DOI:
10.1259/bjr.20190159
10.1259/bjr.20190159

Document Version
Early version, also known as pre-print

Title

Non-invasive classification of non-small cell lung: a comparison between random forest models utilising radiomic and semantic features.

Short title

CT semantic and radiomic classification of NSCLC

1. Abstract

1.1. Purpose

Non-invasive distinction between squamous cell carcinoma (SCCA) and adenocarcinoma (ADCA) subtypes of non small-cell lung cancer (NSCLC) may be beneficial to patients unfit for invasive diagnostic procedures or when tissue is insufficient for diagnosis. The purpose of our study was to compare the performance of random forest algorithms utilizing CT radiomics and / or semantic features in classifying NSCLC.

1.2. Methods

Two thoracic radiologists scored 11 semantic features on CT scans of 106 patients with NSCLC. A set of 115 radiomics features was extracted from the CT scans. Random forest models were developed from semantic (RM-sem), radiomics (RM-rad), and all features combined (RM-all). External validation of models was performed using an independent test dataset (n=100) of CT scans. Model performance was measured with out-of-bag error and area under curve (AUC), and compared using receiver-operating characteristics curve analysis on the test dataset.

1.3. Results

The median (interquartile-range) error rates of the models were: RF-sem 24.5% (22.6%-37.5%), RF-rad 35.8% (34.9%-38.7%), and RM-all 37.7% (37.7-37.7). On training data, both RF-rad and RF-all gave perfect discrimination (AUC=1), which was significantly higher than that achieved by RF-sem (AUC=0.78; $p<0.0001$). On test data, however, RM-sem model (AUC=0.82) out-performed RM-rad and RM-all

(AUC=0.5 and AUC=0.56; $p<0.0001$), neither of which was significantly different from random guess ($p=0.9$ and 0.6 respectively).

1.4. Conclusion

Non-invasive classification of NSCLC can be done accurately using random forest classification models based on well-known CT-derived descriptive features. However, radiomics-based classification models performed poorly in this scenario when tested on independent data and should be used with caution, due to their possible lack of generalizability to new data.

1.5. Advances in knowledge

Our study describes novel CT-derived random forest models based on radiologist-interpretation of CT scans (semantic features) that can assist non small-cell lung cancer classification when histopathology is equivocal or when histopathologic sampling is not possible. It also shows that random forest models based on semantic features may be more useful than those built from computational radiomic features.

1. Introduction

Non-small cell lung cancers (NSCLC) comprise 85% of all primary lung malignancies [1]. Of these, approximately 60% are adenocarcinomas (ADCA) and 35-40% are squamous cell carcinomas (SCCA), with large cell cancers accounting for less than 5%[1]. Conventionally, ADCA and SCCA are differentiated by histopathologic examination of haematoxylin & eosin-stained slides. ADCAs, depending upon the predominant pathologic subtype, may exhibit lepidic, glandular, papillary or micropapillary, or solid sheet-like architecture. SCCAs are characterised by the presence of keratinisation, pearl formation, and intercellular bridges [2]. Frequently, NSCLC is diagnosed on sputum cytology or clinical and radiological features, but adequate tissue is not available to perform histological subtyping and molecular analysis, requiring a multidisciplinary approach for decision-making [2]. Although curative options for both NSCLC subtypes are similar - either surgical or with SABR - the two subtypes differ in prognosis and choice of targeted agents [3]. Hence, an accurate non-invasive test for NSCLC classification could serve as a valuable alternative for prognostication and choosing targeted agents in patients unsuitable for surgical resection.

Radiomics and machine learning (ML) are becoming increasingly popular in imaging research [4]. Radiomics involves computational analysis of a grey-scale image to derive features (e.g., mean, mode, kurtosis, and skewness) which are expected to quantify the tumour pathophysiology [5]. ML is the task of using radiomics and other relevant variables (e.g., age, sex, and air bronchogram) in suitable computational algorithms (e.g., random forests or logistic regression) to infer clinically relevant information, e.g., tumour subtype. CT radiomics has been shown to be moderately to highly accurate in predicting NSCLC subtype, with reported performance of 68% to

90% [6–8]. However, despite its promise [5], widespread acceptance of radiomics is hindered by largely unmet challenges surrounding variable reproducibility, procedure standardisation, and biologic explanation of used variables[4,9,10].

