Validation of the BRODERS classifier (Benign *versus* aggRessive nODule Evaluation using Radiomic Stratification), a novel HRCT-based radiomic classifier for indeterminate pulmonary nodules

Fabien Maldonado[1,7], Cyril Varghese[2,7], Srinivasan Rajagopalan[3,7], Fenghai Duan[4], Aneri B. Balar[1], Dhairya A. Lakhani[1], Sanja L. Antic[1], Pierre P. Massion[1,5], Tucker F. Johnson[6], Ronald A. Karwoski[3], Richard A. Robb[3,†], Brian J. Bartholmai[6] and Tobias Peikert[2]

@ERSpublications
This study reports the independent external validation of the Mayo Clinic BRODERS (Benign *versus* aggRessive nODule Evaluation using Radiomic Stratification) classifier, radiomics model, for the classification into benign and malignant lung nodules. https://bit.ly/2GNUPSL

Cite this article as: Maldonado F, Varghese C, Rajagopalan S, *et al.* Validation of the BRODERS classifier (Benign *versus* aggRessive nODule Evaluation using Radiomic Stratification), a novel HRCT-based radiomic classifier for indeterminate pulmonary nodules. *Eur Respir J* 2021; 57: 2002485 [https://doi.org/10.1183/13993003.02485-2020].

ABSTRACT

Introduction: Implementation of low-dose chest computed tomography (CT) lung cancer screening and the ever-increasing use of cross-sectional imaging are resulting in the identification of many screen- and incidentally detected indeterminate pulmonary nodules. While the management of nodules with low or high pre-test probability of malignancy is relatively straightforward, those with intermediate pre-test probability commonly require advanced imaging or biopsy. Noninvasive risk stratification tools are highly desirable.

Methods: We previously developed the BRODERS classifier (Benign *versus* aggRessive nODule Evaluation using Radiomic Stratification), a conventional predictive radiomic model based on eight imaging features capturing nodule location, shape, size, texture and surface characteristics. Herein we report its external validation using a dataset of incidentally identified lung nodules (Vanderbilt University Lung Nodule Registry) in comparison to the Brock model. Area under the curve (AUC), as well as sensitivity, specificity, negative and positive predictive values were calculated.

Results: For the entire Vanderbilt validation set (n=170, 54% malignant), the AUC was 0.87 (95% CI 0.81–0.92) for the Brock model and 0.90 (95% CI 0.85–0.94) for the BRODERS model. Using the optimal cut-off determined by Youden's index, the sensitivity was 92.3%, the specificity was 62.0%, the positive (PPV) and negative predictive values (NPV) were 73.7% and 87.5%, respectively. For nodules with intermediate pre-test probability of malignancy, Brock score of 5–65% (n=97), the sensitivity and specificity were 94% and 46%, respectively, the PPV was 78.4% and the NPV was 79.2%.

Conclusions: The BRODERS radiomic predictive model performs well on an independent dataset and may facilitate the management of indeterminate pulmonary nodules.

This article has supplementary material available from erj.ersjournals.com

Received: 26 June 2020 | Accepted: 1 Oct 2020

Introduction

Lung cancer remains the deadliest malignancy in the United States (US) and worldwide [1]. While lung cancer 5-year survival has improved over the past decade, >50% of all lung cancer cases continue to be diagnosed at advanced stages. This is at least in part attributable to the lack of widespread implementation of lung cancer screening [2]. Several recent large lung cancer screening studies, the National Lung Screening Trial (NLST) in the US, the European Multicentric Italian Lung Detection (MILD) study and Nederlands-Leuvens Longkanker Screenings ONderzoek (NELSON) trial have demonstrated that low-dose computed tomography (LDCT) screening can reduce lung cancer mortality in high-risk patients [3–5]. However, even in the US, despite endorsement by the Center for Medicare & Medicaid Services and the United States Preventive Services Task Force, the clinical implementation and acceptance of LDCT screening remains suboptimal [6]. One of the main clinical challenges remains the high rate of false positive results, as almost all detected pulmonary nodules are benign. Other obstacles include the diagnosis of indolent lung cancer (overdiagnosis), uncertainty about optimal patient selection, screening intervals and duration, as well as concerns about cost-effectiveness [7]. While high false positive rates (96% of all screen-detected nodules ⩽4 mm were false positives in the NLST) can be improved by the application of Lung Imaging Reporting and Data System (Lung-RADS) criteria and the updated Fleischner Society nodule management guidelines for screen- and incidentally detected indeterminate pulmonary nodules (IPNs), these are associated with a decreased sensitivity [8, 9]. For example, while Lung-RADS reduces the false positive rate to 5.3%, it also reduces sensitivity by ∼10% [10].

