EUROPEAN SOCIETY OF RADIOLOGY
European Radiology

CHEST Open Access

Multicentre performance and consistency of two deep learning models for malignancy probability estimation of incidental pulmonary nodules

Renate Dinnessen[1*], Noa Antonissen[1], Dré Peeters[1], Hester A. Gietema[2,3], Michiel T. H. M. Henkens[4,5], Firdaus A. A. Mohamed Hoesein[6], Ernst Th. Scholten[1], Cornelia Schaefer-Prokop[1,7] and Colin Jacobs[1]

Abstract

Objectives A screening-trained deep learning (DL) model (DL1) for pulmonary nodule malignancy probability estimation on CT previously demonstrated good discrimination on a single-centre dataset of incidental nodules. An updated DL model (DL2) was trained on both screening and clinical data. We aimed to test the performance of both models in a multicentre dataset of incidental nodules.

Materials and methods A retrospective, multicentre, case-control dataset of incidental nodules was collected, sampled across size buckets (5–10 mm, 10–15 mm, 15–30 mm), aiming for 10 malignant and 20 benign nodules per bucket and centre, resulting in 270 nodules. Performance was assessed using AUCs and specificity at a fixed sensitivity. AUCs were compared using the DeLong method. Both DL models were compared with the Brock model. Consistent discrimination was investigated by centre-stratified analyses.

Results The multicentre dataset contained 269 nodules (89 malignant) from 231 patients. DL1 and DL2 achieved AUCs of 0.74 and 0.72, respectively, versus 0.63 for Brock (both $p < 0.01$). Using a 10% threshold for the Brock model (sensitivity 77.5%), specificity was 60% for both DL models compared with 44% for the Brock model. Centre-specific AUCs for DL1, DL2, and Brock were 0.76, 0.75, and 0.74 (centre 1), 0.72, 0.71, and 0.55 (centre 2), and 0.73, 0.71, and 0.59 (centre 3), respectively.

Conclusion The DL models outperformed the Brock model on a multicentre, cancer-enriched size-stratified dataset and performed consistently across centres, though prospective validation in representative cohorts with real-world prevalences is needed. The DL model trained with additional clinical data performed equal to the screening-trained model.

Key Points

Question *How consistent is the performance of a screening-trained deep learning model across multicentre incidental pulmonary nodules, and does an additionally clinically trained model improve performance?*

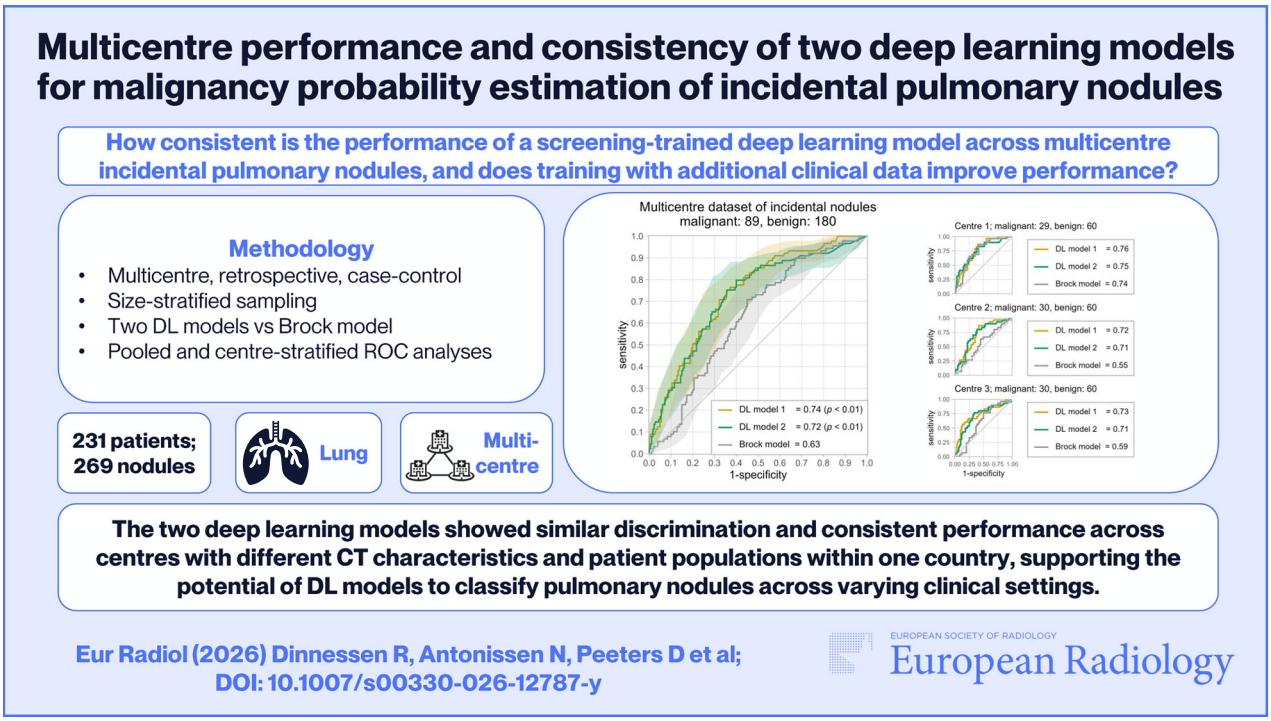
Findings *The two deep learning models showed similar discrimination, outperforming the clinically used Brock model in the pooled cancer-enriched dataset, and consistent performance across multiple centres.*

Clinical relevance *The current study provides evidence of consistent model performance across centres with varying CT characteristics and patient populations within one country, supporting the potential of DL models to classify pulmonary nodules across different clinical settings.*

*Correspondence:
Renate Dinnessen
renate.dinnessen@radboudumc.nl
Full list of author information is available at the end of the article

Keywords Artificial intelligence, Lung, Neoplasms, Solitary pulmonary nodule, Tomography (X-ray computed)

Graphical Abstract



Introduction

Pulmonary nodules are frequently incidentally detected on computed tomography (CT) scans performed for other reasons than the detection of malignancies [1, 2]. Accurate estimation of their malignancy probability is important for guiding appropriate follow-up and intervention [3, 4]. This assessment of the malignancy probability relies on radiological features, information from previous scans, and clinical information [3, 4].

