Lung cancer prediction by Deep Learning to identify benign lung nodules

Heuvelmans, Marjolein A; van Ooijen, Peter M A; Ather, Sarim; Silva, Carlos Francisco; Han, Daiwei; Heussel, Claus Peter; Hickes, William; Kauczor, Hans-Ulrich; Novotny, Petr; Peschl, Heiko

Published in:
Lung Cancer

DOI:
10.1016/j.lungcan.2021.01.027

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2021

Lung Cancer 154 (2021) 1–4

Contents lists available at ScienceDirect

Lung Cancer

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

Lung cancer prediction by Deep Learning to identify benign lung nodules

Marjolein A. Heuvelmans [a,b,*], Peter M.A. van Ooijen [c], Sarim Ather [d], Carlos Francisco Silva [e,g,n], Daiwei Han [f], Claus Peter Heussel [e,g,n], William Hickes [h], Hans-Ulrich Kauczor [e,g,n], Petr Novotny [i,j], Heiko Peschl [k], Mieneke Rook [f,l], Roman Rubtsov [e,g,n], Oyunbileg von Stackelberg [e,g,n], Maria T. Tsakok [k], Carlos Arteta [i], Jerome Declerck [i], Timor Kadir [i], Lyndsey Pickup [i], Fergus Gleeson [h], Matthijs Oudkerk [m]

[a] University of Groningen, University Medical Center Groningen Groningen, Department of Epidemiology, Groningen, The Netherlands
[b] Department of Pulmonary Medicine, Medisch Spectrum Twente, Enschede, The Netherlands
[c] University of Groningen, University Medical Center Groningen Groningen, Department of Radiation Oncology, Groningen, The Netherlands
[d] Department of Radiology, Churchill Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom
[e] Diagnostic and Interventional Radiology, University Hospital Heidelberg, Heidelberg, Germany
[f] University of Groningen, University Medical Center Groningen Groningen, Department of Radiology, Groningen, The Netherlands
[g] Translational Lung Research Center, Member of the German Lung Research Center, Heidelberg, Germany
[h] Oxford University Hospitals NHS Foundation Trust, United Kingdom
[i] Optellum Ltd., Oxford, United Kingdom
[j] Department of Respiratory Sciences, College of Life Sciences, University of Leicester, Leicester, United Kingdom
[k] Department of Radiology, Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[l] Martini Hospital Groningen, Department of Radiology, Groningen, The Netherlands
[m] University of Groningen, Faculty of Medical Sciences, Groningen, The Netherlands
[n] Department of Diagnostic and Interventional Radiology With Nuclear Medicine, Thoraxklinik, Heidelberg University Hospital, Heidelberg, Germany

ARTICLE INFO

Keywords:
Lung cancer
Screening
Pulmonary nodule
Deep learning

ABSTRACT

Introduction: Deep Learning has been proposed as promising tool to classify malignant nodules. Our aim was to retrospectively validate our Lung Cancer Prediction Convolutional Neural Network (LCP-CNN), which was trained on US screening data, on an independent dataset of indeterminate nodules in an European multicentre trial, to rule out benign nodules maintaining a high lung cancer sensitivity.

Methods: The LCP-CNN has been trained to generate a malignancy score for each nodule using CT data from the U.S. National Lung Screening Trial (NLST), and validated on CT scans containing 2106 nodules (205 lung cancers) detected in patients from from the Early Lung Cancer Diagnosis Using Artificial Intelligence and Big Data (LUCINDA) study, recruited from three tertiary referral centers in the UK, Germany and Netherlands. We pre-defined a benign nodule rule-out test, to identify benign nodules whilst maintaining a high sensitivity, by calculating thresholds on the malignancy score that achieve at least 99 % sensitivity on the NLST data. Overall performance per validation site was evaluated using Area-Under-the-ROC-Curve analysis (AUC).

Results: The overall AUC across the European centers was 94.5 % (95 %CI 92.6–96.1). With a high sensitivity of 99.0 %, malignancy could be ruled out in 22.1 % of the nodules, enabling 18.5 % of the patients to avoid follow-up scans. The two false-negative results both represented small typical carcinoids.