In addition to screen-detected IPNs, incidentally discovered IPNs are on the rise. This development is due to increased utilisation of diagnostic cross-sectional chest imaging and the more widespread availability of advanced high-resolution computed tomography. Approximately 12 million chest CT studies are performed annually in the US and based on data from 2006 to 2012, it has been estimated that ∼1.5 million adult Americans will be diagnosed with a pulmonary nodule annually [11]. The magnitude of the clinical challenges of noninvasively classifying screen- and incidentally detected IPNs highlights the urgent need for improved diagnostic tools.

Radiomics is a rapidly emerging field. It involves quantitative image analysis to objectively and reproducibly analyse imaging data [12] to identify predictive and descriptive radiological features not otherwise evident to a human observer that may correlate with the biological behaviour of the lesion analysed. While radiomic approaches were conceived as early as the 1950s [13], the increased availability of inexpensive and powerful computing hardware [14] has generated considerable interest in lung nodule analysis in the past decade [15]. However, there is great variability in image acquisition, feature extraction methodology and statistical modelling across the many radiomic models described in literature, and, so far, no radiomic model has been integrated into routine clinical practice [16]. Furthermore, it is unclear whether conventional radiomic approaches, whereby expert-selected radiomic variables are used to derive a multivariate prediction model *via* regression analysis, unsupervised deep-learning approaches or a combination of these two methodologies will ultimately prove more clinically useful.

Many promising radiomic models for IPNs have been proposed, but few have been successfully validated on independent, external cohorts either due to the lack of access to readily available, well-curated datasets, or because of the risk of overfitting that particularly pervades radiomic models. In addition, CT datasets are typically heterogeneous, characterised by substantial variability in scanner technology, image acquisition and reconstruction [17]. Thus, it is unclear whether such models outperform validated simpler and readily accessible clinical prediction models [18].

Using a training set of 726 IPNs from the NLST database, we previously developed and internally validated the BRODERS classifier (Benign *versus* aggRessive nODule Evaluation using Radiomic Stratification), a radiomic classifier that effectively distinguishes benign from malignant nodules [19]. Herein we report the successful validation of this classifier in an independent dataset of incidentally detected IPNs from a tertiary referral centre. In addition, we compare the performance of our model to the performance of an established clinical prediction model routinely used in clinical practice [15].

Affiliations: [1]Division of Allergy, Pulmonary and Critical Care Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [2]Division of Pulmonary and Critical Care Medicine, Mayo Clinic, Rochester, MN, USA. [3]Dept of Physiology and Biomechanical Engineering, Mayo Clinic, Rochester, MN, USA. [4]Pulmonary Section, Medical Service, Tennessee Valley Healthcare Systems, Nashville Campus, Nashville, TN, USA. [5]Dept of Biostatistics and Center for Statistical Sciences, Brown University School of Public Health, Providence, RI, USA. [6]Dept of Radiology, Mayo Clinic, Rochester, MN, USA. [7]These authors contributed equally to this work.

Correspondence: Tobias Peikert, Division of Pulmonary and Critical Care Medicine, 200 First Street SW, Rochester, MN 55905, USA. E-mail: peikert.tobias@mayo.edu

Methods

Classifier development

The development of our radiomic classifier and Computer-Aided Lung Informatics for Pathology Evaluation and Rating (CALIPER) and Computer-Aided Nodule Assessment and Risk Yield (CANARY) used to analyse the lung and nodule texture has been described and validated previously [19–22]. Briefly, 726 patients with screen-detected IPNs with largest diameter ranging from 7 to 30 mm enrolled into the LDCT arm of the NLST were included in the training set. The first LDCT screening scans to identify the lung nodule were included in the radiomic analysis. A semi-automated region-growing approach was used for nodule segmentation (ANALYSE Biomedical Imaging Resource; Mayo Clinic, Rochester, MN, USA). Manual editing was performed to exclude adjacent intrathoracic structures such as blood vessels and pleura. Receiver operative curves (ROC) were calculated for each of 57 pre-selected radiological features organised in the following broad categories characterising the nodule: spatial location, size, shape, radiodensity, nodule texture, texture of lung tissue surrounding the nodule and nodule surface characteristics. Statistical significance of the area under the curve (AUC) was calculated and adjusted for multiple comparisons using Bonferroni correction. Spearman rank correlations between all pairs of variables were calculated and displayed in a heat map. Multivariate analysis was performed using the least absolute shrinkage and selection operator (LASSO) to enhance the prediction accuracy. LASSO was run 1000 times and variables that were selected by ⩾50% of the runs were included in the final multivariate model. To correct for overfitting bootstrapping was applied to calculate the optimism-corrected AUC for the final model of benign *versus* malignancy prediction which was found to be 0.939 [19]. We identified the optimal cut-off at 0.478 with sensitivity 0.904 and specificity 0.855 using Youden's index.

