for Classification of Lung Nodules: A Review of Clinical Utility, Cost-Effectiveness, and Guidelines, Canadian Agency for Drugs and Technologies in Health, Diagnostic Accuracy, 2020 http://www.ncbi.nlm.nih.gov/pubmed/33074628 (accessed May 5, 2021).
[10] M. Ludwig, E. Chipon, J. Cohen, E. Reymond, M. Medici, A. Cole, A. Moreau Gaudry, G. Ferretti, Detection of pulmonary nodules: a clinical study protocol to compare ultra-low dose chest CT and standard low-dose CT using ASIR-V, BMJ Open 9 (2019) e025661, https://doi.org/10.1136/bmjopen-2018-025661.
[11] M. Oudkerk, A. Devaraj, R. Vliegenthart, T. Henzler, H. Prosch, C.P. Heussel, G. Bastarrika, N. Sverzellati, M. Mascalchi, S. Delorme, D.R. Baldwin, M. E. Callister, N. Becker, M.A. Heuvelmans, W. Rzyman, M.V. Infante, U. Pastorino, J. H. Pedersen, E. Paci, S.W. Duffy, H. de Koning, J.K. Field, European position statement on lung cancer screening, Lancet Oncol. 18 (2017) e754–e766, https://doi.org/10.1016/S1470-2045(17)30861-6.
[12] S.P. Morozov, E.S. Kuzmina, N.N. Vetsheva, V.A. Gombolevskiy, Z.A. Lantukh, N. S. Polishuk, A.Sh. Laipan, S.O. Ermolaev, E.V. Panina, I.A. Blokhin, Moscow Screening: Lung Cancer Screening With Low-Dose Computed Tomography, Probl. Sotsial'noi Gig. Zdr. i Istor. Meditsiny 27 (Special Issue) (2019), https://doi.org/10.32687/0869-866X-2019-27-si1-630-636.
[13] Hans-Ulrich Kauczor, Anne-Marie Baird, Torsten Gerriet Blum, Lorenzo Bonomo, Clementine Bostantzoglou, Otto Burghuber, Blanka Čepická, Alina Comanescu, Sébastien Couraud, Anand Devaraj, Vagn Jespersen, Sergey Morozov, Inbar Nardi Agmon, Nir Peled, Pippa Powell, Helmut Prosch, Sofia Ravara, Janette Rawlinson, Marie-Pierre Revel, Mario Silva, Annemiek Snoeckx, Bram van Ginneken, Jan P. van Meerbeeck, Constantine Vardavas, Oyunbileg von Stackelberg, Mina Gaga, ESR/ERS statement paper on lung cancer screening, Eur. Radiol. 30 (6) (2020) 3277–3294, https://doi.org/10.1007/s00330-020-06727-7.
[14] L. Garzelli, J.M. Goo, S.Y. Ahn, K.J. Chae, C.M. Park, J. Jung, H. Hong, Improving the prediction of lung adenocarcinoma invasive component on CT: Value of a vessel removal algorithm during software segmentation of subsolid nodules, Eur. J. Radiol. 100 (2018) 58–65, https://doi.org/10.1016/j.ejrad.2018.01.016.
[15] J. Jung, H. Hong, J.M. Goo, Ground-glass nodule segmentation in chest CT images using asymmetric multi-phase deformable model and pulmonary vessel removal, Comput. Biol. Med. 92 (2018) 128–138, https://doi.org/10.1016/j.compbiomed.2017.11.013.
[16] H. MacMahon, D.P. Naidich, J.M. Goo, K.S. Lee, A.N.C. Leung, J.R. Mayo, A. C. Mehta, Y. Ohno, C.A. Powell, M. Prokop, G.D. Rubin, C.M. Schaefer-Prokop, W. D. Travis, P.E. Van Schil, A.A. Bankier, Guidelines for management of incidental pulmonary nodules detected on CT images: From the Fleischner Society 2017, Radiology 284 (2017) 228–243, https://doi.org/10.1148/radiol.2017161659.
[17] N. Horeweg, E.T. Scholten, P.A. De Jong, C.M. Van Der Aalst, C. Weenink, J.-W.J. Lammers, K. Nackaerts, R. Vliegenthart, K. Ten Haaf, U.A. Yousaf-Khan, M.A. Heuvelmans, E. Thunnissen, M. Oudkerk, W. Mali, H.J. De Koning, Detection of lung cancer through low-dose CT screening (NELSON): a prespecifi ed analysis of screening test performance and interval cancers, Lancet Oncol. 15 (2014) 1342–1350. 10.1016/S1470-2045(14)70387-0.
[18] H.L. Lancaster, M.A. Heuvelmans, G.J. Pelgrim, M. Rook, M.G.J. Kok, A. Aown, G. H. de Bock, M. van den Berge, H.J.M. Groen, R. Vliegenthart, Seasonal prevalence and characteristics of low-dose CT detected lung nodules in a general Dutch population, Sci. Rep. 11 (2021) 9139, https://doi.org/10.1038/s41598-021-88328-y.
[19] Keith M. Sigel, Dongming Xu, Jonathan Weber, Juan P. Wisnivesky, Juan C. Celedón, Rafael E. de la Hoz, Prevalence of Pulmonary Nodules Detected by Computed Tomography in World Trade Center Rescue and Recovery Workers, Ann. Am. Thorac. Soc. 17 (1) (2020) 125–128, https://doi.org/10.1513/AnnalsATS.201907-517RL.
[20] H. Kim, H.Y. Kim, J.M. Goo, Y. Kim, Lung Cancer CT Screening and Lung-RADS in a Tuberculosis-endemic Country: The Korean Lung Cancer Screening Project (K-LUCAS), Radiology 296 (2020) 181–188, https://doi.org/10.1148/radiol.2020192283.
[21] Mark Petticrew, Amanda Sowden, Deborah Lister-Sharp, False-negative results in screening programs, Int. J. Technol. Assess. Health Care 17 (2) (2001) 164–170, https://doi.org/10.1017/S0266462300105021.
[22] K. Liu, Q. Li, J. Ma, Z. Zhou, M. Sun, Y. Deng, W. Tu, Y. Wang, L. Fan, C. Xia, Y. Xiao, R. Zhang, S. Liu, Evaluating a Fully Automated Pulmonary Nodule Detection Approach and Its Impact on Radiologist Performance, Radiol. Artif. Intell. 1 (2019) e180084, https://doi.org/10.1148/ryai.2019180084.
[23] M.A. Heuvelmans, P.M.A. van Ooijen, S. Ather, C.F. Silva, D. Han, C.P. Heussel, W. Hickes, H.-U. Kauczor, P. Novotny, H. Peschl, M. Rook, R. Rubtsov, O. von Stackelberg, M.T. Tsakok, C. Arteta, J. Declerck, T. Kadir, L. Pickup, F. Gleeson, M. Oudkerk, Lung cancer prediction by Deep Learning to identify benign lung nodules, Lung Cancer 154 (2021) 1–4, https://doi.org/10.1016/j.lungcan.2021.01.027.
[24] Ramandeep Singh, Mannudeep K. Kalra, Fatemeh Homayounieh, Chayanin Nitiwarangkul, Shaunagh McDermott, Brent P. Little, Inga T. Lennes, Jo-Anne O. Shepard, Subba R. Digumarthy, Artificial intelligence-based vessel suppression for detection of sub-solid nodules in lung cancer screening computed tomography, Quant. Imaging Med. Surg. 11 (4) (2021) 1134–1143, https://doi.org/10.21037/qims10.21037/qims-20-630.
[25] Yung-Liang Wan, Patricia Wu, Pei-Ching Huang, Pei-Kwei Tsay, Kuang-Tse Pan, Nguyen Trang, Wen-Yu Chuang, Ching-Yang Wu, ShihChung Lo, The Use of Artificial Intelligence in the Differentiation of Malignant and Benign Lung Nodules on Computed Tomograms Proven by Surgical Pathology, Cancers (Basel) 12 (8) (2020) 2211, https://doi.org/10.3390/cancers12082211.