Semantic features, i.e. features derived from subjective interpretation of CT scans by a radiologist, have been shown in numerous independent studies to be related to tumour subtype and histopathology [11–17]. Air-bronchogram, and ground-glass opacification are more common in ADCA, whereas cavitation and spiculation are more common in SCCA [16,17]. To our knowledge however, despite these well-known associations, semantic features have not been modelled in ML algorithms to predict tumour sub-type and therefore help clinical decision making in a quantitative manner. Furthermore, no studies have compared or combined radiomic features with semantic features (e.g., air bronchogram and cavitation) in differentiating ADCA from SCCA.

We hypothesised that multivariate predictive models combining the strengths of semantic and radiomic features could yield potentially higher accuracy in NSCLC classification than either class of variables alone. Such non-invasive classification would benefit patients for whom an adequate histopathological subtyping cannot be obtained. Therefore, the objective of this study was to develop and compare NSCLC classification models based on semantic features, radiomic features, and combination of both.

2. Methods and patients

2.1. Patient population

The training dataset comprised patients referred to a single institution as follows: We identified pre-treatment CT scans of pathologically-proved NSCLC patients referred

to our tertiary care centre from 1/1/2011 to 31/12/2015. Patients were excluded if it was not possible to accurately determine tumour boundaries on CT, e.g. due to adjacent atelectasis. The final dataset comprised 106 studies (42 SCCA, 64 ADCA; figure 1). The independent validation cohort (n=100) comprised 65 ADCAs and 35 SCCAs downloaded from the Cancer Imaging archive, subsampled with respect to ADCAs to ensure balanced proportions [18–20]. Local ethics committee waived informed written consent for this retrospective study of anonymised data.

2.2. Imaging

Imaging of patients in the training dataset was performed on one of three Philips scanners: MX8000, Brilliance iCT 256, or Brilliance 40 (Philips Medical Systems, Best, Netherlands). Patients were imaged in the supine position at full inspiration. Scanning parameters were as follows: detector collimation: 0.625-0.75; rotation time: 0.5-0.75 seconds; tube voltage: 120 kVp; tube current: 34-229 mAs. 100-150 mL iopromide 300 (300 mg I/mL Ultravist, Bayer Pharma, Berlin, Germany) was administered intravenously at a rate of 2-4 mL/s after a 30-70 second delay.

2.3. Semantic features

Two thoracic radiologists (AN and MM, with 14 and 9 years' experience, respectively), blinded to histopathologic diagnosis, independently recorded 9 nodule semantic feature (table 1) and two background parenchymal features, i.e., emphysema (present or absent) and airway thickening (present or absent) [11,12,21–26].

Discrepant findings were resolved by consensus. Annotation of the validation dataset was performed by a separate blinded reader, UB (10 years' radiology experience), using the same descriptions.

2.4. Radiomic features

Tumours were delineated by UB open-source software ITK-Snap (version 3.4.0; supplemental data)[27]. From the segmented volumes-of-interest, 756 radiomic features were derived using an in-house feature extraction tool developed in MATLAB (Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States). Highly correlated redundant features (showing pairwise correlation coefficient >0.8; n=641) were removed to yield a final set of independent 115 radiomic features.

2.5. Random forests model development and validation

In this study, we used random forests as machine learning classifier. Random forests are known for their high performance and generalisability [28]. Here we present a summary of random forest model development; technical details are provided in the supplemental data.

A random forest model is a group of a large number of decision trees, e.g., 2000. The name 'random' alludes to the fact that each split of an individual decision tree is developed from a random subset of input variables. Each member tree is also trained on a slightly different variation of the dataset by using bootstrap sampling, i.e., sampling with replacement whereby several cases are sampled more than once and others omitted altogether (labelled 'out-of-bag' (OOB) samples). Since the OOB samples have not been used in training the particular tree, they are used for internal

validation and the proportion of misclassified cases in the OOB sample serves as a performance metric: OOB error. After training of all 2000 decision trees is complete, a new case is classified by the entire 'random forest' by obtaining votes from member trees. A decision threshold is set based on the preferred degree of sensitivity, to provide a final classification of each new case; for example, using a 50% probability threshold, a case may be classifying as ADCA if > 50% trees classify it as ADCA, and SCCA otherwise.