External validation database

The study was approved or exempted by the institutional review boards of the two participating institutions (Vanderbilt University (IRB# 151 500) and Mayo Clinic (IRB# 15–002674)). The validation dataset included consecutive patients with incidentally identified IPNs enrolled into the Vanderbilt University pulmonary nodule registry. The Digital Imaging and Communications in Medicine images of the CT scans were transferred to the Mayo Clinic (Rochester, MN, USA) for radiomic analysis. All the investigators at Mayo Clinic were blinded to the clinical information available for each patient, including baseline patient information (demographics, smoking status, prior cancer history), pathological information (benign *versus* malignant, histopathological type, staging) and long-term outcomes (death, alive with or without evidence of disease). Semi-automated segmentation was performed by the ANALYSE software described earlier. The BRODERS radiomics classifier was then used to predict the probability of malignancy of the included nodules.

Comparison of the BRODERS classifier with Brock model

The probability of malignancy calculated for each nodule using the Brock model, a well validated nodule malignancy probability calculator widely used in clinical practice [20], was compared with the BRODERS classifier in both the subset of our previously published screen-detected nodule NLST dataset for which the variables to calculate Brock model were available and the incidentally detected nodule Vanderbilt dataset (supplementary figure S1). For these cases, Brock model prediction was compared with the BRODERS classifier using ROC analysis. In addition, comparative ROC analysis was performed on subsets of nodules classified based on pre-test malignancy probability as follows. Low probability: Brock score <5%, NLST n=257, Vanderbilt n=42; intermediate probability: Brock score ⩾5% but <65%, NLST n=416, Vanderbilt n=126; and high probability: Brock score ⩾65%, NLST n=12, Vanderbilt n=2.

Statistical analyses

MedCalc Statistical Software version 19.0.7 (MedCalc Software, Ostend, Belgium; www.medcalc.org; 2019) was used for statistical analysis. Comparison of ROC curves was done using the nonparametric method described by DeLong *et al.* [21] for AUC calculation, exact binomial confidence intervals were used.

Results

The baseline characteristics of the patients in the subset of our NLST cohort and the Vanderbilt cohort are shown in table 1. The Vanderbilt external validation set included 170 consecutive patients with incidentally identified IPNs (diameter 7–30 mm) enrolled into the Vanderbilt University pulmonary nodule registry. Although the distribution of malignant *versus* benign nodules is similar in both cohorts, many of the other baseline characteristics including smoking status, nodule size and spiculation is different between the two groups, as would be expected in comparing a screen-detected nodule cohort with an incidentally discovered nodule cohort. In the Vanderbilt University cohort, the mean diameter of the malignant nodules was larger than the benign nodules, 10.3 mm (CI 9.4–11.3 mm) *versus* 17.5 mm

TABLE 1 Baseline characteristics of the two cohorts described in the study

	NLST	Vanderbilt
Subjects	685	170
Age years	63±5.3	66±7.6
Sex		
Male	392 (57.2)	113 (66.5)
Female	293 (42.8)	57 (33.5)
Race		
Caucasian	632 (92.3)	152 (89.4)
Black, Asian, other	53 (7.7)	18 (10.6)
Smoking		
Current	362 (52.8)	108 (64)
Former	327 (47.2)	58 (34)
Never	0	4 (2)
Smoking pack-years	61±27.1	57±34.2
Mode of nodule detection	Screening	Incidental
Nodule diagnosis		
Benign	313 (45.7)	79 (46)
Malignant	372 (54.3)	91 (54)
Nodule size mm	12.2±6.5	14.6±6.9
Spiculation	199 (29.1)	20 (11.8)

Data are presented as n, mean±SD or n (%). NLST: National Lung Screening Trial.

(CI 16.2–17.8 mm), respectively ($p<0.001$) (supplementary figure S2). Supplementary figures S3 and S4 show high-resolution axial scout images formatted into truth tables comparing the ground-truth histology with radiomic predictions using BRODERS. Confusion tables comparing the clinical/histological ground truth to the Brock model and the BRODERS classifier for the NLST and Vanderbilt datasets are shown in tables 2 and 3, respectively. The distribution of malignancies and their BRODERS classifications at various Brock score categories are displayed in supplementary tables S1 and S2.

Using the optimal cut-off of 0.478 identified *via* Youden's index, the sensitivity and specificity of the BRODERS classifier were 88.7% and 86.2%, respectively, in the NLST screen-detected nodule cohort (n=685). For nodules with intermediate pre-test probability of malignancy (5–65%) by the Brock model (n=416) sensitivity was 91.9% and specificity was 71.6% using the same cut-off.

