EUROPEAN RESPIRATORY *journal*

FLAGSHIP SCIENTIFIC JOURNAL OF ERS

Early View

Original article

Validation of the BRODERS classifier (Benign *versus* aggressive nODule Evaluation using Radiomic Stratification), a novel high-resolution computed tomography-based radiomic classifier for indeterminate pulmonary nodules

Fabien Maldonado, Cyril Varghese, Srinivasan Rajagopalan, Fenghai Duan, Aneri Balar, Dhairya A. Lakhani, Sanja B. Antic, Pierre Massion, Tucker F. Johnson, Ronald A. Karwoski, Richard A. Robb, Brian J. Bartholmai, Tobias Peikert

Please 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 high-resolution computed tomography-based radiomic classifier for indeterminate pulmonary nodules. *Eur Respir J* 2020; in press (https://doi.org/10.1183/13993003.02485-2020).

Validation of the BRODERS classifier (Benign versus aggressive nODule Evaluation using Radiomic Stratification), a novel high-resolution computed tomography-based radiomic classifier for indeterminate pulmonary nodules

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

[1] Division of Allergy, Pulmonary and Critical Care Medicine, Vanderbilt University Medical Center, Nashville, TN

[2] Division of Pulmonary and Critical Care Medicine, Mayo Clinic, Rochester, MN

[3] Department of Physiology and Biomechanical Engineering, Mayo Clinic, Rochester, MN

[4] Pulmonary section, Medical Service, Tennessee Valley Healthcare Systems, Nashville Campus, Nashville, TN

[5] Department of Biostatistics and Center for Statistical Sciences, Brown University School of Public Health, Providence, RI

[6] Department of Radiology, Mayo Clinic, Rochester, MN

[*] Authors contributed equally to this work

Corresponding author:

Tobias Peikert, MD

Division of Pulmonary and Critical Care Medicine

200 First Street SW

Rochester, MN 55905

Peikert.Tobias@mayo.edu

Keywords: lung cancer, radiomics, benign, malignant, risk.

Competing Interests

Some of the authors (Fabien Maldonado, Tobias Peikert, Brian Bartholmai, Srinivasan Rajagopalan and Ronald Karwoski) are co-inventors of a CT-based radiomic classifier for lung adenocarcinomas (distinct from the present work). None of the other co-authors have any disclosure.

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 No. W81XWH-15-1-0110 (Fabien Maldonado) and by CA 152662 and CA 186145 (Pierre 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.

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 pretest probability of malignancy is relatively straightforward, those with intermediate pretest probability commonly require advanced imaging or biopsy. Non-invasive risk stratification tools are highly desirable.

Methods: We previously developed the BRODERS classifier (**B**enign vs agg**R**essive n**OD**ule **E**valuation using **R**adiomic **S**tratification), a conventional predictive radiomic model based on 8 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 cutoff 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%, the PPV was 78.4% and the NPV was 79.2%, respectively.

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

Introduction: Lung cancer remains the deadliest malignancy in the US and worldwide [1]. While lung cancer 5-year survival has improved over the past decade, more than 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 United States (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 the endorsement by the Center for Medicare Services (CMS) and the United States Preventive Services Task Force (USPSTF) 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 the 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 nodule $\leq$ 4mm 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 about 10% [10].

In addition to screen-detected IPNs, incidentally discovered IPNs are also on the rise. This development is due to increased utilization of diagnostic cross-sectional chest imaging and the more widespread availability of advanced high-resolution CT (HRCT). 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 around 1.5 million adult Americans will be diagnosed with a pulmonary nodule annually [11]. The magnitude of the clinical challenges of non-invasively 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 analyze imaging data [12] to identify predictive and descriptive radiologic features not otherwise evident to a human observer that may correlate with the biological behavior of the lesion analyzed. 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 last decade [15]. However, there is great variability in image acquisition, feature extraction methodology and statistical modeling 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, characterized by substantial variability in scanner technology, image acquisition and reconstruction [17]. It is thus 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 (**B**enign vs agg**R**essive n**OD**ule **E**valuation using **R**adiomic **S**tratification), 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 center. We also compare the performance of our model to the performance of an established clinical prediction model routinely used in clinical practice [15].

Methods

Classifier Development

The development of our radiomic classifier and CALIPER and CANARY used to analyze the lung and nodule texture has been described and validated previously [19-22]. The development of our radiomic classifier has been previously published [19]. 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 (ANALYZE, Biomedical Imaging Resource, Mayo Clinic, Rochester, MN). 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 preselected radiological features organized in the following broad categories characterizing 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 curves (AUCs) were 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 at least 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 cutoff 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# 151500) 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 DICOM images of the CT scans were transferred to Mayo Clinic Rochester, Minnesota 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 ANALYZE software described above. 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. (Supplemental **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 bv, Ostend, Belgium; https://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 (9.4-11.3mm) versus 17.5 mm CI (16.2-17.8 mm), respectively ($p<0.001$) (**Supplemental Figure S2)**. **Supplemental 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 **Supplemental Table S1.** and **S2.**

Using the optimal cutoff of 0.478 identified via Youden's index, the sensitivity and specificity of the BRODERS classifier were 88.7% and 86.2% in the NLST screen-detected nodule cohort (n=685), respectively. For nodules with intermediate pre-test probability of

malignancy (5-65%) by the Brock model (n=416) the Sensitivity was 91.9% and the Specificity was 71.6% using the same cutoff.