We developed three random forests classifiers using the training dataset: One classifier comprising semantic variables only (RF-sem), one comprising radiomic features only (RF-rad), and one comprising both semantic and radiomic features (RF-all). Model validation was performed on the independent validation cohort.

2.6. Statistical analysis

R version 3.3.2 was used for statistical analysis [29]. Continuous variables were reported as means and standard deviations. For descriptive analysis, differences between ADCAs and SCCAs were determined using Wilcoxon ranked sum test for continuous variables and using Fisher's exact test for categorical variables. Inter-observer agreement between the two radiologists with regards to semantic variables was measured with Cohen's kappa test and summarised as estimated weighted kappa scores and their 95% CIs. A p-value cut-off of 0.05 was used to determine statistical significance.

The performance of random forests models was reported in terms of two metrics: The OOB error of random forests models was reported as error rate of decision trees during internal validation. The second metric - Area under curve (AUC) – was used as performance metric of fully trained models and reported separately for training

and validation data. We used two metrics instead of one to illustrate both the robustness of individual trees (OOB error) and that of the forest as a whole (AUC). Both are related, and an ideal classifier should have both a low OOB error and a high AUC.

Since our random forests used large numbers of variables, we also measured the importance of individual variables in the training dataset using the 'mean decrease in accuracy' (MDA) metric, i.e., decrease in classifier accuracy by removing the variable in question. The higher the MDA of a variable the more important the variable is. A variable with MDA of zero has no association with the outcome (tumour subtype) and there is no decrease in classifier accuracy if that variable is removed. Variables with low but non-zero MDA are still useful since random forests by design work well when individual variables are weakly related to the outcome, and mitigate their weak association by pooling them into a robust final classifier [28]

3. Results

The mean interval between pathologic diagnosis and CT chest imaging was 21 days (range 5 – 41 days). Patients were aged from 40.3 to 85.5 years (median: 71.4 years), with similar gender proportions (50 females: 56 males). There were no significant differences between patients with ADCA versus SCCA in terms of age ($p=0.6$), smoking ($p=0.67$), or gender (0.55) (Table 2).

Of the 13 tested semantic variables, 3 were significantly more common in ADCAs, i.e. air bronchogram ($p <0.0001$), ground-glass component ($p=0.0006$), and satellite nodules ($p=0.004$). Cavitation was present in relatively few cases (n=9), of which 8 were SCCAs ($p=0.002$). Table 3 describes the frequencies of semantic variables in both NSCLC subtypes.

3.1. Comparison of random forest models

The semantic random forest (RF-sem) performed equally well on training and test datasets with AUC of 0.78 and 0.82 respective (figure 2). The radiomics-only and combined models gave performed tumour subtype discrimination on the training data (AUC 1), but very low performance on validation data of AUC 0.5 and 0.56 respectively, similar to random chance (figure 2). The OOB error of RF-sem (25.5%) was also lower than that of RF-rad (40.6%) and RF-all (37.7%). Figure 3 shows example tumours of each type with class probabilities, highlighting the probabilistic nature of the random forest model that can be exploited in clinical decision making to balance probability of tumour type against individual patient circumstances.

In terms variable importance, air bronchogram (MDA=0.039), ground-glass component (MDA=0.023), and cavitation (MDA=0.019) were the top-ranking semantic variables, whereas tumour location, spiculation, and tumour margins did not have any discriminatory value. Of the radiomic variables, the highest ranking variables were grey-level size-zone matrix (GLSZM) short zone low intensity emphasis (GLSZM-SZLIE; MDA=0.005), co-efficient of variation (MDA=0.004), and neighbourhood grey-tone difference matrix (NGTDM) coarseness (MDA=0.003). Variable importance of semantic features and top 10 ranking radiomic features (total=756) is given in table 4.

4. Discussion

We developed 3 NSCLC classification models. RF-sem semantic features obtained by consensus between two thoracic radiologists from training data and by a separate radiologist, from the validation data. RF-rad was based on computer-aided extraction

of radiomic features from CT images of NSCLCs, whereas RF-all was a combination of semantic and radiomic features. RF-sem performed well on both training and validation data despite both data-sets having been annotated by separate radiologists, indicating the robustness of random forests models developed with semantic features to inter-observer variability. RF-rad and RF-all gave perfect predictions on training data but performed no better than random guess on validation data – indicating a high degree of overfitting of random forests developed using radiomic features.