For the entire Vanderbilt incidental nodule dataset (n=170), sensitivity was 92.3%, specificity was 62.0%, the positive predictive value (PPV) was 73.7% and the negative predictive value (NPV) was 87.5%. For nodules with intermediate pre-test probability of malignancy by the Brock model (n=97), sensitivity was 94%, specificity was 46%, PPV was 78.4% and NPV was 79.2%. The performance of the BRODERS classifier across different Brock-probability cut-offs for the intermediate lung nodules are shown in supplementary tables S3 and S4.

TABLE 2 Truth tables comparing histology *versus* Benign *versus* aggRessive nODule Evaluation using Radiomic Stratification (BRODERS) classifier *versus* Brock model probability categories in the National Lung Screening Trial (NLST) cohort

Brock model probability of malignancy	Subjects	Clinical/ histological classification		BRODERS benign	BRODERS malignant
Low <5%	257	Benign	204	192	12
		Malignant	53	17	36
Intermediate 5–<65%	416	Benign	109	78	31
		Malignant	307	25	282
High ⩾65%	12	Benign	0	0	0
		Malignant	12	0	12

Data are presented as n.

TABLE 3 Truth tables comparing histology *versus* Benign *versus* aggRessive nODule Evaluation using Radiomic Stratification (BRODERS) classifier *versus* Brock model probability categories in the Vanderbilt cohort

Brock model probability of malignancy	Subjects	Clinical/histological classification		BRODERS benign	BRODERS malignant
Low <5%	42	Benign	38	30	8
		Malignant	4	2	2
Intermediate 5–<65%	126	Benign	41	19	22
		Malignant	85	5	80
High ⩾65%	2	Benign	0	0	0
		Malignant	2	0	2

Data are presented as n.

The direct correlation between the Brock model and the BRODERS classifier for the Vanderbilt University cohort are shown in supplementary figure S5. Figures 1 and 2 show the ROC comparing Brock model *versus* BRODERS for the entire NLST and Vanderbilt cohorts, and subsets of the cohort classified as low and intermediate pre-test malignancy risk. In both cohorts the AUC are significantly greater for the BRODERS model compared to the Brock model at all pre-test malignancy probabilities ($p<0.001$). The difference is most pronounced in the intermediate pre-test malignancy risk group. The benign resection rates based on the hypothetical application of the BRODERS classifier to the NLST and the Vanderbilt datasets are 12% and 26%, respectively, for the entire cohorts and 10% and 22%, respectively, for the Brock model intermediate probability nodules (5–65%).

Discussion

In this study, we validated the BRODERS classifier on an independent dataset of incidentally identified lung nodules, and report excellent diagnostic test performance, with the potential to clarify the clinical significance of IPNs, using a novel radiomic model applicable to existing CT images.

Several notable studies have described the use of radiomics for pulmonary nodule characterisation. Some of them used large datasets like the NLST [22–24] or the Lung Image Database Consortium image collection [25], while others used institution-specific datasets as their training sets [26]. While some of these studies include validation cohorts, the majority of them are either internal validation sets or represent a subset of the cohort used for training (split sample validation), and thus do not truly provide external validation [15]. External validation in truly independent datasets is critical for radiomic models, which typically explore large numbers of candidate predictive variables in regression analyses with limited datasets. This introduces a substantial risk of overfitting, which is compounded when deep-learning methods are used. In addition, it is important to take into consideration the potential differences between screen- and incidentally identified lung nodules, as models derived from screening cohorts may perform