For the entire Vanderbilt incidental nodule dataset (n=170), the Sensitivity was 92.3%, the 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), the Sensitivity was 94%, Specificity was 46%, the PPV was 78.4% and the NPV was 79.2%. The performance of the BRODERS classifier across different Brock-probability cut offs for the intermediate lung nodules are shown in **Supplemental Table S3**. and **S4**.

The direct correlation between the Brock Model and the BRODERS classifier for the Vanderbilt University cohort are shown in **Supplemental 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% for the entire cohorts and 10% and 22% 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 recently described the use of radiomics for pulmonary nodule characterization. Some of them used large datasets like the NLST[22-24] or the Lung Image Database Consortium image collection (LIDC) [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. It is also important to take into consideration the potential differences between screen and incidentally identified lung nodules, as models derived from screening cohorts may perform well in similar cohorts, but may not be generalizable 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 and colleagues reported the development of their deep learning-based Lung Cancer Prediction

Convolutional Neural Network (LCP-CNN) model.[28] The reported AUCs of 0.92, 0.84 and 0.92 in the NLST (training set, screen-detected), a Vanderbilt University and an Oxford University validation sets (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 also 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 characterization were selected for their direct relevance to predictive biological features, such as nodule texture, surface characteristics and location.

It is also important to recognize that to a variety of factors, including strict inclusion criteria and healthy volunteer effect, subjects enrolled in screening studies tend to be substantially different than patients presenting at lung nodule clinics or even patients eligible for lung cancer screening [33]. 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 generalizability 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 [34, 35]. In addition, a study by Van Riel et al. suggested that the Brock model may be preferable to both Lung-RADS and the National Comprehensive Cancer Network guidelines to classify nodules [36]. 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 pretest probability of malignancy in the Vanderbilt cohort, a nodule with a 50% pretest probability could be reclassified as low posttest probability after negative radiomic analysis (7.7%) or high posttest 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 (ANALYZE, Biomedical Imaging Resource, Mayo Clinic, Rochester, MN) is fast, 1-5 minutes 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.[37] At Mayo Clinic and Vanderbilt University, we currently effectively utilize radiology technician in the 3D-laboratory to clinically segment pulmonary nodules for other radiomics application. 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 [38]. 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 multi-center 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 individualized management decisions for screen- or incidentally identified lung nodules.

References

1. Bade BC, Dela Cruz CS. Lung Cancer 2020: Epidemiology, Etiology, and Prevention. *Clin Chest Med* 2020: 41(1): 1-24.
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin* 2020: 70(1): 7-30.
3. de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, Lammers JJ, Weenink C, Yousaf-Khan U, Horeweg N, van 't Westeinde S, Prokop M, Mali WP, Mohamed Hoesein FAA, van Ooijen PMA, Aerts J, den Bakker MA, Thunnissen E, Verschakelen J, Vliegenthart R, Walter JE, Ten Haaf K, Groen HJM, Oudkerk M. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *N Engl J Med* 2020: 382(6): 503-513.
4. National Lung Screening Trial Research T, Aberle DR, Berg CD, Black WC, Church TR, Fagerstrom RM, Galen B, Gareen IF, Gatsonis C, Goldin J, Gohagan JK, Hillman B, Jaffe C, Kramer BS, Lynch D, Marcus PM, Schnall M, Sullivan DC, Sullivan D, Zylak CJ. The National Lung Screening Trial: overview and study design. *Radiology* 2011: 258(1): 243-253.
5. Pastorino U, Silva M, Sestini S, Sabia F, Boeri M, Cantarutti A, Sverzellati N, Sozzi G, Corrao G, Marchiano A. Prolonged lung cancer screening reduced 10-year mortality in the MILD trial: new confirmation of lung cancer screening efficacy. *Ann Oncol* 2019: 30(7): 1162-1169.
6. Triplette M, Thayer JH, Pipavath SN, Crothers K. Poor Uptake of Lung Cancer Screening: Opportunities for Improvement. *J Am Coll Radiol* 2019: 16(4 Pt A): 446-450.
7. Dama E, Melocchi V, Colangelo T, Cuttano R, Bianchi F. Deciphering the Molecular Profile of Lung Cancer: New Strategies for the Early Detection and Prognostic Stratification. *Journal of clinical medicine* 2019: 8(1).
8. MacMahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, Mehta AC, Ohno Y, Powell CA, Prokop M, Rubin GD, Schaefer-Prokop CM, Travis WD, Van Schil PE, Bankier AA. Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. *Radiology* 2017: 284(1): 228-243.
9. Manos D, Seely JM, Taylor J, Borgaonkar J, Roberts HC, Mayo JR. The Lung Reporting and Data System (LU-RADS): a proposal for computed tomography screening. *Can Assoc Radiol J* 2014: 65(2): 121-134.
10. Pinsky PF, Gierada DS, Black W, Munden R, Nath H, Aberle D, Kazerooni E. Performance of Lung-RADS in the National Lung Screening Trial: a retrospective assessment. *Annals of internal medicine* 2015: 162(7): 485-491.
11. Gould MK, Tang T, Liu I-LA, Lee J, Zheng C, Danforth KN, Kosco AE, Di Fiore JL, Suh DE. Recent trends in the identification of incidental pulmonary nodules. *American journal of respiratory and critical care medicine* 2015: 192(10): 1208-1214.
12. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Cognitive modeling* 1988: 5(3): 1.
13. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* 1958: 65(6): 386.
14. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, Kadoury S, Tang A. Deep learning: a primer for radiologists. *Radiographics : a review publication of the Radiological Society of North America, Inc* 2017: 37(7): 2113-2131.