We found several semantic features highly predictive of NSCLC subtype (table 3), of which air-bronchogram, ground-glass component, cavitation, and satellite nodules ranked highest in terms of discriminatory capability (table 4). Our findings regarding the relative proportions of the various semantic features follow previously reported trends with a few differences [13,30–32]: Several clinical variables including older age, male gender, and smoking history are known to be more frequent in SCCA, in addition to semantic features such as spiculation and central location [32]. In our cohort, none of these variables were significantly different between ADCA and SCCA and did not make a substantial contribution to the classifier.

The most important radiomic features in our study were GLSZM-SZLIE (MDA=0.005), coefficient of variation (MDA=0.004), and NGTDM coarseness (MDA=0.003). The biologic counterparts of these features are poorly understood; here we attempt an intuitive explanation of what these features might represent in tumour CT images: The GLSZM, described originally for texture characterisation of cell nuclei [33], quantifies image heterogeneity in terms of zones of contiguous voxels sharing the same grey level intensity. A relatively homogeneous tumour would have large zones of voxels sharing similar grey level intensity and vice versa. The derived

quantity GLSZM-SZLIE, as the name implies, would be expected to be high in tumours with heterogeneous distribution of low grey-level (e.g., ground-glass density) voxels. NGTDM coarseness, originally tested on various natural (e.g., pebbles, grass) and synthetic materials (e.g., cloth) [34], would be high in tumours exhibiting similar intensities in neighbouring voxels with a low spatial rate of change in voxel intensities. In other words, they would comprise clusters of similar intensity voxels which would stand out against the background and give a 'coarse' appearing texture to the tumour. Coefficient of variation (ratio of standard deviation over mean) is a first-order statistical texture feature which is high in tumours exhibiting high variation in grey-level intensities and low mean intensities. All three features were slightly more common in ADCAs versus SCCAs in our cohort.

A few authors have previously explored radiomics in NSCLC classification: In their proof of concept study, Basu et al. trained a classifier (accuracy: 68%) on CT-derived radiomic features from 74 cases of NSCLC [7]. Their study focused on differentiating the efficacy of 2D radiomic features versus 3D radiomic features and presented a comparison of various model categories including random forests, support vector machines, decision trees, and nearest neighbours. Their best model accuracy of 68% was obtained by employing all 215 features in a leave-one-out cross-validation scheme. However, the authors did not report the best performing variables and a comparison with our radiomic features can therefore not be performed. Two recent studies done by Wu et al. (n=300) and Zhu et al (n=129) have reported higher performance of radiomics-models (AUC 0.72 and 0.9 respectively) [6,8].

Other than that, neither study compared radiomic features with semantic features, the most important difference between our study and either two is that the subset of highest performing radiomic features is different in all three studies. It is possible that

since there are hundreds of radiomic features with majority inter-correlated, some of the different high-ranking features might merely be variations of the same feature. A second possibility is that some of the radiomic models developed by other authors may have overfit, as seen in our study, although Wu et al used an external validation cohort making this unlikely in their study. Overfitting is a common design problem in ML studies, especially in studies with a large number of variables with respect to cases and lack of external validation cohort. Radiomics is doubly challenged in gaining widespread acceptance due to the common use of hundreds of variables and issues surrounding reproducibility, although efforts are underway to standardise radiomics [35].

Our study has several potential limitations: Because this was a CT study, we could not completely eliminate the possibility of including small regions of normal tissue, e.g., opacification due to adjacent atelectasis. However, we minimised such cases by excluding lesions that were difficult to delineate from adjacent collapsed lung. As a result, there may have been an under-representation of centrally located SCCAs because such tumours were frequently inseparable from adjacent atelectasis. Central location is a known feature of SCCAs and including more centrally located tumours, expected to be majority SCCA, may have improved model performance [33]. Secondly, as in most radiomics studies, our original radiomic feature space comprised a large number (n=756) of features derived from CT scans with varying data acquisition parameters, especially those obtained from TCIA. Radiomic features are variable in terms of reproducibility and are dependent on tumour segmentation and image post-processing steps [27]. Hence, we believe that future studies using a more refined selection of radiomic features, especially features engineered specifically for chosen classification tasks, may provide more useful results.