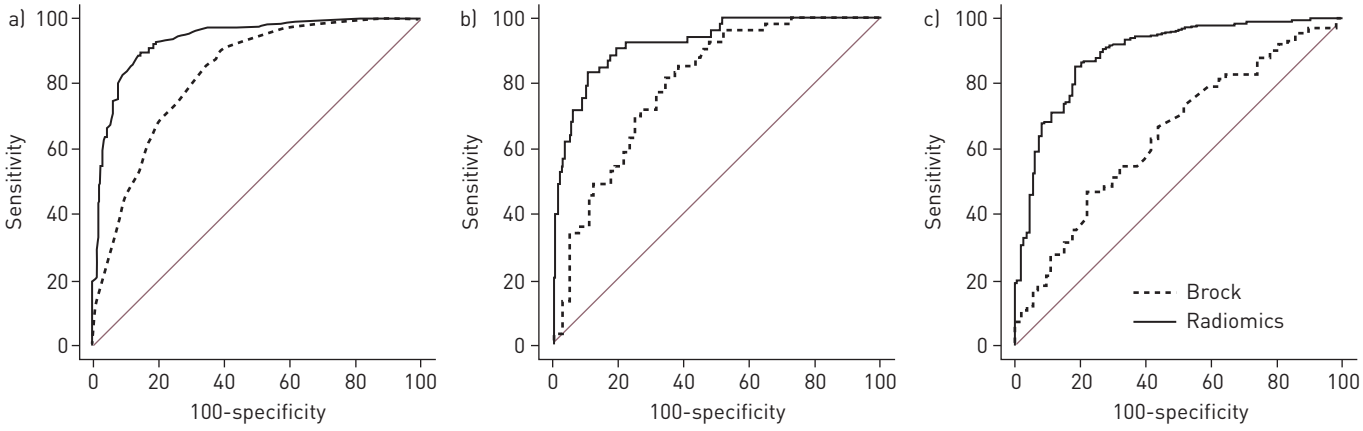

FIGURE 1 Receiver operating characteristic (ROC) for the National Lung Screening Trial cohort comparing the Brock and radiomics classifications. a) Entire cohort: area under the curve (AUC) Brock 0.833 (95% CI 0.803–0.860), AUC radiomics 0.939 (0.918–0.955); b) low-risk (Brock score <5%) group: AUC Brock 0.795 (0.74–0.842), AUC radiomics 0.925 (0.886–0.954); c) intermediate-risk (5% Brock score <65%) group: AUC Brock 0.648 (0.599–0.694), AUC radiomics 0.893 (0.859–0.922).

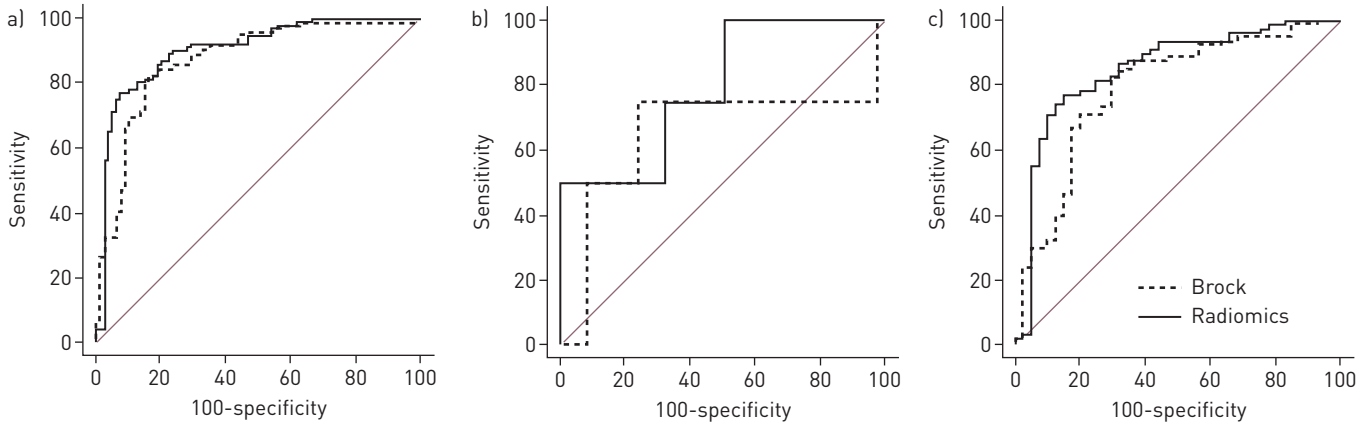

FIGURE 2 Receiver operating characteristic (ROC) for the Vanderbilt cohort comparing the Brock and radiomics classifications. a) Entire cohort: area under the curve (AUC) Brock 0.872 (95% CI 0.812–0.918), AUC radiomics 0.904 (0.849–0.943); b) low-risk (Brock score <5%) group: AUC Brock 0.658 (0.496–0.797), AUC radiomics: 0.796 (0.644–0.904); c) intermediate-risk (5% Brock score <65%) group: AUC Brock 0.798 (0.717–0.864), AUC radiomics 0.856 (0.782–0.912).

well in similar cohorts, but may not be generalisable to all lung nodules. In 2019, Ardila *et al.* [27] developed a deep-learning radiomic tool using the NLST dataset as a training cohort, and validated it on an independent cohort from an academic institution with comparable diagnostic test performance. However, the validation dataset was also a screening cohort, which may limit the model's external validity, and specifically its applicability to incidentally discovered IPNs. More recently, Massion *et al.* [28] reported the development of their deep-learning based Lung Cancer Prediction Convolutional Neural Network (LCP-CNN) model. The reported AUCs of 0.92, 0.84 and 0.92 in the NLST (training set, screen-detected), a Vanderbilt University validation set and an Oxford University validation set (incidentally identified nodules), respectively, are comparable to the performance of our conventional radiomic classifier and outperformed the clinical Mayo Lung Nodule prediction model. Ultimately, the clinical utility of the LCP-CNN will need to be clarified with prospective validation.