Abbreviations: AI, arteficial intelligence; AUC, area-under-the-ROC-curve analysis; CT, computed tomography; LCP-CNN, lung cancer prediction convolutional neural network; NLST, National Lung Screening Trial.

* Corresponding author at: University of Groningen, University Medical Center Groningen Groningen, Department of Epidemiology, Groningen, The Netherlands.
E-mail addresses: m.a.heuvelmans@umcg.nl (M.A. Heuvelmans), p.m.a.van.ooijen@umcg.nl (P.M.A. van Ooijen), Sarim.Ather@ouh.nhs.uk (S. Ather), Carlos.DaSilva@med.uni-heidelberg.de (C.F. Silva), d.han@umcg.nl (D. Han), heussel_elsevier2019@contbay.com (C.P. Heussel), William.Hickes@ouh.nhs.uk (W. Hickes), Hans-Ulrich.Kauczor@med.uni-heidelberg.de (H.-U. Kauczor), pnovotnyq@gmail.com (P. Novotny), Heiko.Peschl@ouh.nhs.uk (H. Peschl), m.rook@umcg.nl (M. Rook), roman.rubtsov.med@gmail.com (R. Rubtsov), Oyunbileg.Stackelberg@med.uni-heidelberg.de (O. von Stackelberg), mariatsakok@ouh.nhs.uk (M.T. Tsakok), carlos.arteta@optellum.com (C. Arteta), jerome.declerck@optellum.com (J. Declerck), timor.kadir@optellum.com (T. Kadir), lyndsey.pickup@optellum.com (L. Pickup), fergus.gleeson@ouh.nhs.uk (F. Gleeson), m.oudkerk@umcg.nl (M. Oudkerk).

https://doi.org/10.1016/j.lungcan.2021.01.027
Received 3 November 2020; Received in revised form 22 January 2021; Accepted 25 January 2021
Available online 31 January 2021

Conclusion: The LCP-CNN, trained on participants with lung nodules from the US NLST dataset, showed excellent performance on identification of benign lung nodules in a multi-center external dataset, ruling out malignancy with high accuracy in about one fifth of the patients with 5−15 mm nodules.

1. Introduction

Adequate differentiation of benign and malignant small-to-intermediate sized, 5–15 mm, pulmonary nodules detected by computed tomography (CT) is a challenge for radiologists. With the improvement of CT scanners, an increasing number of these pulmonary nodules are detected, both in routine clinical care and in a lung cancer screening setting. Approximately 50 % of smokers have a pulmonary nodule, [1] and 25 % have more than one, although less than 1% of these nodules are malignant [1]. Nodule classification for both incidentally detected and screening detected nodules are based on nodule type, size, and growth, according to Fleischner and Lung-RADS™ guidelines [2,3]. Despite their widespread adoption, these nodule management protocols still result in a rather high false-positive rate.

Recently, multiple studies have proposed using artificial intelligence (AI), and in particular Deep Learning, as a tool to characterize pulmonary nodules, and thereby potentially decrease the numbers of scans needed to achieve a benign or malignant diagnosis. [4] Currently, the sensitivity of these tools is only moderate, with most of these studies including larger nodules up to 30 mm in diameter. The large nodule size results in biased data sets, as most lung cancers are generally larger than benign nodules. Therefore, it remains debatable whether such Deep Learning methods identify nodule characteristics specific to lung cancer, or predominantly simply stratify nodules based on size.

By using a convolutional neural network (CNN) to specifically identify benign nodules as opposed to only possibly-malignant nodules, many follow-up CT examinations could potentially be avoided. We developed a CNN trained for the task of Lung Cancer Prediction (LCP-CNN), using the National Lung Sceening Trial (NLST) data, and configured it to specifically identify benign nodules, enabling them to be "ruled out"of unnecessary follow-up with a high degree of certainty. [5] Recently, validating this LCP-CNN on a retrospective UK dataset it was shown that the LCP-CNN outperformed the Brock University model for lung nodule risk categorization [6]. A second study showed that using the LCP-CNN, nodules could be classified into low (5% malignancy treshold) and high-risk (65 % malignancy treshold) categories, with improved accuracy compared to traditional risk prediction models [7]. The aim of this study was to validate the LCP-CNN on an independent dataset of small-to-intermediate sized (5−15 mm) nodules in a European multicentre trial, to rule out benign nodules whilst maintaining a high lung cancer sensitivity.