5. Conclusions

Our study showed that non-invasive classification of NSCLCs using semantic features is possible and can be done with good accuracy (AUC: 0.82) using machine learning algorithms. However, CT-scan radiomic features performed poorly on independent validation data (AUC 0.5 and 0.56 for RF-tex and RF-all respectively), despite perfect classification on test data, and may be unsuitable for this task.

6. References

1. Wang H, Xing F, Su H, Stromberg A, Yang L. Novel image markers for non-small cell lung cancer classification and survival prediction. *BMC Bioinformatics* 2014; **15**: 310. doi: https://doi.org/10.1186/1471-2105-15-310
2. Travis WD, Brambilla E, Noguchi M, Nicholson AG, Geisinger K, Yatabe Y, *et al.* Diagnosis of Lung Adenocarcinoma in Resected Specimens: Implications of the 2011 International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society Classification. *Arch Pathol Lab Med* 2012; **137**: 685–705. doi: https://doi.org/10.5858/arpa.2012-0264-RA
3. Sculier J-P, Chansky K, Crowley JJ, Van Meerbeeck J, Goldstraw P. The Impact of Additional Prognostic Factors on Survival and their Relationship with the Anatomical Extent of Disease Expressed by the 6th Edition of the TNM Classification of Malignant Tumors and the Proposals for the 7th Edition. *J Thorac Oncol* 2008; **3**: 457–66. doi: https://doi.org/10.1097/JTO.0b013e31816de2b8
4. Bashir U, Siddique MM, Mclean E, Goh V, Cook GJ. Imaging Heterogeneity in Lung Cancer: Techniques, Applications, and Challenges. *AJR Am J Roentgenol* 2016; **207**: 534–43. doi: https://doi.org/10.2214/AJR.15.15864
5. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* 2015; **278**: 563–77. doi: https://doi.org/10.1148/radiol.2015151169
6. Wu W, Parmar C, Grossmann P, Quackenbush J, Lambin P, Bussink J, *et al.* Exploratory Study to Identify Radiomics Classifiers for Lung Cancer Histology. *Front Oncol* 2016; **6**: 71. doi: https://doi.org/10.3389/fonc.2016.00071

7. Basu S, Hall LO, Goldgof DB, Gu Y, Kumar V, Choi J, *et al.* Developing a Classifier Model for Lung Tumors in CT-Scan Images
8. Zhu X, Dong D, Chen Z, Fang M, Zhang L, Song J, *et al.* Radiomic signature as a diagnostic factor for histologic subtype classification of non-small cell lung cancer. *Eur Radiol* 2018; **28**: 2772–8. doi: https://doi.org/10.1007/s00330-017-5221-1
9. Chalkidou A, O'Doherty MJ, Marsden PK. False Discovery Rates in PET and CT Studies with Texture Features: A Systematic Review. *PloS One* 2015; **10**: e0124165. doi: https://doi.org/10.1371/journal.pone.0124165
10. Sanduleanu S, Woodruff HC, de Jong EEC, van Timmeren JE, Jochems A, Dubois L, *et al.* Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. *Radiother Oncol J Eur Soc Ther Radiol Oncol* 2018; **127**: 349–60. doi: https://doi.org/10.1016/j.radonc.2018.03.033
11. Lim H, Ahn S, Lee KS, Han J, Shim YM, Woo S, *et al.* Persistent pure ground-glass opacity lung nodules ≥ 10 mm in diameter at CT scan: histopathologic comparisons and prognostic implications. *Chest* 2013; **144**: 1291–9. doi: https://doi.org/10.1378/chest.12-2987
12. Lee SM, Park CM, Goo JM, Lee H-J, Wi JY, Kang CH. Invasive Pulmonary Adenocarcinomas versus Preinvasive Lesions Appearing as Ground-Glass Nodules: Differentiation by Using CT Features. *Radiology* 2013; **268**: 265–73. doi: https://doi.org/10.1148/radiol.13120949
13. Kunihiro Y, Kobayashi T, Tanaka N, Matsumoto T, Okada M, Kamiya M, *et al.* High-resolution CT findings of primary lung cancer with cavitation: a comparison between adenocarcinoma and squamous cell carcinoma. *Clin Radiol* 2016; **71**: 1126–31. doi: https://doi.org/10.1016/j.crad.2016.06.110