The Brock model is a widely used malignancy prediction model, which combines imaging and patient features through predefined variables [5]. This constrained model may limit accurate risk stratification, particularly in the more heterogeneous clinical setting. Artificial intelligence models that estimate the malignancy probability of pulmonary nodules on CT scans have demonstrated strong performance, showing the potential to support decision-making [6–9].

Although several studies have reported high diagnostic accuracy for pulmonary nodule malignancy probability estimation, most have focused on more homogeneous lung cancer screening cohorts or single-centre samples [6–9]. Models used across centres may encounter differences in patient populations, scanners, and acquisition and reconstruction protocols that can affect performance. Especially, validation in incidental nodules from routine clinical care has been reported less frequently and largely in single-centre samples. One study reported good multicentre performance of a deep learning (DL) algorithm in incidental nodules [8, 9]; however, evidence on robustness across centres remains limited.

In addition, most existing models have been trained primarily on data from lung cancer screening studies, which may not reflect the heterogeneity in patient characteristics and underlying risk factors encountered in routine clinical practice. A review of DL models for pulmonary nodule malignancy probability estimation reported substantial heterogeneity in performance, with data source (screen-detected versus incidentally detected nodules) identified as an important contributing factor [10]. In previous work, we demonstrated that a screening-trained DL model maintained good diagnostic performance when applied to a clinical population with incidental nodules in a single-centre dataset, indicating that translation from screening to clinical settings is possible [11]. Nevertheless, reliance on screening data alone for training may limit performance, as clinical datasets include a broader range of imaging heterogeneity and

patient variability. Incorporating CT scans from clinical populations into model training could therefore improve performance among patients with incidental nodules.

The aim of this multicentre study was to evaluate the discriminative performance and consistency across centres for two versions of a DL model for pulmonary nodule malignancy probability estimation of incidental nodules: a model trained exclusively on screening data and an updated model trained on both screening and clinical data.

Methods

Study design and population

The study protocol was reviewed by the local Medical Ethics Committee, which waived the need for informed consent due to the retrospective use of anonymized data. The study was conducted in accordance with the General Data Protection Regulation (GDPR) and the Declaration of Helsinki.

A retrospective dataset was collected from three Dutch centres, two University hospitals, and one non-university teaching hospital, using a case-control design. We aimed to collect 30 malignant and 60 benign nodules per centre. Since nodule size is strongly associated with lung cancer risk, size-based buckets were used to ensure the final dataset covered the complete size range for both benign and malignant nodules. The buckets were defined using the maximum diameter as 5–10 mm, 10–15 mm, and 15–30 mm, with each bucket containing 10 malignant and 20 benign nodules per centre. For selection, the size of the nodule as measured by the reporting radiologist was used.

Patients were eligible when they were 18 years or older, diagnosed with a pulmonary nodule at CT scan, and had at least a 2-year follow-up showing stability or resolution of the nodule or received a diagnostic work-up with a definite outcome. Cases were defined as patients with a stage I lung cancer and controls as patients without any cancer or cancer history. The patients were excluded if they were in a different diagnostic pathway in which pulmonary nodules were a potential manifestation of the underlying disease (i.e., patients with a history of cancer, MEN1 syndrome, benign nodular diseases (i.e., rheumatoid arthritis, common variable immunodeficiency, granulomatosis with polyangiitis)), had more than 10 nodules, had advanced fibrosis, or had a carcinoid cancer. Nodules were excluded when they had clear benign features (i.e., tree-in-bud configuration, calcified nodules, typical perifissural nodules, hamartomas containing macroscopic fatty areas) or were cystic nodules. CT scans were eligible when they contained the full thorax, had a reconstruction matrix of 512 × 512 or 1024 × 1024, and a slice thickness of 3 mm or less.

Data collection

Patients were selected by sampling data based on the reported diameter for each bucket until the buckets were filled. Cases were identified from pathology records. Controls were identified differently across centres. A Natural Language Processing algorithm was used to analyse the radiology reports for a mention of a pulmonary nodule [2]. The positive CT studies were manually checked. At centres where structured export of radiology reports was not supported, a keyword search in PACS was used to identify patients with a reported nodule.

Anonymized CT scans and patient data were retrieved. The first CT scan on which the incidental nodule was visible for each participant was collected. The patient data included age at the time of the CT scan, gender (male/female), reported diameter of the nodule, and lobe location. Age and gender were retrieved from the Electronic Medical Record System, and lobe location was extracted from the reports and verified by the local radiologist of the participating centre, without blinding to outcome. Nodules were located using the acquired information on lobe and size, and all other nodules were annotated to determine the nodule count. Uncertain cases were discussed with the local radiologist of the participating centre. Emphysema was not annotated, as it has high inter-reader variability, partly due to its binarisation in the Brock model, and its contribution to the Brock model has been shown to vary across cohorts [12].

Nodule size was measured in a standardised manner for all centres using semi-automated software, recording the maximum diameter in any imaging plane. These measurements could differ from those reported in the original radiology reports, which may have used average or maximum diameters measured in different planes. Discrepancies between the measurements could lead to nodules being reassigned to adjacent size buckets, and therefore, the final bucket count may differ from the target count.

Malignancy probability estimation models

Three models were compared: the clinically used Brock model (model 2a) and two in-house developed deep learning models. The Brock model, a logistic regression model containing patient and nodule characteristics, was used as a comparison model. All models calculate a malignancy probability per nodule between 0 and 100%.

The first deep learning model (DL1) is a model that is trained on data from a screening setting and validated on both screening data and clinical data [6, 7, 11]. The model was trained on 16,077 nodules (1249 malignant) from the National Lung Cancer Screening Trial. The model was

validated on screen-detected nodules from the Danish Lung Cancer Screening Trial, the Dutch Belgian lung cancer screening trial (NELSON), and the Multicentric Italian Lung Detection trial, and incidentally detected indeterminate nodules from a Dutch centre. This deep learning model uses a $5 \times 5 \times 5$ cm block around the nodule as input.

The second deep learning model (DL2) extended the screening-trained model by adding data from a clinical cohort and by redefining the benign nodule annotation method. For both data sources, the model was trained on histologically verified malignant nodules (screening: 1249 malignant; incidental: 127) and automatically detected benign nodules (screening: 82,283, incidental: 86,541). These benign nodules were collected using a previously published in-house nodule detection algorithm at an operating point of 1 false positive per scan, which was used on scans of patients without any cancer diagnosis [13]. The clinical data were obtained from a Dutch centre and included patients who had a chest CT scan between 2010 and 2017. The cancer diagnoses had been retrieved from the Netherlands Comprehensive Cancer Organisation. This DL model has not been validated previously.