15. Hassani C, Varghese BA, Nieva J, Duddalwar V. Radiomics in Pulmonary Lesion Imaging. *AJR American journal of roentgenology* 2019: 212(3): 497-504.
16. Carter BW, Godoy MC, Erasmus JJ. Predicting malignant nodules from screening CTs. *Journal of Thoracic Oncology* 2016: 11(12): 2045-2047.
17. Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, Bellomi M. Radiomics: the facts and the challenges of image analysis. *Eur Radiol Exp* 2018: 2(1): 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(10): 1117-1126.
19. Peikert T, Duan F, Rajagopalan S, Karwoski RA, Clay R, Robb RA, Qin Z, Sicks J, Bartholmai BJ, Maldonado F. Novel high-resolution computed tomography-based radiomic classifier for screen-identified pulmonary nodules in the National Lung Screening Trial. *PLoS One* 2018: 13(5): e0196910.
20. McWilliams A, Tammemagi MC, Mayo JR, Roberts H, Liu G, Soghrati K, Yasufuku K, Martel S, Laberge F, Gingras M. Probability of cancer in pulmonary nodules detected on first screening CT. *New England Journal of Medicine* 2013: 369(10): 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(3): 837-845.
22. Hawkins S, Wang H, Liu Y, Garcia A, Stringfield O, Krewer H, Li Q, Cherezov D, Gatenby RA, Balagurunathan Y, Goldgof D, Schabath MB, Hall L, Gillies RJ. Predicting Malignant Nodules from Screening CT Scans. *J Thorac Oncol* 2016: 11(12): 2120-2128.
23. Huang P, Park S, Yan R, Lee J, Chu LC, Lin CT, Hussien A, Rathmell J, Thomas B, Chen C, Hales R, Ettinger DS, Brock M, Hu P, Fishman EK, Gabrielson E, Lam S. 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(1): 286-295.
24. Paul R, Hawkins SH, Schabath MB, Gillies RJ, Hall LO, Goldgof DB. Predicting malignant nodules by fusing deep features with classical radiomics features. *J Med Imaging (Bellingham)* 2018: 5(1): 011021.
25. Choi W, Oh JH, Riyahi S, Liu CJ, Jiang F, Chen W, White C, Rimner A, Mechalakos JG, Deasy JO. Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer. *Medical physics* 2018: 45(4): 1537-1549.
26. Chen C-H, Chang C-K, Tu C-Y, Liao W-C, Wu B-R, Chou K-T, Chiou Y-R, Yang S-N, Zhang G, Huang T-C. Radiomic features analysis in computed tomography images of lung nodule classification. *PloS one* 2018: 13(2): e0192002.
27. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, Tse D, Etemadi M, Ye W, Corrado G. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine* 2019: 1.
28. Massion PP, Antic S, Ather S, Arteta C, Brabec J, Chen H, Declerck J, Dufek D, Hickes W, Kadir T, Kunst J, Landman BA, Munden RF, Novotny P, Peschl H, Pickup LC, Santos C, Smith GT, Talwar A, Gleeson F. Assessing the Accuracy of a Deep Learning Method to Risk Stratify Indeterminate Pulmonary Nodules. *Am J Respir Crit Care Med* 2020.
29. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, Forster K, Aerts HJ, Dekker A, Fenstermacher D. Radiomics: the process and the challenges. *Magnetic resonance imaging* 2012: 30(9): 1234-1248.
30. Welch ML, McIntosh C, Haibe-Kains B, Milosevic MF, Wee L, Dekker A, Huang SH, Purdie TG, O'Sullivan B, Aerts H, Jaffray DA. Vulnerabilities of radiomic signature

development: The need for safeguards. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology* 2019: 130: 2-9.
31. Zhang C, Vinyals O, Munos R, Bengio S. A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:180406893* 2018.
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(6): e0217541.
33. MacMahon H, Naidich DP, Goo JM, Lee KS, Leung AN, Mayo JR, Mehta AC, Ohno Y, Powell CA, Prokop M. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. *Radiology* 2017: 284(1): 228-243.
34. Al-Ameri A, Malhotra P, Thygesen H, Plant PK, Vaidyanathan S, Karthik S, Scarsbrook A, Callister ME. Risk of malignancy in pulmonary nodules: a validation study of four prediction models. *Lung cancer (Amsterdam, Netherlands)* 2015: 89(1): 27-30.
35. Uthoff J, Koehn N, Larson J, Dilger SKN, Hammond E, Schwartz A, Mullan B, Sanchez R, Hoffman RM, Sieren JC. Post-imaging pulmonary nodule mathematical prediction models: are they clinically relevant? *European radiology* 2019.
36. van Riel SJ, Ciompi F, Jacobs C, Wille MMW, Scholten ET, Naqibullah M, Lam S, Prokop M, Schaefer-Prokop C, van Ginneken B. Malignancy risk estimation of screen-detected nodules at baseline CT: comparison of the PanCan model, Lung-RADS and NCCN guidelines. *European radiology* 2017: 27(10): 4019-4029.
37. Nakajima EC, Frankland MP, Johnson TF, Antic SL, Chen H, Chen SC, Karwoski RA, Walker R, Landman BA, Clay RD, Bartholmai BJ, Rajagopalan S, Peikert T, Massion PP, Maldonado F. Assessing the inter-observer variability of Computer-Aided Nodule Assessment and Risk Yield (CANARY) to characterize lung adenocarcinomas. *PLoS One* 2018: 13(6): e0198118.
38. Wahidi MM, Govert JA, Goudar RK, Gould MK, McCrory DC. Evidence for the treatment of patients with pulmonary nodules: when is it lung cancer?: ACCP evidence-based clinical practice guidelines. *Chest* 2007: 132(3): 94S-107S.