While deep-learning radiomic models and machine learning have received disproportionate attention in recent years, they have significant limitations. These include the need for very large training sets [29], redundancy of features that are thought to be significant [30], overfitting [31] and the inability for external research groups to replicate results [32]. Deep-learning models are often compared to a "black box", in that predictive variables are unknown, limiting reproducibility and transparency, may have no direct correlation with underlying relevant biological features, or may be heavily weighted by features easily identified during clinical CT evaluation, such as nodule size. Conversely, in our conventional radiomic model, variables with known relevance to nodule characterisation were selected for their direct relevance to predictive biological features, such as nodule texture, surface characteristics and location.

It is important to recognise that due to a variety of factors, including strict inclusion criteria and healthy volunteer effect, subjects enrolled in screening studies tend to be substantially different to patients presenting at lung nodule clinics or even patients eligible for lung cancer screening [8]. In this study we validated our model, the BRODERS classifier, which was trained using the NLST screening dataset [19], on an external dataset of consecutively identified incidentally detected lung nodules collected at the Vanderbilt lung nodule clinic. The excellent performance of our model supports its generalisability to other populations of patients with IPNs.

A variety of clinical prediction models have been proposed to assist clinicians in lung nodule management using readily available data [18]. These models are relatively easy to use and while some may be better suited for selected populations, comparative studies suggest that the Brock model may perform better than the others [33, 34]. In addition, a study by van Riel *et al.* [35] suggested that the Brock model may be preferable to both Lung-RADS and the National Comprehensive Cancer Network guidelines to classify nodules. The BRODERS classifier outperformed the Brock model in both the NLST and Vanderbilt cohorts at all pre-test malignancy risk levels. Notably, our model had high NPV at low pre-test malignancy risk and good PPV and NPV at intermediate pre-test malignancy risk. Hence, applying the BRODERS radiomic model to screen- or incidentally identified lung nodules may effectively reclassify nodules with intermediate probability of malignancy into high or low post-test probability, obviating the need for advanced imaging, invasive biopsy or benign surgical resections. For example, using the calculated sensitivity and specificity for the nodules with intermediate pre-test probability of malignancy in the

Vanderbilt cohort, a nodule with a 50% pre-test probability could be reclassified as low post-test probability after negative radiomic analysis (7.7%), or high post-test probability (74.7%), which may alter the clinical management.

The clinical implementation of the BRODERS classifier should be highly feasible. Our semi-automated region-growing approach nodule segmentation approach (ANALYSE) is fast (1–5 min for most nodules), and does not require the operator to be a trained radiologist. We have successfully evaluated the reproducibility of our segmentation approach across different institutions and various operators [36]. At Mayo Clinic and Vanderbilt University, we currently effectively utilise radiology technician in the 3D laboratory to clinically segment pulmonary nodules for other radiomics applications. After segmentation, the BRODERS classifier can be calculated within a few seconds.

Our study has several limitations. First, it is a retrospective study with the limitations inherent in this type of study design. Second, the CT scans for the Vanderbilt cohort were largely obtained at a single institution using similar scanners and acquisition protocols, and all nodules were incidentally rather than screen-detected. In addition, our validation cohort included 79 benign and 91 malignant nodules, which may not reflect typical nodule cohorts as encountered in all clinical practice settings and is certainly not reflective of the disease prevalence encountered in a screening cohort [37]. Populations with different proportions of malignant nodules may affect our model's positive and negative predictive values. Finally, the validation cohort is relatively small. However, the paucity of radiomic studies using external, well-curated validation cohorts, strengthens the significance of our work. Lastly, the diagnostic performance of the Brock model, which was originally derived from a cohort of screen-detected nodules, may have been altered by applying it to the incidentally discovered nodules in the Vanderbilt dataset.

To mitigate these potential issues, we are planning to prospectively validate the performance of the BRODERS classifier in a representative mixed multicentre dataset of incidentally and screen-detected lung nodules.

In conclusion, herein we present the validation of the BRODERS classifier. Additional validation in other external datasets and further prospective validation may prove the value of the BRODERS classification as guidance to clinicians. In the near future, BRODERS might be used in practice to leverage the wealth of features readily available in CT datasets and facilitate individualised management decisions for screen- or incidentally identified lung nodules.

We would like to dedicate this manuscript to honor the life and the extraordinary contributions to medical research of our dear friend and long-time mentor and coinvestigator Professor Richard A. Robb, PhD (2 December, 1942 to 29 August, 2020).