Statistical analysis

Baseline characteristics of the patients and nodules were compared by malignancy status and centre. Continuous variables were compared using one-way analysis of variance (ANOVA; normally distributed) or Kruskal–Wallis Rank Sum test (non-normally distributed) and categorical variables using Fisher's exact test.

Nodule diameter was analysed in depth, both in the multicentre cohort and within each centre, because of its strong influence on dataset composition. Nodule diameter was summarised and compared between benign and malignant nodules. The distribution of nodule diameter was visualised using a distribution plot stratified by malignancy status. Associations between nodule diameter and model malignancy probability scores were assessed using Spearman's rank correlation and visualised using scatterplots. Spearman's rank correlation coefficients were interpreted following Mukaka [14].

The performance of the models was assessed using Receiver Operating Characteristic (ROC) curves and Area Under the ROC Curve (AUC) with their 95% confidence intervals. The 95% CIs for the ROC curves were determined using bootstrapping with 10,000 samples, and 95% CIs for the AUCs and comparison between the AUCs were assessed using the DeLong method.

Specificity was calculated at a target sensitivity equal to Brock's sensitivity at a 10% threshold as used in guidelines by the British Thoracic Society [3]. By using a fixed sensitivity, the models' specificities were compared while ensuring that the DL models would not miss more cancers than the Brock model.

Consistency of discrimination was assessed by repeating the ROC analysis and specificity analysis stratified by centre. A sensitivity analysis at the patient level was performed, selecting the highest risk score per model per patient, resulting in one nodule per patient.

Statistical analyses were performed using R (version 4.4.0). Descriptive statistics were generated using 'tableone' (version 0.13.2) and ROC analyses using 'pROC' (version 1.19.0) for the ROC analyses and sensitivity and specificity analyses. Statistical significance was defined as $p < 0.05$.

Results

A total of 269 nodules from 231 patients were collected, including 89 malignant nodules from 85 patients. Figure 1 presents the inclusion into the study dataset from data transfer to the final study protocol, which shows that the predefined target of 90 malignant nodules was not reached, due to exclusion of cases with ineligible CT scans or inability to localise the malignant lesion. Table 1 shows that malignant nodules were on average larger (maximum diameter 14 mm vs 12 mm) and more often located in the upper lobes than benign nodules (67% vs 47%). Additionally, nodule type was different per centre ($p < 0.01$): centre 2 only had solid nodules (100%) and centre 3 had a larger number of non-solid nodules compared to centre 1 (non-solid: 13% vs 5%; Table 2). CT characteristics of the full dataset and stratified by centre are presented in Table 3, showing that the CT characteristics differ per centre on study protocol, matrix, slice thickness, and manufacturer.

Malignant nodules were larger for centre 1 than benign nodules, specifically in the 15–30 mm size bucket. Nodule size per centre and malignancy status for each size bucket separately can be found in Table 4 for all centres combined and Supplementals 1 per centre. The distribution of nodule sizes and the correlation between the malignancy probability scores and nodule sizes are shown in Fig. 2. The diameter distribution showed a greater separation between malignant and benign nodules for centre 1 than the other centres. The Brock model is strongly to very strongly correlated to the nodule size for the centres (centre 1: 0.89; centre 2: 0.86; and centre 3: 0.91; all $p < 0.01$), while the DL models are strongly correlated for centre 1 (DL1: 0.73; DL2: 0.73; all $p < 0.01$), and weak to moderately correlated for centre 2 and 3 (centre 2—DL1: 0.40; DL2: 0.52; centre 3—DL1: 0.40; DL2: 0.59; all $p < 0.01$).

Both DL models achieved higher AUCs than the Brock model in the multicentre dataset (Fig. 3). DL1 achieved an AUC of 0.74 (95% CI: 0.68, 0.80) and DL2 achieved 0.72