Tables and Figure Legends

Figure 1. ROC for the NLST cohort comparing the Brock and Radiomics classifications. Panel (**A**) is for the entire cohort. AUC Brock 0.833 (95% CI = 0.803-0.860); AUC Radiomics 0.939 (0.918-0.955). Panel (**B**) is for the low risk (Brock score < 5%) group, AUC Brock 0.795 (0.74-0.842); AUC Radiomics 0.925 (0.886-0.954). Panel (**C**) is for the 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. ROC for the Vanderbilt cohort comparing the Brock and Radiomics classifications. Panel (**A**) is for the entire cohort. AUC Brock 0.872 (95% CI = 0.812-0.918); AUC Radiomics 0.904 (0.849-0.943). Panel (**B**) is for the low risk (Brock score < 5%) group, AUC Brock 0.658 (0.496-0.797); AUC Radiomics 0.796 (0.644-0.904). Panel (**C**) is for the intermediate risk (5% Brock Score < 65) group. AUC Brock 0.798 (0.717-0.864); AUC Radiomics 0.856 (0.782-0.912).

Table 1. Baseline characteristics of the two cohorts described in the study

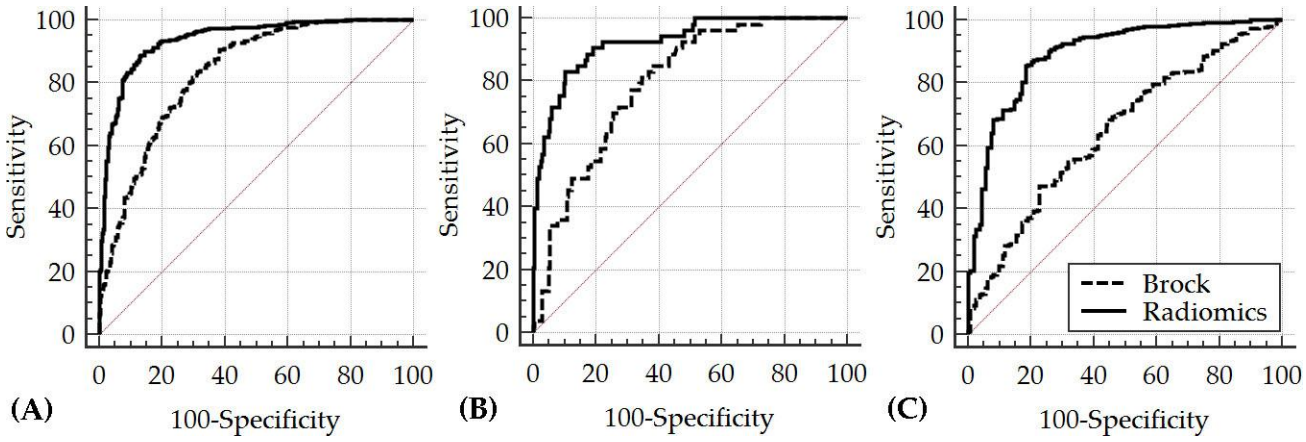
	NLST (N=685)	VANDERBILT (N=170)
AGE MEAN YEARS [SD]	63 [5.3]	66 [7.6]
GENDER [%]		
MEN	392 [57.2]	113 [66.5]
WOMEN	293 [42.8]	57 [33.5]
RACE [%]		
CAUCASIAN	632 [92.3]	152 [89.4]
BLACK, ASIAN, OTHER	53 [7.7]	18 [10.6]
SMOKING [%]		
CURRENT	362 [52.8]	108 [64]
FORMER	327 [47.2]	58 [34]
NEVER	0	4 [2]
SMOKING PACK YEARS MEAN [SD]	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 MEAN MM [SD]	12.2 [6.5]	14.6 [6.9]
SPICULATION [%]	199 [29.1]	20 [11.8]