Conflict of interest: F. Maldonado reports grants from Department of Defense (Lung Cancer Research Program under award number W81XWH-15-1-0110), during the conduct of the study; and holds intellectual property rights as an inventor of the CANARY software and BRODERS classifier, but does not receive any financial relationships regarding this software. C. Varghese has nothing to disclose. S. Rajagopalan reports grants from Department of Defense (Lung Cancer Research Program under award number W81XWH-15-1-0110 to F. Maldonado), during the conduct of the study; CALIPER software is licensed to Imbio, LLC, from whom The Mayo Clinic and B.J. Bartholmai receive royalties related to CALIPER (also known as Imbio Lung Texture Analysis, LTA); and S. Rajagoplan holds intellectual property rights as an inventor of the CANARY software and BRODERS classifier, but does not receive any financial relationships regarding this software. F. Duan reports grants from Department of Defense (Lung Cancer Research Program under award number W81XWH-15-1-0110 to F. Maldonado), during the conduct of the study. A.B. Balar has nothing to disclose. D.A. Lakhani has nothing to disclose. S.L. Antic has nothing to disclose. P.P. Massion has nothing to disclose. T.F. Johnson reports grants from Department of Defense, during the conduct of the study. R.A. Karwoski reports grants from Department of Defense (Lung Cancer Research Program under award number W81XWH-15-1-0110 to F. Maldonado), during the conduct of the study; CALIPER software is licensed to Imbio, LLC, from whom The Mayo Clinic and B.J. Bartholmai receive royalties related to CALIPER (also known as Imbio Lung Texture Analysis, LTA); and R.A. Karwoski holds intellectual property rights as an inventor of the CANARY software and BRODERS classifier, but does not receive any financial relationships regarding this software. R.A. Robb reports grants from Department of Defense (Lung Cancer Research Program under award number W81XWH-15-1-0110 to F. Maldonado), during the conduct of the study; CALIPER software is licensed to Imbio, LLC, from whom The Mayo Clinic and B.J. Bartholmai receive royalties related to CALIPER (also known as Imbio Lung Texture Analysis, LTA); and R.A. Robb holds intellectual property rights as an inventor of the CANARY software and BRODERS classifier, but does not receive any financial relationships regarding this software. B.J. Bartholmai reports grants from Department of Defense (Lung Cancer Research Program under award number W81XWH-15-1-0110 to F. Maldonado), during the conduct of the study; personal fees for advisory board work from Promedior, LLC, outside the submitted work; CALIPER software is licensed to Imbio, LLC, from whom The Mayo Clinic and B.J. Bartholmai receive royalties related to CALIPER (also known as Imbio Lung Texture Analysis, LTA); and B.J. Bartholmai holds intellectual property rights as an inventor of the CANARY software and BRODERS classifier, but does not receive any financial relationships regarding this software. T. Peikert reports grants from Department of Defense (Lung Cancer Research Program under award number W81XWH-15-1-0110 to F. Maldonado), during the conduct of the study; fees paid to institution for advisory board work from AstraZeneca and Novocure, outside the submitted work; and holds intellectual property rights as an inventor of the CANARY software and BRODERS classifier, but does not receive any financial relationships regarding this software.

Support statement: This work was supported in part by the Office of the Assistant Secretary of Defense for Health Affairs, through the Lung Cancer Research Program under award number W81XWH-15-1-0110 (F. Maldonado) and by CA 152662 and CA 186145 (P. Massion). Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the Department of Defense. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Funding information for this article has been deposited with the Crossref Funder Registry.