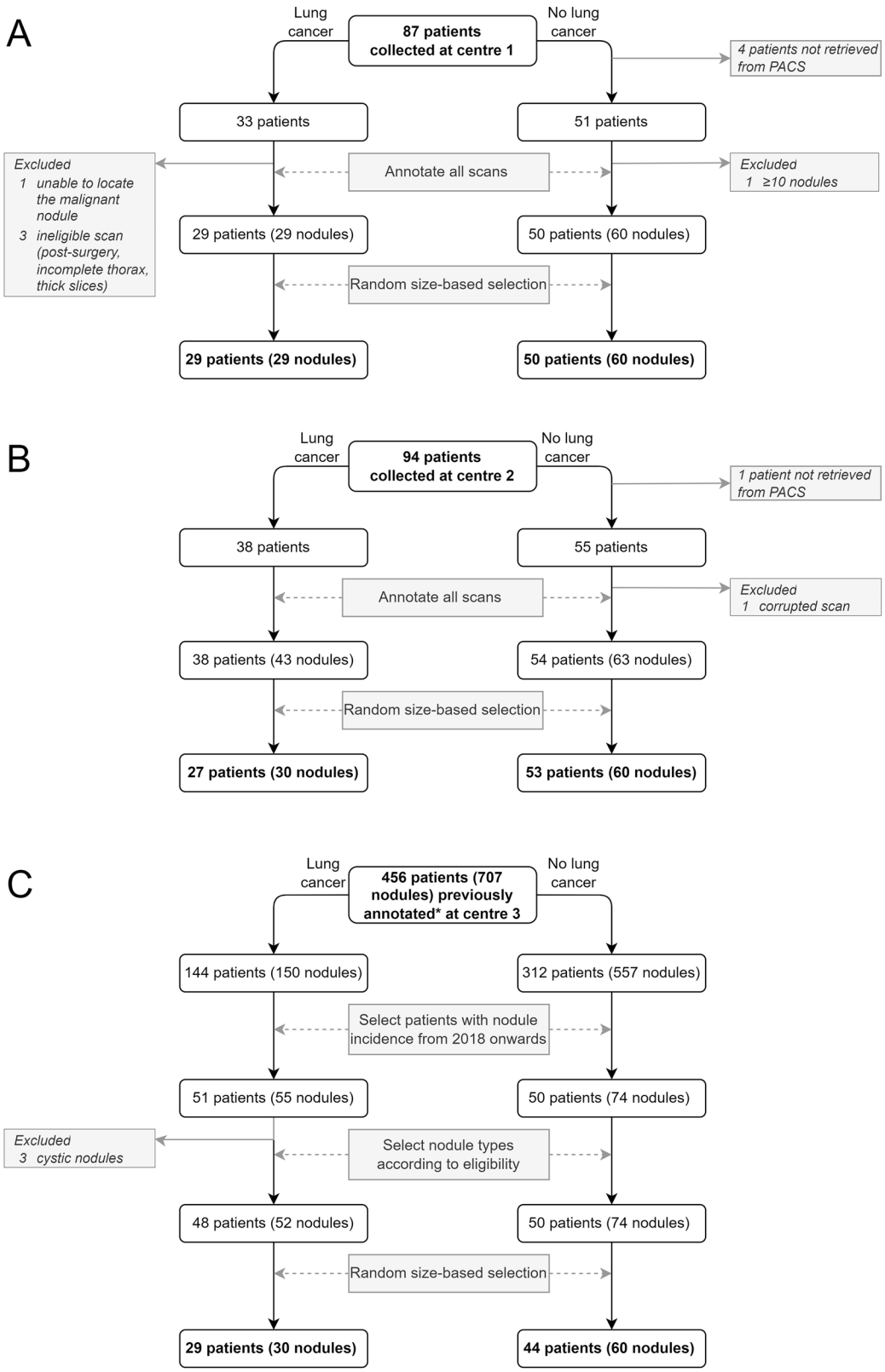


Fig. 1 Nodule selection from data transfer to the final study dataset for (**A**) centre 1, (**B**) centre 2, and (**C**) centre 3. * A nodule dataset was previously annotated according to the study protocol [11] and extended to include nodules measuring 5–30 mm. A subset of this dataset was reused in the current study

Table 1 Baseline characteristics for the full dataset and stratified by malignancy status, reported at the nodule level

Variable	Benign (180)	Malignant (89)	*p*-value	Overall (269)
Age (median [IQR])	68 [60, 74]	67 [62, 73]	0.88	67 [61, 74]
Sex (%)			0.09	
Female	84 (46.7)	52 (58.4)		136 (50.6)
Male	96 (53.3)	37 (41.6)		133 (49.4)
Maximum diameter (mm; median [IQR])	12.1 [9.3, 18.0]	14.1 [10.6, 20.4]	**0.03**	13.1 [9.6, 19.1]
Nodule type (%)			0.37	
Non-solid	14 (7.8)	3 (3.4)		17 (6.3)
Part-solid	7 (3.9)	4 (4.5)		11 (4.1)
Solid	159 (88.3)	82 (92.1)		241 (89.6)
Nodule count (median [IQR])	2 [1, 4]	2 [1, 3]	0.22	2 [1, 4]
Lobe location(%)			**0.02**	
Left upper	30 (16.7)	26 (29.2)		56 (20.8)
Left lower	30 (16.7)	12 (13.5)		42 (15.6)
Right upper	55 (30.6)	34 (38.2)		89 (33.1)
Right middle	18 (10.0)	5 (5.6)		23 (8.6)
Right lower	47 (26.1)	12 (13.5)		59 (21.9)

Data are the number of nodules with percentages in parentheses or median values with interquartile ranges in brackets. Significant *p*-values are presented in bold

(95% CI: 0.66, 0.79), both higher than the Brock model at 0.63 (95% CI: 0.56, 0.70; both $p < 0.01$). At the patient level, selecting the highest risk score per patient (85 malignant, 146 benign), DL1 achieved an AUC of 0.71 (95% CI: 0.65, 0.78), DL2 achieved 0.70 (95% CI: 0.63, 0.77), and Brock achieved 0.61 (95% CI: 0.54, 0.68; Supplementals 2).

In centres 2 and 3, both DL models achieved higher AUCs than the Brock model, whereas AUCs were similar across models in centre 1 (Fig. 4). In centre 1 (university hospital), AUCs were 0.76 for DL1 ($p = 0.60$ vs Brock), 0.75 for DL2 ($p = 0.76$ vs Brock), and 0.74 for the Brock model. In centre 2 (non-university teaching hospital), AUCs were 0.72 for DL1, 0.71 for DL2, and 0.55 for the Brock model, where the AUC of both DL models was higher than the Brock model (DL1: $p < 0.01$; DL2: $p = 0.01$). In centre 3 (university hospital), AUCs were 0.73 for DL1, 0.71 for DL2, and 0.59 for the Brock model (DL1: $p = 0.03$; DL2: $p = 0.04$).

At the sensitivity corresponding to the Brock model's sensitivity at a 10% threshold, both deep learning models achieved specificities that were equal to or higher than those of the Brock model across all datasets (Table 5). For the complete multicentre dataset, sensitivities of 77.5% for DL1 and DL2 were reached at thresholds of 11% and 7.8%, resulting in specificities of 60% for both DL models, compared to 44.4% for the Brock model. In centre 1, DL1 showed higher specificity than the Brock model (60% vs 48% at a target sensitivity of 86%), while DL2 was similar (48%). In centres 2 and 3, both deep learning models showed higher specificities than the Brock model. In centre 2 (target sensitivity 70%), specificities were 63% and 68% for DL1 and DL2 versus 37% for Brock. In centre 3 (target sensitivity 77%), specificities were 57% and 67% for DL1 and DL2 versus 48% for Brock.

Discussion

In this study, we evaluated the performance and consistency of two deep learning models for pulmonary nodule malignancy probability estimation in a cancer-enriched, size-stratified multicentre dataset. DL1 was trained on screening data alone and DL2 on screening and clinical data.

In the pooled multicentre dataset, both DL models outperformed the widely used Brock model in AUC and specificity at a fixed sensitivity. The performance found in the current study was lower than shown in other studies on DL models for incidental pulmonary nodule malignancy probability estimation, which is likely due to the composition of our dataset, in which nodule size had reduced discriminative value [8, 9]. Our size-based case-control collection strategy reduced the influence of diameter on performance, making risk stratification more challenging. While this makes our findings more relevant for daily clinical practice, as this is challenging data where false predictions are more common than in a dataset representative in size, it also limits the generalisability of our results. The thresholds used for this comparison are specific to this study population and should be interpreted as illustrative rather than as validated operating points for clinical use.