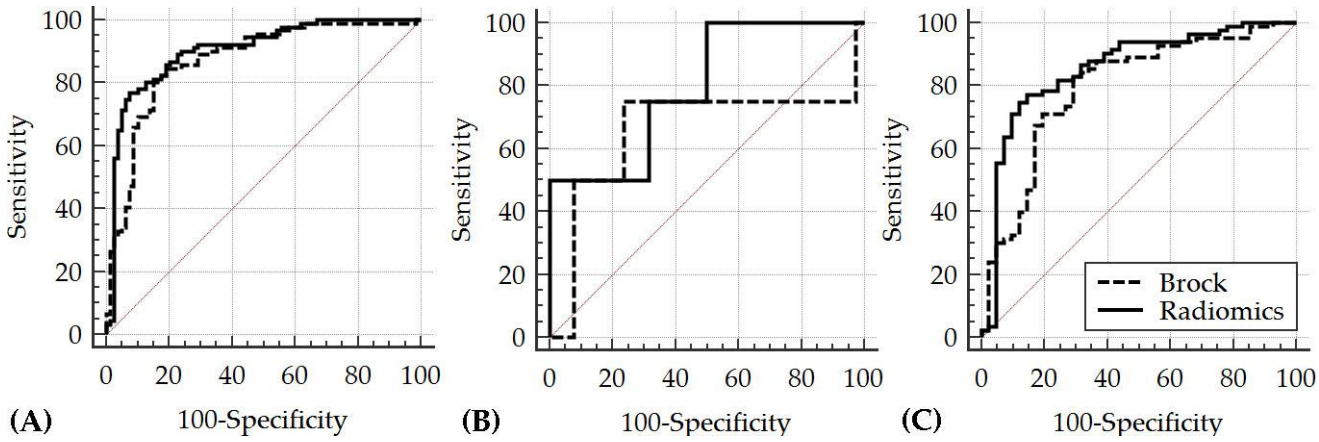
Table 2. Truth tables comparing histology versus BRODERS classifier versus Brock model probability categories in the NLST cohort.

Brock Model probability of malignancy	Clinical/ Histological Classification	BRODERS Benign	BRODERS Malignant
Low < 5 % (N = 257)	Benign (N = 204)	192	12
	Malignant (N = 53)	17	36
Intermediate 5 □ to < 65 (N = 416)	Benign (N = 109)	78	31
	Malignant (N = 307)	25	282
High □ 65 (N = 12)	Benign (N =0)	0	0
	Malignant (N = 12)	0	12

Table 3. Truth tables comparing histology versus BRODERS classifier versus Brock model probability categories in the Vanderbilt cohort.

Brock Model probability of malignancy	Clinical/ Histological Classification	BRODERS Benign	BRODERS Malignant
Low < 5 % (N = 42)	Benign (N = 38)	30	8
	Malignant (N = 4)	2	2
Intermediate 5 □ to < 65 (N = 126)	Benign (N = 41)	19	22
	Malignant (N = 85)	5	80
High □ 65 (N = 2)	Benign (N =0)	0	0
	Malignant (N = 2)	0	2


Sensitivity
100-Specificity
(A)
(B)
(C)
Brock
Radiomics


Sensitivity
100-Specificity
(A)
(B)
(C)
Brock
Radiomics

Supplemental Figure 1. Flow-diagram for the patient selection from the NLST (**A.**) and Vanderbilt University (**B.**) databases.

Supplemental Figure S2. Comparison of Nodule Diameter between malignant and benign nodules in the Vanderbilt University Cohort

Supplemental Figure S3: 2x2 contingency table with the axial scout images for the NLST cohort comparing the clinical/histological ground truth and the BRODERS radiomics prediction.

Supplemental Figure S4: 2x2 contingency table with the axial scout images for the Vanderbilt cohort comparing the clinical/histological ground truth and the BRODERS radiomics prediction.

Supplemental Figure S5 Comparison between Brock Model and BRODERS classifier for the Vanderbilt University cohort

Supplemental Table S1. Types of malignancies in the NLST cohort and distribution across the Brock and BRODERS classification

Histology	Brock < 5%	5% <= Brock < 65%	Brock >= 65%	BRODERS Benign	BRODERS Malignant
Adenocarcinoma (N=268)	34	224	10	39	229
Squamous cell carcinoma (N=71)	14	55	2	2	69
Large cell carcinoma (N=18)	3	15	0	1	17
Small Cell carcinoma (N=11)	2	9	0	0	11
Carcinoid (N=4)	0	4	0	0	4

Supplemental Table S2. Types of malignancies in the Vanderbilt cohort and distribution across the Brock and BRODERS classification

Histology	Brock < 5%	5% <= Brock < 65%	Brock >= 65%	BRODERS Benign	BRODERS Malignant
Adenocarcinoma (N=60)	2	57	1	3	57
Squamous cell carcinoma (N=24)	2	21	1	1	23
Large cell carcinoma (N=3)	0	3	0	0	3
Small Cell carcinoma (N=3)	0	3	0	1	2
Carcinoid (N=1)	0	1	0	1	0

Supplemental Table S3 Effect of different Brock cutoffs the intermediate probability group on BRODERS diagnostic performance in the NLST cohort.