2. Methods

2.1. Training of the LCP-CNN

The LCP-CNN was trained using the NLST dataset. Study design and inclusion criteria of the NLST have been previously described.[7] In total, this dataset consisted of 10,368 participants: 9310 participants with only benign lung nodules and 1058 participants with lung cancer. In the training set, we included all the nodules as malignant that could be categorically linked in retrospect to a diagnosed lung cancer (N = 932 in 575 patients). We included all nodules in patients without a lung cancer diagnosis during the NLST (screening and follow-up until seven years after baseline) as benign nodules (N = 14,761 in 5972 participants). Details on the training set have been published [7].

The LCP-CNN was trained on this data to differentiate between benign and malignant nodules directly from the CT image by using a supervised learning approach. The algorithm learns by changing its parameters until its predictions agree with the nodules' known true diagnosis. Once trained, the model can predict how likely a new nodule is to be malignant or benign (i.e. the LCP-CNN classifier produces a malignancy score per nodule). We then defined a benign nodule rule-out test by calculating a threshold on the malignancy score on the NLST data, with a target sensitivity of 100 % (i.e. no cancers missed). The rule-out thresholds were defined using an eight-fold cross-validation technique.

2.2. Validation of the LCP-CNN

We validated the LCP-CNN software on lung nodules incidentally detected on thoracic CT images in patients from the Early Lung Cancer Diagnosis Using Artificial Intelligence and Big Data (LUCINDA) study. Patients were recruited from three hospitals; University Medical Center Groningen, Groningen, The Netherlands (site A), Heidelberg University Hospital (including Thoraxklinik Heidelberg), Heidelberg, Germany (site B), which is a tertiary referral center for lung cancer patients, and Oxford University Hospitals (including Royal Berkshire Hospital), Oxford, United Kingdom (site C). Inclusion and exclusion criteria are described in the *Appendix*.

Each site obtained local ethics committee approval. Data was retrospectively selected based on the presence of pulmonary nodules on a thoracic CT. All nodules measured between 5 and 15 mm diameter. Each reported lung nodule and lung cancer was located, contoured and labeled benign or malignant by clinicians, who had access to the ground truth diagnosis as decided based on the trial protocol (see *Appendix*). The CT data included heterogeneous scan parameters, and a variety of scanner manufacturers and clinical indications. Both normal dose, low-dose, contrast-enhanced and non-contrast enhanced thoracic CT scans were included. CT acquisitions are described in the *Appendix*.

2.3. Statistics

The LCP-CNN, trained on the NLST data, generated a malignancy score for each nodule in the external validation dataset. Overall performance was evaluated using Area-Under-the-ROC-Curve analysis (AUC). Benign rule-out performance, i.e. the proportion of benign nodules correctly stratified as benign by the software, was calculated by determing sensitivity and specificity at the pre-determined score threshold.

3. Results

3.1. Nodule characteristics

In total, 2106 unique nodules (205 malignant, 9.7 %) from 1650 unique patients (201 lung cancer patients, 12.2 %) were included in the validation set (Fig. 1). Median patient age was 63.0 (range 19–94), and 489 (29.6 %) were female. Nodule characteristics are presented in Table 1. More detailed information on nodule size and location is in Tables A1 and A2 in the *Appendix*.

3.2. Performance per site

Performance of the LCP-CNN for the identification of benign nodules differed per site. The overall AUC (Fig. 2) across the three centers was 94.5 % (95 %CI 92.6–96.1). For sites A, B and C respectively, AUCs were 98.5 % (95 %CI 96.6–99.6), 88.1 % (95 %CI 84.5–91.3) and 97.7 % (95 %CI 95.9–99.0). When comparing AUCs between the centers, center A and C perfomed significantly better compared to center B ($p < 0.01$ and