The DL models showed consistent discrimination across centres despite differences in scanners, CT acquisition parameters, and case mix. In two centres, both DL models outperformed the Brock model, whereas in one centre (centre 1), discrimination was similar across all models. Nonetheless, operating-point performance still differed, with DL1 achieving higher specificity at the Brock-matched sensitivity. This performance difference is likely due to the difference in nodule diameter distribution across the centres. In the centre with similar AUCs, malignant and benign nodules were more clearly separated by diameter, specifically in the 15–30 mm range. As diameter is a very strong factor in the Brock model, this clear difference in size between larger malignant and

Table 2 Baseline characteristics stratified by centre, reported at the nodule level

Variable	Centre 1 (89)	Centre 2 (90)	Centre 3 (90)	*p*-value
Age (median [IQR])	67 [60, 72]	68 [61, 76]	67 [60, 71]	0.41
Sex (%)				0.62
Female	48 (53.9)	42 (46.7)	46 (51.1)	
Male	41 (46.1)	48 (53.3)	44 (48.9)	
Maximum diameter (mm; median [IQR])	13.8 [10.1, 18.8]	13.1 [9.4, 19.5]	12.2 [9.3, 19.1]	0.69
Nodule type (%)				**< 0.01**
Non-solid	5 (5.6)	0 (0.0)	12 (13.3)	
Part-solid	6 (6.7)	0 (0.0)	5 (5.6)	
Solid	78 (87.6)	90 (100.0)	73 (81.1)	
Nodule count (median [IQR])	2 [1, 3]	2 [1, 4]	2 [1, 4]	0.14
Lobe location (%)				0.44
Left upper	20 (22.5)	14 (15.6)	22 (24.4)	
Left lower	17 (19.1)	14 (15.6)	11 (12.2)	
Right upper	22 (24.7)	35 (38.9)	32 (35.6)	
Right middle	9 (10.1)	9 (10.0)	5 (5.6)	
Right lower	21 (23.6)	18 (20.0)	20 (22.2)	

Data are the number of nodules with percentages in parentheses or median values with interquartile ranges (IQR) in brackets. Significant *p*-values are presented in bold

Table 3 CT characteristics, reported at the scan level

Variable	Centre 1 (78)	Centre 2 (80)	Centre 3 (74)	*p*-value	Overall (232)
Study protocol (%)				**0.01**	
Standard	57 (73.1)	64 (80.0)	40 (54.1)		161 (69.4)
HR	9 (11.5)	9 (11.2)	16 (21.6)		34 (14.7)
low dose	4 (5.1)	4 (5.0)	13 (17.6)		21 (9.1)
CTA	7 (9.0)	3 (3.8)	5 (6.8)		15 (6.5)
Other (cardiac)	1 (1.3)	0 (0.0)	0 (0.0)		1 (0.4)
Contrast usage (yes, %)	36 (46.2)	48 (60.0)	44 (59.5)	0.14	128 (55.2)
Matrix (%)				**< 0.01**	
1024 × 1024	0 (0.0)	4 (5.0)	20 (27.0)		24 (10.3)
512 × 512	78 (100.0)	76 (95.0)	54 (73.0)		208 (89.7)
Slice thickness (%)				**< 0.01**	
< 0.5	0 (0.0)	0 (0.0)	5 (6.8)		5 (2.2)
0.5	0 (0.0)	0 (0.0)	50 (67.6)		50 (21.6)
0.5–1.0	0 (0.0)	0 (0.0)	1 (1.4)		1 (0.4)
1.0	18 (23.1)	75 (93.8)	14 (18.9)		107 (46.1)
1.25–1.5	4 (5.1)	4 (5.0)	1 (1.4)		9 (3.9)
2.0–2.5	52 (66.7)	0 (0.0)	0 (0.0)		52 (22.4)
3.0	4 (5.1)	1 (1.2)	3 (4.1)		8 (3.4)
Manufacturer (%)				**< 0.01**	
Canon Medical Systems	0 (0.0)	32 (40.0)	3 (4.1)		35 (15.1)
GE Medical Systems	3 (3.8)	0 (0.0)	2 (2.7)		5 (2.2)
Philips	5 (6.4)	0 (0.0)	7 (9.5)		12 (5.2)
Siemens	70 (89.7)	48 (60.0)	3 (4.1)		121 (52.2)
Toshiba	0 (0.0)	0 (0.0)	59 (79.7)		59 (25.4)

Data are the number of CT studies with percentages in parentheses or median values with interquartile ranges in brackets. Significant *p*-values are presented in bold
HR high resolution, *CTA* computed tomography angiography

smaller benign nodules improves the Brock model's performance.

In this study, the model trained on screening and clinical data did not perform better than the model trained on screening data alone, both in the pooled dataset and for each centre separately. These results suggest that, in our setting, training on screening CT scans alone may be sufficient for malignancy probability estimation in incidental nodules. One possible explanation is that DL2 was trained on relatively few additional malignant nodules ($n = 127$) compared to the screening data ($n = 1249$), and that the number of malignant clinical nodules may have been insufficient to meaningfully improve model performance. Alternatively, potential population differences may not be included in the model input, as the models only look at a block of 50 × 50 × 50mm around the nodule; thus, they do not include information from parenchymal pathology elsewhere in the lungs that could be correlated with lung cancer risk. Lastly, pulmonary nodules demonstrate similar imaging characteristics across screening and clinical settings, even though the patient populations differ. As DL1 had already reached peak performance on data from a screening setting [15], it might have learned sufficient features to generalise to clinical nodules, such that the additional training data, whether clinical or screening, would have provided limited added value. Using screening data is practical, as screening scans are typically acquired with more standardised protocols, and large public datasets exist. Nevertheless, when screening-trained models are applied in clinical practice, differences in calibration and optimal decision thresholds should be considered.