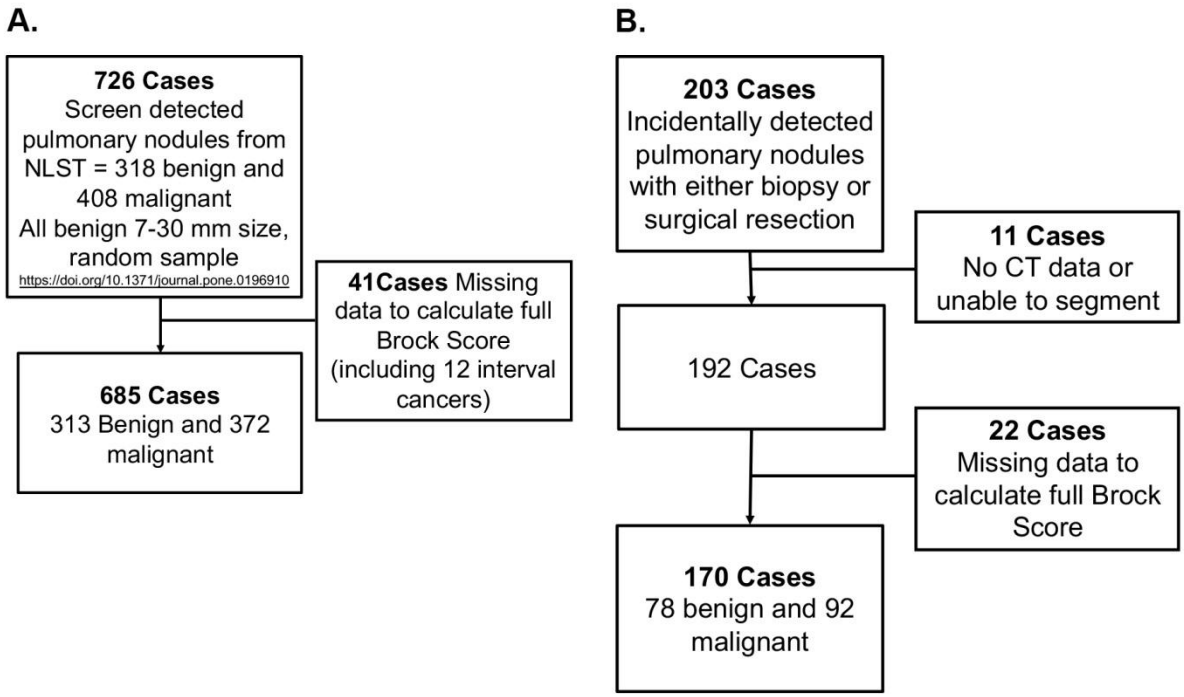
	Brock Score Cutoffs for the Intermediate Group					
	5-60	5-65	5-70	10-60	10-65	10-70
TN	78	78	78	47	47	47
FP	31	31	31	23	23	23
FN	25	25	25	15	15	15
TP	273	282	287	229	238	243
Sensitivity	0.92 (0.92-0.92)	0.9185 (0.918-0.919)	0.92 (0.919-0.921)	0.939 (0.938-0.94)	0.941 (0.939-0.942)	0.942 (0.941-0.943)
Specificity	0.72 (0.71-0.72)	0.7156 (0.713-0.718)	0.7156 (0.713-0.718)	0.671 (0.668-0.68)	0.67 (0.667-0.675)	0.671 (0.668-0.675)
PPV	0.898 (0.896-0.899)	0.9009 (0.899-0.902)	0.9025 (0.902-0.904)	0.909 (0.907-0.91)	0.912 (0.911-0.913)	0.914 (0.912-0.915)
NPV	0.757 (0.754-0.759)	0.757 (0.754-0.76)	0.757 (0.755-0.76)	0.758 (0.755-0.76)	0.758 (0.755-0.761)	0.758 (0.755-0.761)
BRR %	10.2	9.9	9.75	9.13	8.8	8.65
Brock-AUC 95% CI	0.65 (0.6-0.69)	0.66 (0.61-0.70)	0.66 (0.62-0.71)	0.61 (0.55-0.66)	0.62 (0.57-0.68)	0.63 (0.58-0.68)
Rad- AUC 95 % CI	0.89 (0.86- 0.92)	0.90 (0.86-0.92)	0.9 (0.86-0.92)	0.88 (0.84-0.92)	0.89 (0.85-0.92)	0.89 (0.85-0.92)

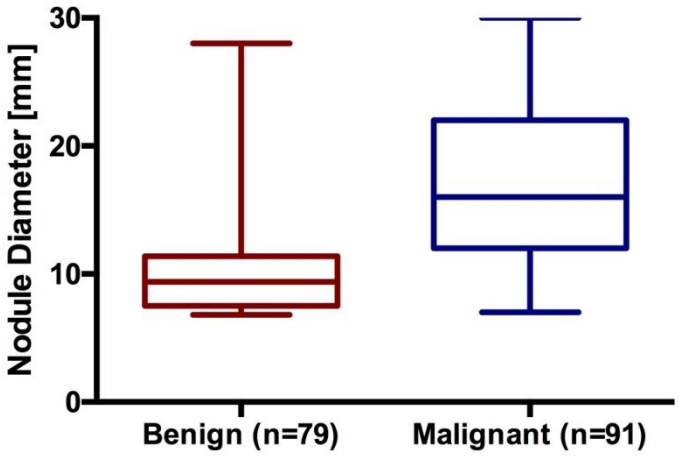
TN – True Negatives; FP – False Positives; FN – False Negatives; TP – True Positives; PPV – Positive Predictive Value, NPV – Negative Predictive Value; BRR – Benign Resection Rate; Rad-AUC – Radiomics AUC