References

1 Bade BC, Dela Cruz CS. Lung cancer 2020: epidemiology, etiology, and prevention. *Clin Chest Med* 2020; 41: 1–24.
2 Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin* 2020; 70: 7–30.
3 de Koning HJ, van der Aalst CM, de Jong PA, *et al.* Reduced lung-cancer mortality with volume CT screening in a randomized trial. *N Engl J Med* 2020; 382: 503–513.
4 National Lung Screening Trial Research Trial, Aberle DR, Berg CD, *et al.* The National Lung Screening Trial: overview and study design. *Radiology* 2011; 258: 243–253.
5 Pastorino U, Silva M, Sestini S, *et al.* Prolonged lung cancer screening reduced 10-year mortality in the MILD trial: new confirmation of lung cancer screening efficacy. *Ann Oncol* 2019; 30: 1162–1169.
6 Triplette M, Thayer JH, Pipavath SN, *et al.* Poor uptake of lung cancer screening: opportunities for improvement. *J Am Coll Radiol* 2019; 16: 446–450.
7 Dama E, Melocchi V, Colangelo T, *et al.* Deciphering the molecular profile of lung cancer: new strategies for the early detection and prognostic stratification. *J Clin Med* 2019; 8: 108.
8 MacMahon H, Naidich DP, Goo JM, *et al.* Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. *Radiology* 2017; 284: 228–243.
9 Manos D, Seely JM, Taylor J, *et al.* The Lung Reporting and Data System (LU-RADS): a proposal for computed tomography screening. *Can Assoc Radiol J* 2014; 65: 121–134.
10 Pinsky PF, Gierada DS, Black W, *et al.* Performance of Lung-RADS in the National Lung Screening Trial: a retrospective assessment. *Ann Intern Med* 2015; 162: 485–491.
11 Gould MK, Tang T, Liu I-LA, *et al.* Recent trends in the identification of incidental pulmonary nodules. *Am J Respir Crit Care Med* 2015; 192: 1208–1214.
12 Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Cogn Model* 1988; 5: 1.
13 Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev* 1958; 65: 386–408.
14 Chartrand G, Cheng PM, Vorontsov E, *et al.* Deep learning: a primer for radiologists. *Radiographics* 2017; 37: 2113–2131.
15 Hassani C, Varghese BA, Nieva J, *et al.* Radiomics in pulmonary lesion imaging. *AJR Am J Roentgenol* 2019; 212: 497–504.
16 Carter BW, Godoy MC, Erasmus JJ. Predicting malignant nodules from screening CTs. *J Thorac Oncol* 2016; 11: 2045–2047.
17 Rizzo S, Botta F, Raimondi S, *et al.* Radiomics: the facts and the challenges of image analysis. *Eur Radiol Exp* 2018; 2: 36.
18 Choi HK, Ghobrial M, Mazzone PJ. Models to estimate the probability of malignancy in patients with pulmonary nodules. *Ann Am Thorac Soc* 2018; 15: 1117–1126.
19 Peikert T, Duan F, Rajagopalan S, *et al.* Novel high-resolution computed tomography-based radiomic classifier for screen-identified pulmonary nodules in the National Lung Screening Trial. *PLoS One* 2018; 13: e0196910.
20 McWilliams A, Tammemagi MC, Mayo JR, *et al.* Probability of cancer in pulmonary nodules detected on first screening CT. *N Engl J Med* 2013; 369: 910–919.
21 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; 44: 837–845.
22 Hawkins S, Wang H, Liu Y, *et al.* Predicting malignant nodules from screening CT scans. *J Thorac Oncol* 2016; 11: 2120–2128.
23 Huang P, Park S, Yan R, *et al.* Added value of computer-aided CT image features for early lung cancer diagnosis with small pulmonary nodules: a matched case-control study. *Radiology* 2018; 286: 286–295.
24 Paul R, Hawkins SH, Schabath MB, *et al.* Predicting malignant nodules by fusing deep features with classical radiomics features. *J Med Imaging* 2018; 5: 011021.
25 Choi W, Oh JH, Riyahi S, *et al.* Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer. *Med Phys* 2018; 45: 1537–1549.
26 Chen C-H, Chang C-K, Tu C-Y, *et al.* Radiomic features analysis in computed tomography images of lung nodule classification. *PLoS One* 2018; 13: e0192002.
27 Ardila D, Kiraly AP, Bharadwaj S, *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med* 2019; 25: 954–961.
28 Massion PP, Antic S, Ather S, *et al.* Assessing the accuracy of a deep learning method to risk stratify indeterminate pulmonary nodules. *Am J Respir Crit Care Med* 2020; 202: 241–249.
29 Kumar V, Gu Y, Basu S, *et al.* Radiomics: the process and the challenges. *Magn Reson Imaging* 2012; 30: 1234–1248.
30 Welch ML, McIntosh C, Haibe-Kains B, *et al.* Vulnerabilities of radiomic signature development: the need for safeguards. *Radiother Oncol* 2019; 130: 2–9.
31 Zhang C, Vinyals O, Munos R, *et al.* A study on overfitting in deep reinforcement learning. *arXiv* 2018; preprint [https://arxiv.org/abs/1804.06893].
32 Voets M, Møllersen K, Bongo LA. Reproduction study using public data of: development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PLoS One* 2019; 14: e0217541.

33 Al-Ameri A, Malhotra P, Thygesen H, *et al.* Risk of malignancy in pulmonary nodules: a validation study of four prediction models. *Lung Cancer* 2015; 89: 27–30.
34 Uthoff J, Koehn N, Larson J, *et al.* Post-imaging pulmonary nodule mathematical prediction models: are they clinically relevant? *Eur Radiol* 2019; 29: 5367–5377.
35 van Riel SJ, Ciompi F, Jacobs C, *et al.* Malignancy risk estimation of screen-detected nodules at baseline CT: comparison of the PanCan model, Lung-RADS and NCCN guidelines. *Eur Radiol* 2017; 27: 4019–4029.
36 Nakajima EC, Frankland MP, Johnson TF, *et al.* Assessing the inter-observer variability of Computer-Aided Nodule Assessment and Risk Yield (CANARY) to characterize lung adenocarcinomas. *PLoS One* 2018; 13: e0198118.
37 Wahidi MM, Govert JA, Goudar RK, *et al.* Evidence for the treatment of patients with pulmonary nodules: when is it lung cancer?: ACCP evidence-based clinical practice guidelines. *Chest* 2007; 132: Suppl. 3, 94S–107S.