While both DL models demonstrated consistent performance across centres, several aspects remain relevant

Table 4 Maximum nodule size stratified by centre and malignancy status

	Overall	Benign	Malignant	*p*-value	*n* (benign/malignant)
All sizes					
Centre 1	13.8 [10.1, 18.8]	11.7 [9.5, 16.3]	16.4 [13.6, 23.4]	**< 0.01**	60/29
Centre 2	13.1 [9.4, 19.5]	13.0 [9.4, 19.7]	13.1 [10.0, 18.4]	0.79	60/30
Centre 3	12.2 [9.3, 19.1]	12.1 [9.1, 17.5]	12.5 [9.6, 20.1]	1.00	60/30

Data are median values with interquartile ranges in brackets. Significant *p*-values are presented in bold

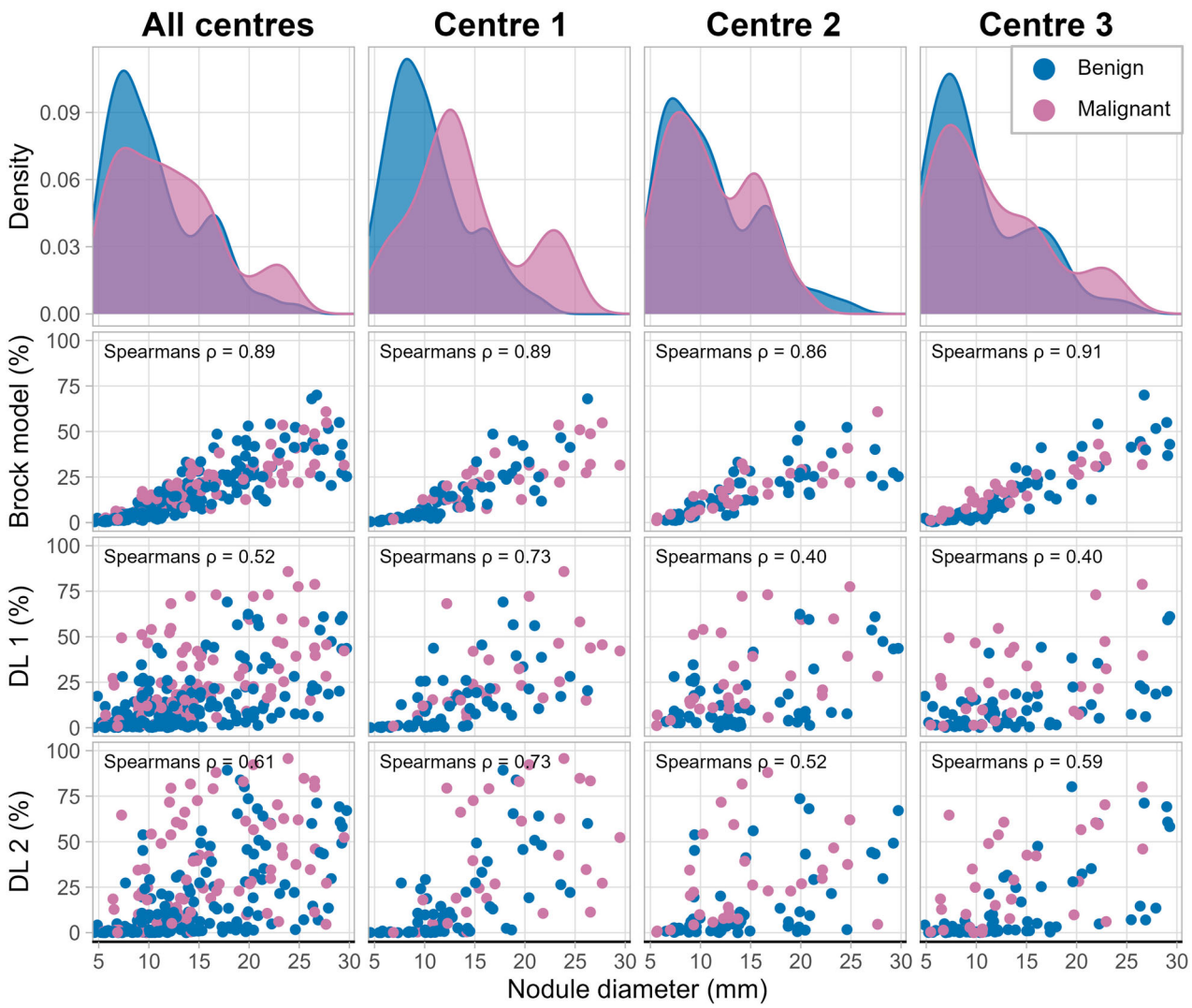


Fig. 2 Density distributions of nodule sizes and correlation between nodule size and model risk scores for the multicentre dataset and per centres. The density curves represent normalised distributions (area = 1) and do not reflect sample sizes. The Spearman's rho for all correlation curves is $p < 0.01$

for future development and clinical implementation. First, the models were based on a single CT scan, so they cannot take longitudinal information into account or use patient clinical characteristics. Incorporating such factors into the model may improve performance. Second, although DL models can achieve a high performance,