[26] G. Milanese, M. Eberhard, K. Martini, V. De Martini, T. Frauenfelder, Vessel suppressed chest Computed Tomography for semi-automated volumetric measurements of solid pulmonary nodules (2018), https://doi.org/10.1016/j.ejrad.2018.02.020.

[27] Michelle Claire Williams, Ausami Abbas, Erica Tirr, Shirjel Alam, Edward Nicol, James Shambrook, Matthias Schmitt, Gareth Morgan Hughes, James Stirrup, Ben Holloway, Deepa Gopalan, Aparna Deshpande, Jonathan Weir-McCall, Bobby Agrawal, Jonathan C L Rodrigues, Adrian J B Brady, Giles Roditi, Graham Robinson, Russell Bull, Reporting incidental coronary, aortic valve and cardiac calcification on non-gated thoracic computed tomography, a consensus statement from the BSCI/BSCCT and BSTI, Br. J. Radiol. 94 (1117) (2021) 20200894, https://doi.org/10.1259/bjr.20200894.

[28] J.C.M. van de Wiel, Y. Wang, D.M. Xu, H.J. van der Zaag-Loonen, E.J. van der Jagt, R.J. van Klaveren, M. Oudkerk, Neglectable benefit of searching for incidental findings in the Dutch-Belgian lung cancer screening trial (NELSON) using low-dose multidetector CT, Eur. Radiol. 17 (6) (2007) 1474–1482, https://doi.org/10.1007/s00330-006-0532-7.

[29] M. Vonder, S. Zheng, M.D. Dorrius, C.M. van der Aalst, H.J. de Koning, J. Yi, D. Yu, J.W.C. Gratama, D. Kuijpers, M. Oudkerk, Deep Learning for Automatic Calcium Scoring in Population-Based Cardiovascular Screening, JACC Cardiovasc. Imag. (2021), https://doi.org/10.1016/j.jcmg.2021.07.012.