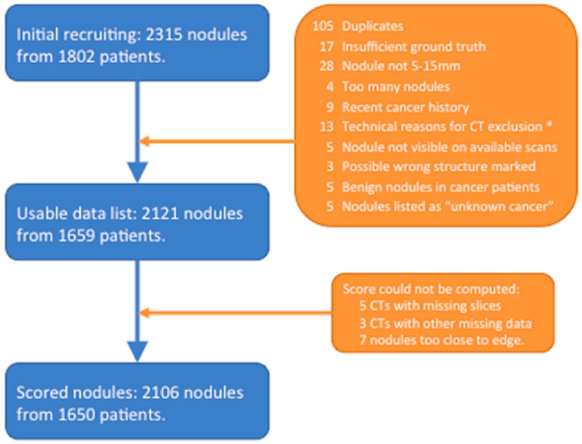

Fig. 1. Flowchart* Technical exclusion happened when the CT was not an original/primary/axial CT, had missing slices, or extreme motion artefacts around the nodule.

Table 1
Nodule information per site.

	Total	A	B	C
Number of Nodules	2106	855	567	684
Lung cancer (N, %)	205 (9.6)	25 (2.9)	136 (24.0)	44 (6.4)
Nodule size (median, IQR)	6.0 (5.0–8.4)	5.2 (5.0–6.6)	9.0 (7.0–12.0)	7.0 (5.0–8.0)
Lung cancer size (median, IQR)	12.0 (11.0–14.0)	13.0 (9.0–15.0)	12.0 (11.0–14.0)	12.5 (11.0–14.0)
Benign size (median, IQR)	6.0 (5.0–8.0)	5.2 (5.0–6.3)	8.0 (6.0–11.0)	6.0 (5.0–8.0)

Site A = Groningen, Site B = Heidelberg, Site C = Oxford.

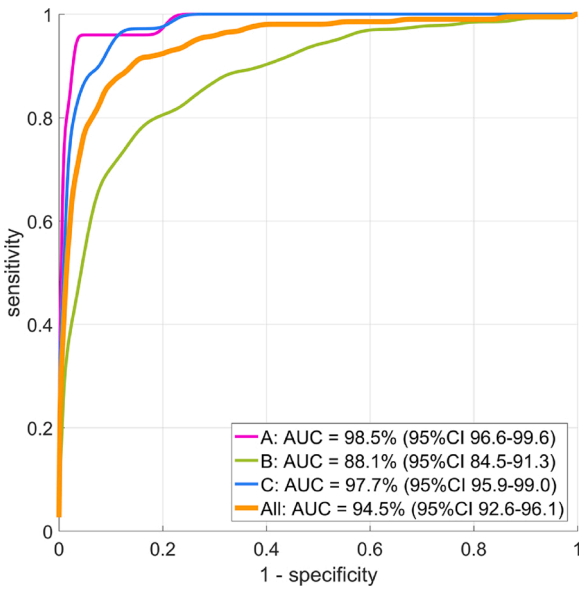

Fig. 2. Performance of the LCP-CNN in the test group. AUC is Area Under the ROC Curve.
Site A = Groningen, Site B = Heidelberg, Site C = Oxford

p < 0.001, respectively), which was a tertiary referral center where patients usually presented with a much higher pre-test probability of lung cancer. No significant difference was found between centers A and C (p = NS), both with patients with mainly incidentally detected nodules.

3.3. Benign-rule out performance

The pre-defined score threshold for a benign nodule achieved overall sensitivity of 99.0 % (95 %CI: 97.5 %–100.0 %) and specificity of 22.1 % (95 %CI: 20.2 %–24.9 %), examples see Figure A1 *Appendix*. Negative predictive value was 99.5 %. Using the threshold score for benignity, 420 benign nodules were correctly ruled out (11 had originally been diagnosed based on histology, 5 on resolution at follow-up, 191 by expert opinion [i.e. perifissural nodules], 133 by 1-year volumetric stability, and 80 by 2-year diameter stability). Two cancers (Fig A2 *Appendix*) received a false-negative result. Both cases represented typical carcinoids, sized 7 and 8 mm, both spherical with smooth surfaces. When considering patients instead of only nodules, the software identified 18.5 % (95 %CI: 16.5 %–20.6 %) of the patients correctly as benign, where the highest-scoring nodule in a patient was used for decision-making. Ruled-out nodules were often <8.0 mm in size and located in the middle lobe (Table A2 *Appendix)*.