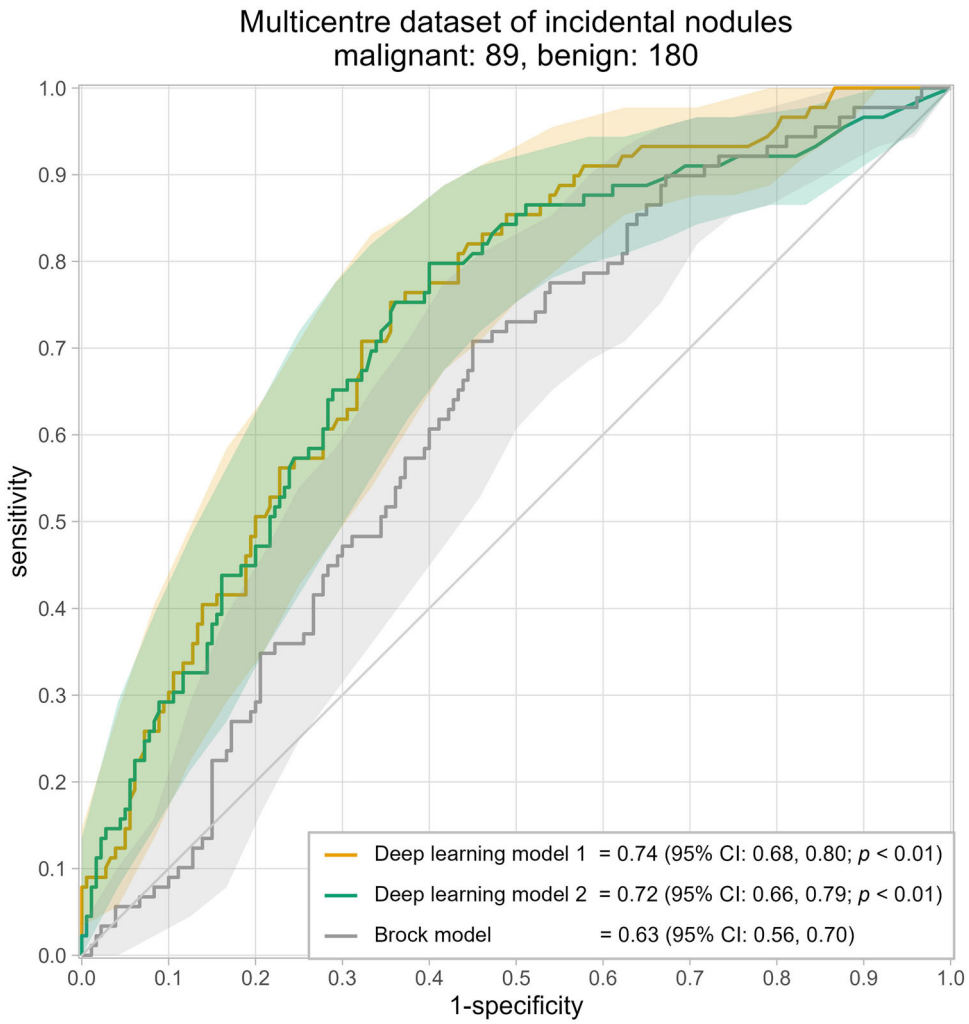


Fig. 3 ROC curves with 95% confidence bands and AUCs with 95% confidence intervals for the malignancy risk models. *p*-values correspond to comparisons with the Brock model. The comparison between the deep learning models (not shown) resulted in a *p*-value of 0.51

their limited interpretability may limit clinical use and adoption. Lastly, the models output a continuous score. In the current study, specificity was compared at matched sensitivity, suggesting that the DL models may have the potential to reduce false positives without missing more cancers than Brock. This is a first step toward how to use DL malignancy probability scores, but how to interpret the malignancy probability scores and translate them into patient management requires further research.

This study has several limitations that should be considered. First, the retrospective size-based case-control dataset does not reflect the prevalence of clinical practice; thus, we did not have the possibility to assess prevalence-based metrics, like positive predictive values or calibration metrics, or the generalisability to consecutive clinical populations. Second, all centres are in the same country, limiting the assessment of consistent performance across different healthcare systems and patient populations. However, substantial variation in scanner types, acquisition protocols, and case mix was present, allowing evaluation of consistency across centres and imaging conditions. In combination with a prior evaluation across several European screening cohorts [7], the results suggest consistency of the models within a selective Western European population. Third, this study only included stage I lung cancers, meaning that the validation set did not include patients with positive lymph nodes or invasion in the main bronchus, limiting generalisability to higher stages. However, as this model assesses the malignancy probability of the nodule itself, and the malignancy features are similar between these lower stages, this likely does not influence the performance of the models.

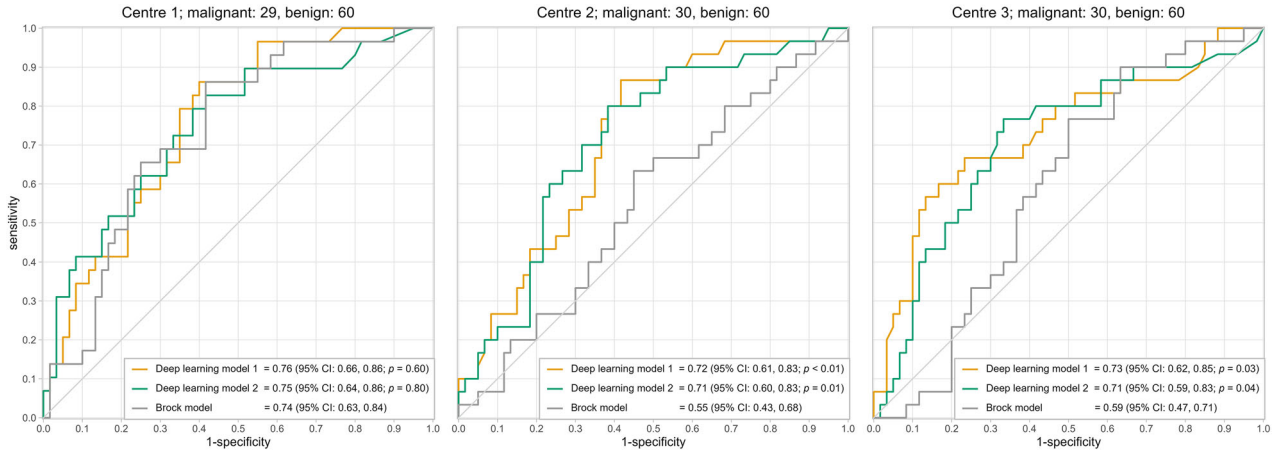


Fig. 4 ROC curves and AUCs with 95% confidence intervals stratified by centre. *p*-values correspond to comparisons with the Brock model. The comparison between the deep learning models (not shown) resulted in a *p*-value of 0.68 for centre 1, 0.91 for centre 2, and 0.70 for centre 3

Table 5 Thresholds and specificities for all models evaluated at a target sensitivity equal to or higher than the sensitivity achieved by the Brock model at its 10% risk threshold, for the multicentre dataset and stratified by centre

	Threshold (%)	Sensitivity (target; %)	Specificity (%)
Multicentre dataset			
Brock model	10.0	77.5	44.4
Deep learning model 1	11.0	77.5	60.0
Deep learning model 2	7.8	77.5	60.0
Centre 1			
Brock model	10.0	86.2	48.3
Deep learning model 1	12.0	86.2	60.0
Deep learning model 2	5.0	86.2	48.3
Centre 2			
Brock model	10.0	70.0	36.7
Deep learning model 1	13.6	70.0	63.3
Deep learning model 2	9.8	70.0	68.3
Centre 3			
Brock model	10.0	76.7	48.3
Deep learning model 1	9.1	76.7	56.7
Deep learning model 2	7.4	76.7	66.7

By evaluating two deep learning models across multiple centres, this study provides evidence of consistent model performance across centres with varying CT and patient populations within one country, providing further evidence of the potential of DL models to classify pulmonary nodules in varying circumstances. Future work should look at how to integrate such DL malignancy probability scores into management protocols, the interaction between these models and clinicians, and the acceptance of such models, as within a clinical setting, these DL models could assist the clinician. Additionally, prospective studies in representative clinical populations are needed to assess the real-world impact and potential usefulness of these models in routine healthcare.