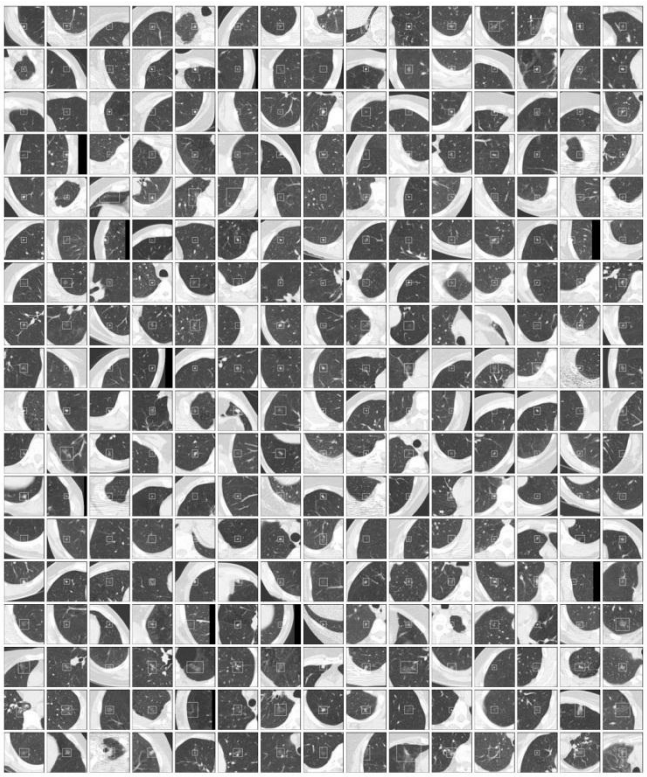
Supplemental Table S4 Effect of different Brock cutoffs the intermediate probability group on BRODERS diagnostic performance in the Vanderbilt University cohort.

	Brock Score Cutoffs for the Intermediate Group					
	5-60	5-65	5-70	10-60	10-65	10-70
TN	19	19	19	10	10	10
FP	22	22	22	13	13	13
FN	5	5	5	4	4	4
TP	78	80	81	70	72	73
Sensitivity	0.94 (0.938-0.941)	0.94 (0.939-0.943)	0.942 (0.94-0.943)	0.946 (0.9440.948)	0.947 (0.946-0.949)	0.948 (0.946-0.95)
Specificity	0.463 (0.458-0.468)	0.463 (0.459-0.468)	0.463 (0.456-0.468)	0.435 (0.428-0.441)	0.435 (0.428-0.441)	0.435 (0.428-0.441)
PPV	0.78 (0.777-0.783)	0.784 (0.782-0.786)	0.786 (0.784-0.789)	0.843 (0.841-0.846)	0.847 (0.845-0.85)	0.849 (0.846-0.851)
NPV	0.792 (0.786-0.797)	0.792 (0.786-0.797)	0.792 (0.786-0.797)	0.714 (0.707-0.722)	0.714 (0.707-0.722)	0.714 (0.707-0.722)
BRR %	22	21.57	21.36	15.66	15.29	15.11
Brock-AUC 95% CI	0.793 (0.711-0.861)	0.798 (0.717-0.864)	0.8 (0.72-0.866)	0.743 (0.644-0.826)	0.749 (0.652-0.831)	0.753 (0.656-0.834)
Rad- AUC 95 % CI	0.854 (0.779-0.911)	0.856 (0.782-0.912)	0.857 (0.784-0.913)	0.85 (0.763-0.915)	0.852 (0.767-0.916)	0.854 (0.769-0.916)

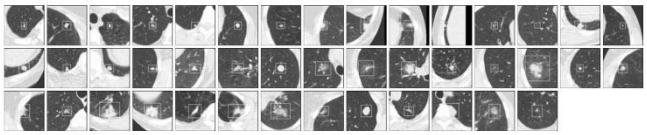
TN – True Negatives; FP – False Positives; FN – False Negatives; TP – True Positives; PPV – Positive Predictive Value, NPV – Negative Predictive Value; BRR – Benign Resection Rate; Rad-AUC – Radiomics AUC


A.
726 Cases
Screen detected pulmonary nodules from NLST = 318 benign and 408 malignant
All benign 7-30 mm size, random sample
https://doi.org/10.1371/journal.pone.0196910
41Cases Missing data to calculate full Brock Score (including 12 interval cancers)
685 Cases
313 Benign and 372 malignant
B.
203 Cases
Incidentally detected pulmonary nodules with either biopsy or surgical resection
11 Cases
No CT data or unable to segment
192 Cases
22 Cases
Missing data to calculate full Brock Score
170 Cases
78 benign and 92 malignant

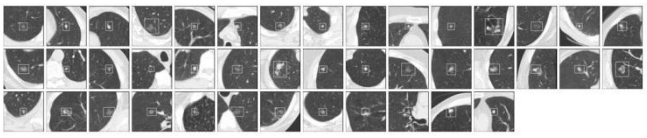

Nodule Diameter [mm]
30
20
10
0
Benign (n=79)
Malignant (n=91)