4. Discussion

Our LCP-CNN, trained on participants with lung nodules from the NLST dataset, demonstrated excellent performance in identifying benign nodules, ruling out malignancy with a high degree of accuracy in one-fifth of patients with small-to-intermediate sized nodules. This indicates the potential value of using CNNs for lung cancer risk prediction decision support for incidentally detected lung nodules. By ruling out CT scans with a very high sensitivity (99.0 % in our study) unnecessary workup including imaging and invasive procedures may be avoided in a large number of patients. It requires prospective validation in a lung cancer screening program before considering whether it might also be used in this setting.

Previous AI studies have focused on maximizing the proportion of cancers correctly characterized (i.e. high positive-predictive value) and have shown promise. [8] However, sensitivity of these tools is still moderate, even with most of these studies including not the clinically relevant small-to-intermediate sized nodules, but large nodules up to 30 mm in diameter, limiting its clinical applicability. In contrast to using AI to identify lung cancers, the approach we have validated in this work was to use our AI system, the LCP-CNN, to identify benign nodules with a high degree of certainty, and to suggest that these nodules may be ruled out, preventing the need for further workup. Using a threshold set independently in advance from the NLST data, we have shown that it is possible to correctly identify 22.1 % of incidentally-detected benign nodules, and that over 18.5 % of patients would consequently require no further work-up for this nodule. When validating the same LCP-CNN in independent datasets, it already has been shown that the LCP-CNN outperforms the nodule classification model of the Brock University, and that it is able to stratify nodules into high- and low malignancy risk categories [6,7].

Although the LCP-CNN ruled out 22 % of benign nodules, two malignancies were missed. Both represented small (7 and 8 mm diameter, respectively) typical carcinoids. These are low-grade lung neuroendocrine tumors, representing 1–2 % of all lung malignancies. [9] Both carcinoids were spherical, and had smooth surfaces, which may be the possible explanation for the false-negative result of the LCP-CNN.

Although the model risk prediction score thresholds used for the target sensitivity of 99 % were the same for the three centers, performance of our LCP-CNN slightly varied across patients of the centers. This difference might be explained by the use of different scan parameters, and the varying lung cancer rate among patients ranging between 2.9 %

and 24.0 % from the three centers. In center A and C, mainly patients with incidentally detected nodules and lung cancers were included, while center B is a tertiary referral center investigating the more difficult cases with high pre-test probability of malignancy. Best performance was found for the centers with lowest lung cancer rate and largest difference in median size between lung cancers and benign nodules. Still, overall performance of the software was very good, considering that 25 % of the cases in the external dataset came from a tertiary referral center with mainly patients with nodules highly suspicious of being malignant. In a lung cancer screening program, quality assurance and standardized scanning parameters are mandated, [10] and it is likely that our LCP-CNN may perform even better.

In conclusion, the LCP-CNN, trained on participants with lung nodules from the NLST dataset, showed excellent performance on identification of benign lung nodules in a multi-center external dataset, ruling out malignancy with high sensitivity in about one fifth of patients with intermediate-sized nodules.

CRediT authorship contribution statement

Marjolein A. Heuvelmans: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Recources, Writing - original draft, Writing - review & editing. **Peter M.A. van Ooijen:** Conceptualization, Funding acquisition, Investigation, Recources, Writing - original draft. **Sarim Ather:** Conceptualization, Funding acquisition, Investigation, Recources, Writing - original draft. **Carlos Francisco Silva:** Conceptualization, Funding acquisition, Investigation, Recources, Writing - original draft. **Daiwei Han:** Conceptualization, Investigation, Recources, Writing - original draft. **Claus Peter Heussel:** Conceptualization, Funding acquisition, Investigation, Recources, Writing - original draft. **William Hickes:** Conceptualization, Funding acquisition, Investigation, Recources, Writing - original draft. **Hans-Ulrich Kauczor:** Conceptualization, Funding acquisition, Investigation, Recources, Writing - original draft. **Petr Novotny:** Conceptualization, Investigation, Recources, Writing - original draft. **Heiko Peschl:** Conceptualization, Funding acquisition, Investigation, Recources, Writing - original draft. **Mieneke Rook:** Conceptualization, Investigation, Recources, Writing - original draft. **Roman Rubtsov:** Conceptualization, Funding acquisition, Investigation, Recources, Writing - original draft. **Oyunbileg von Stackelberg:** Conceptualization, Funding acquisition, Investigation, Recources, Writing - original draft. **Maria T. Tsakok:** Conceptualization, Funding acquisition, Investigation, Recources, Writing - original draft. **Carlos Arteta:** Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Jerome Declerck:** Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Timor Kadir:** Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Lyndsey Pickup:** Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Fergus Gleeson:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Recources, Writing - original draft, Writing - review & editing. **Matthijs Oudkerk:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Recources, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