Abbreviations

AUC	Area under the ROC curve
CI	Confidence interval
CT	Computed tomography
DL	Deep learning
IQR	Interquartile range
ROC	Receiver operating characteristic

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1007/s00330-026-12787-y.

Acknowledgements

The authors would like to thank the registration team of the Netherlands Comprehensive Cancer Organization (IKNL) for data collection from the National Cancer Registry, as well as IKNL staff for scientific advice. We thank Ward Hendrix and Kiran Vaidhya Venkadesh for their technical expertise and valuable advice regarding the application of the algorithms used in this study. Furthermore, we thank Veerle Souverein and Henk de Marie for their expertise and work regarding data stewardship and technical support. The authors used ChatGPT (versions 5.3 and 5.4, OpenAI) for language editing and style refinement of the manuscript.

Funding

This study has received funding from the Dutch Cancer Society (KWF Kankerbestrijding, project number 14113).

Compliance with ethical standards

Guarantor

The scientific guarantor of this publication is Colin Jacobs.

Conflict of interest

The authors of this manuscript declare relationships with the following companies: Cornelia Schaefer-Prokop received speaker fees from Canon and Boehringer and royalties from Thieme and Elsevier unrelated to nodule research; Colin Jacobs receives research grants and royalties to the host institution from MeVis Medical Solutions, Bremen, Germany, and payment for lectures from Canon Medical Systems and Johnson & Johnson. Colin Jacobs is a collaborator in a public-private research project where Radboudumc collaborates with Philips Medical Systems (Best, the Netherlands) and another public-private research project where Radboudumc collaborates with Siemens Healthineers (Erlangen, Germany). He reports no other relationships that are related to the subject matter of the article. The other authors of this manuscript declare no relationship with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry

No complex statistical methods were necessary for this paper.

Informed consent

Written informed consent was waived by the Institutional Review Board.

Ethical approval

Institutional Review Board approval was obtained.

Study subjects or cohorts overlap

Some study subjects or cohorts have been previously reported in Dinnessen et al [11]. This concerns 12 nodules in 12 patients from centre 3, selected from the larger dataset according to the eligibility criteria defined in the current study.

Methodology

- Retrospective
- Case-control study, diagnostic study
- Multicenter study

Author details

[1]Department of Medical Imaging, Radboud University Medical Center, Nijmegen, The Netherlands. [2]Department of Radiology and Nuclear Medicine, Maastricht University Medical Center, Maastricht University, Maastricht, The Netherlands. [3]GROW, School of Oncology and Reproduction, Maastricht University, Maastricht, The Netherlands. [4]Department of Pathology, Maastricht University Medical Centre, Maastricht, The Netherlands. [5]Department of Cardiology, CARIM, Maastricht University Medical Centre, Maastricht, The Netherlands. [6]Department of Radiology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands. [7]Department of Radiology, Meander Medical Center, Amersfoort, The Netherlands.

Received: 7 March 2026 Revised: 16 June 2026 Accepted: 9 July 2026
Published online: 08 August 2026

References

1. Gould MK, Tang T, Liu ILA et al (2015) Recent trends in the identification of incidental pulmonary nodules. Am J Respir Crit Care Med. https://doi.org/10.1164/rccm.201505-0990oc
2. Hendrix W, Rutten M, Hendrix N et al (2023) Trends in the incidence of pulmonary nodules in chest computed tomography: 10-year results from two Dutch hospitals. Eur Radiol. https://doi.org/10.1007/s00330-023-09826-3
3. Callister MEJ, Baldwin DR, Akram AR et al (2015) British Thoracic Society guidelines for the investigation and management of pulmonary nodules. Thorax. https://doi.org/10.1136/thoraxjnl-2015-207168
4. MacMahon H, Naidich DP, Goo JM et al (2017) Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. Radiology. https://doi.org/10.1148/radiol.2017161659
5. McWilliams A, Tammemagi MC, Mayo JR et al (2013) Probability of cancer in pulmonary nodules detected on first screening CT. N Engl J Med. https://doi.org/10.1056/nejmoa1214726
6. Venkadesh KV, Setio AAA, Schreuder A et al (2021) Deep learning for malignancy risk estimation of pulmonary nodules detected at low-dose screening CT. Radiology. https://doi.org/10.1148/radiol.2021204433
7. Antonissen N, Venkadesh KV, Dinnessen R et al (2025) External test of a deep learning algorithm for pulmonary nodule malignancy risk stratification using European screening data. Radiology. https://doi.org/10.1148/radiol.250874
8. Massion PP, Antic S, Ather S et al (2020) Assessing the accuracy of a deep learning method to risk stratify indeterminate pulmonary nodules. Am J Respir Crit Care Med. https://doi.org/10.1164/rccm.201903-0505oc
9. Baldwin DR, Gustafson J, Pickup L et al (2020) External validation of a convolutional neural network artificial intelligence tool to predict malignancy in pulmonary nodules. Thorax. https://doi.org/10.1136/thoraxjnl-2019-214104
10. Wulaningsih W, Villamaria C, Akram A, Benemile J, Croce F, Watkins J (2024) Deep learning models for predicting malignancy risk in CT-detected pulmonary nodules: a systematic review and meta-analysis. Lung. https://doi.org/10.1007/s00408-024-00706-1
11. Dinnessen R, Peeters D, Antonissen N et al (2025) Performance of a screening-trained DL model for pulmonary nodule malignancy estimation of incidental clinical nodules. Eur Radiol. https://doi.org/10.1007/S00330-025-11829-1
12. Winter A, Aberle DR, Hsu W (2019) External validation and recalibration of the Brock model to predict probability of cancer in pulmonary nodules using NLST data. Thorax. https://doi.org/10.1136/thoraxjnl-2018-212413
13. Hendrix W, Hendrix N, Scholten ET et al (2023) Deep learning for the detection of benign and malignant pulmonary nodules in non-screening chest CT scans. Commun Med. https://doi.org/10.1038/s43856-023-00388-5
14. Mukaka MM (2012) A guide to appropriate use of correlation coefficient in medical research. Malawi Med J 24:69–71
15. Obreja B, Bosma J, Venkadesh KV, Saghir Z, Prokop M, Jacobs C (2025) Characterizing the impact of training data on generalizability: application in deep learning to estimate lung nodule malignancy risk. Radiol Artif Intell. https://doi.org/10.1148/ryai240636

Publisher's Note