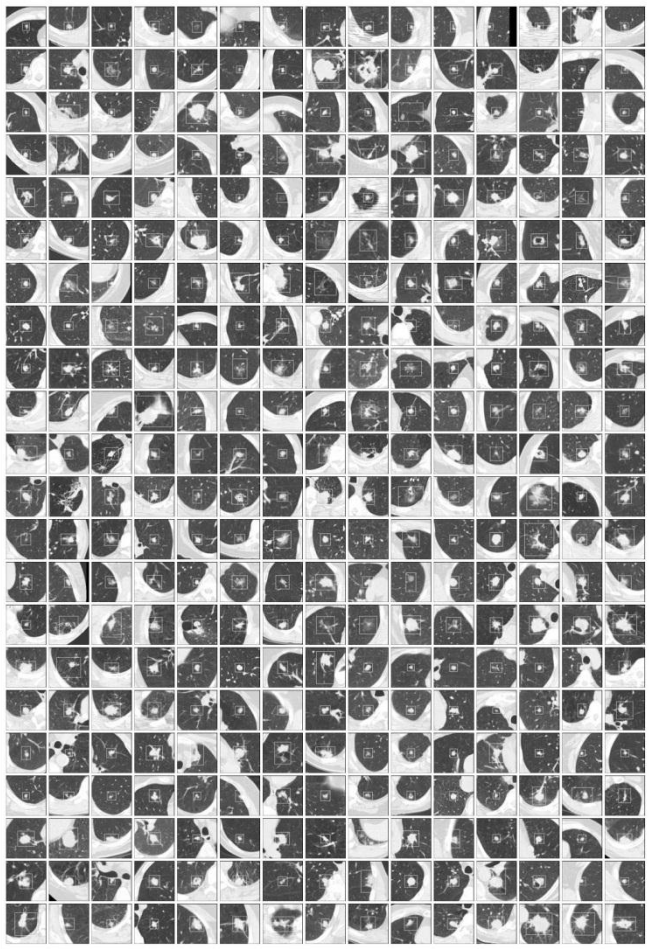
Benign Radiomics: Benign



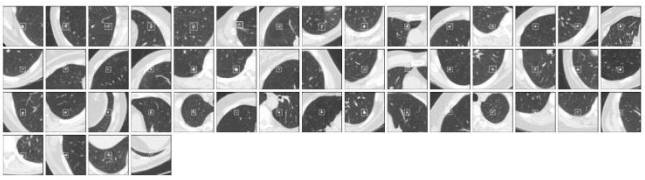
Benign Radiomics: Malignant



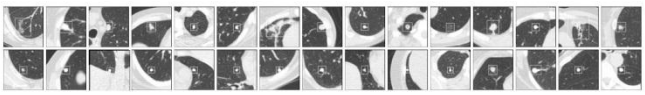
Malignant Radiomics: Benign



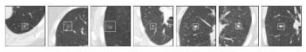
Malignant Radiomics: Malignant



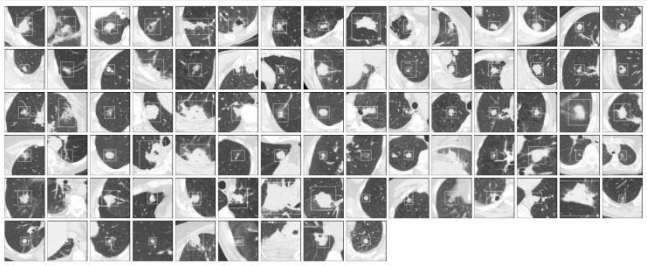
Benign Radiomics: Benign



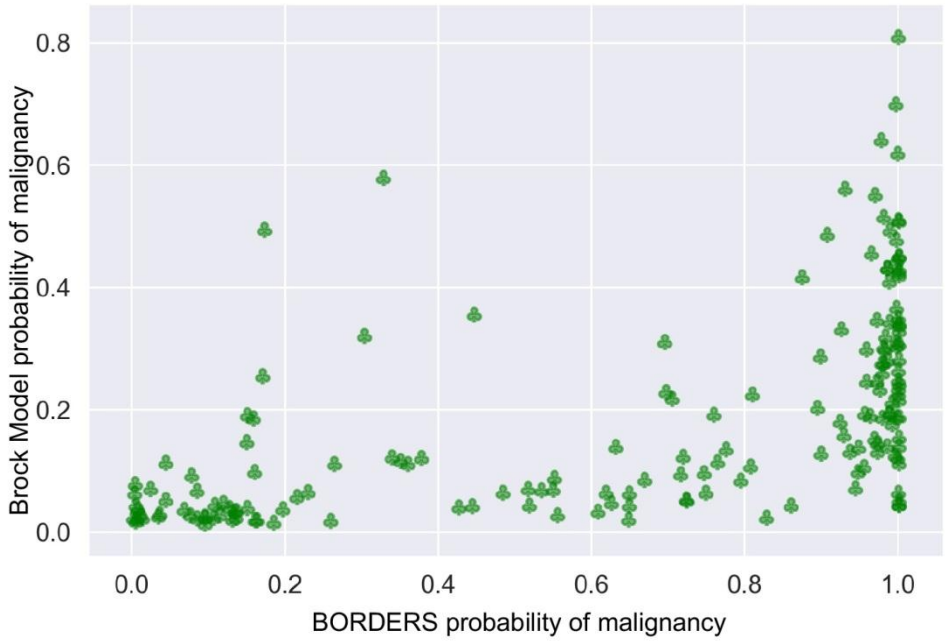
Benign Radiomics: Malignant



Malignant Radiomics: Benign



Malignant Radiomics: Malignant


Brock Model probability of malignancy
0.0
0.2
0.4
0.6
0.8
0.0
0.2
0.4
0.6
0.8
1.0
BORDERS probability of malignancy