None.

Acknowledgements

This activity has received funding from the European Institute of Innovation and Technology (EIT) under grant agreement No 17189. This European body receives support from the Horizon 2020 research and innovation programme.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.lungcan.2021.01.027.

References

[1] N. Horeweg, J. van Rosmalen, M.A. Heuvelmans, C.M. van der Aalst, R. Vliegenthart, E.Th Scholten, K. Ten Haaf, K. Nackaerts, J.W. Lammers, C. Weenink, H.J. Groen, P. van Ooijen, et al., Lung cancer probability in patients with CT-detected pulmonary nodules: a prespecified analysis of data from the NELSON trial of low-dose CT screening, Lancet Oncol. 15 (2014) 1332–1341, https://doi.org/S1470-2045(14)70389-4 [pii].

[2] 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 161659 (2017), https://doi.org/10.1148/radiol.2017161659.

[3] American College of Radiology, Lung CT Screening Reporting and Data System (Lung-RADS), July 7, 20, 2014, http://www.acr.org/Quality-Safety/Resources/LungRADS.

[4] F. Ciompi, K. Chung, S.J. Van Riel, A.A.A. Setio, P.K. Gerke, C. Jacobs, E. Th Scholten, C. Schaefer-Prokop, M.M.W. Wille, A. Marchianò, U. Pastorino, M. Prokop, B. Van Ginneken, Towards automatic pulmonary nodule management in lung cancer screening with deep learning, Sci. Rep. 7 (2017), https://doi.org/10.1038/srep46479.

[5] H. Peschl, D. Han, P. Van Ooijen, M. Oudkerk, et al., Lung Cancer prediction using deep learning software: validation on independent multi-centre data, J. Thorac. Oncol. 13 (2018) S428, https://doi.org/10.1016/j.jtho.2018.08.489.

[6] D.R. Baldwin, J. Gustafson, L. Pickup, C. Arteta, P. Novotny, J. Declerck, T. Kadir, C. Figueiras, A. Sterba, A. Exell, V. Potesil, P. Holland, H. Spence, A. Clubley, E. O'Dowd, M. Clark, V. Ashford-Turner, M.E.J. Callister, F.V. Gleeson, External validation of a convolutional neural network artificial intelligence tool to predict malignancy in pulmonary nodules, Thorax (2020), https://doi.org/10.1136/thoraxjnl-2019-214104.

[7] P.P. Massion, S. Antic, S. Ather, C. Arteta, J. Brabec, H. Chen, J. Declerck, D. Dufek, W. Hickes, T. Kadir, J. Kunst, B.A. Landman, R.F. Munden, P. Novotny, H. Peschl, L.C. Pickup, C. Santos, G.T. Smith, A. Talwar, F. Gleeson, Assessing the accuracy of a deep learning method to risk stratify indeterminate pulmonary nodules, Am. J. Respir. Crit. Care Med. (2020), https://doi.org/10.1164/rccm.201903-0505oc.

[8] D. Ardila, A.P. Kiraly, S. Bharadwaj, B. Choi, J.J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D.P. Naidich, S. Shetty, End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography, Nat. Med. 25 (2019) 954–961, https://doi.org/10.1038/s41591-019-0447-x.

[9] B. Melosky, Advanced typical and atypical carcinoid tumours of the lung: management recommendations, Curr. Oncol. 25 (2018) S86–S93, https://doi.org/10.3747/co.25.3808.

[10] 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.