14. Zhang Y, Qiang JW, Ye JD, Ye XD, Zhang J. High resolution CT in differentiating minimally invasive component in early lung adenocarcinoma. *Lung Cancer* 2014; **84**: 236–41. doi: https://doi.org/10.1016/j.lungcan.2014.02.008
15. Koo HJ, Xu H, Choi C-M, Song JS, Kim HR, Lee JB, *et al.* Preoperative CT Predicting Recurrence of Surgically Resected Adenocarcinoma of the Lung. *Medicine (Baltimore)* 2016; **95**: e2513. doi: https://doi.org/10.1097/MD.0000000000002513
16. Jiang B, Takashima S, Miyake C, Hakucho T, Takahashi Y, Morimoto D, *et al.* Thin-section CT findings in peripheral lung cancer of 3 cm or smaller: are there any characteristic features for predicting tumor histology or do they depend only on tumor size? *Acta Radiol* 2014; **55**: 302–8. doi: https://doi.org/10.1177/0284185113495834
17. Kakinuma R, Kodama K, Yamada K, Yokoyama A, Adachi S, Mori K, *et al.* Performance evaluation of 4 measuring methods of ground-glass opacities for predicting the 5-year relapse-free survival of patients with peripheral nonsmall cell lung cancer: a multicenter study. *J Comput Assist Tomogr* 2008; **32**: 792–8. doi: https://doi.org/10.1097/RCT.0b013e31815688ae
18. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, *et al.* The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J Digit Imaging* 2013; **26**: 1045–57. doi: https://doi.org/10.1007/s10278-013-9622-7
19. Gevaert O, Xu J, Hoang CD, Leung AN, Xu Y, Quon A, *et al.* Non–Small Cell Lung Cancer: Identifying Prognostic Imaging Biomarkers by Leveraging Public Gene Expression Microarray Data—Methods and Preliminary Results. *Radiology* 2012; **264**: 387–96. doi: https://doi.org/10.1148/radiol.12111607

20. Bakr S, Gevaert O, Echegaray S, Ayers K, Zhou M, Shafiq M, *et al.* Data for NSCLC Radiogenomics Collection 2017. doi: https://doi.org/10.7937/K9/TCIA.2017.7hs46erv
21. Aoki T, Tomoda Y, Watanabe H, Nakata H, Kasai T, Hashimoto H, *et al.* Peripheral Lung Adenocarcinoma: Correlation of Thin-Section CT Findings with Histologic Prognostic Factors and Survival. *Radiology* 2001; **220**: 803–9. doi: https://doi.org/10.1148/radiol.2203001701
22. Li F, Sone S, Abe H, MacMahon H, Doi K. Malignant versus Benign Nodules at CT Screening for Lung Cancer: Comparison of Thin-Section CT Findings. *Radiology* 2004; **233**: 793–8. doi: https://doi.org/10.1148/radiol.2333031018
23. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Müller NL, Remy J. Fleischner Society: Glossary of Terms for Thoracic Imaging. *Radiology* 2008; **246**: 697–722. doi: https://doi.org/10.1148/radiol.2462070712
24. Hsu J-S, Han I-T, Tsai T-H, Lin S-F, Jaw T-S, Liu G-C, *et al.* Pleural Tags on CT Scans to Predict Visceral Pleural Invasion of Non–Small Cell Lung Cancer That Does Not Abut the Pleura. *Radiology* 2015; **279**: 590–6. doi: https://doi.org/10.1148/radiol.2015151120
25. Lee SM, Kim S-K, Lim I, Lee H, Hwangbo B, Zo JI. Different diagnostic performance and characteristics of FDG PET/CT between pulmonary adenocarcinoma (ADC) and squamous cell carcinoma (SCC) in lymph node staging. *J Nucl Med* 2008; **49**: 56P-56P
26. Lee E-S, Son D-S, Kim S-H, Lee J, Jo J, Han J, *et al.* Prediction of Recurrence-Free Survival in Postoperative Non–Small Cell Lung Cancer Patients by Using an Integrated Model of Clinical Information and Gene Expression. *Clin Cancer Res* 2008; **14**: 7397–404. doi: https://doi.org/10.1158/1078-0432.CCR-07-4937

27. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, *et al.* User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage* 2006; **31**: 1116–28. doi: https://doi.org/10.1016/j.neuroimage.2006.01.015
28. Breiman L. Random Forests. *Mach Learn* 2001; **45**: 5–32. doi: https://doi.org/10.1023/A:1010933404324
29. R Core Team. *R: A Language and Environment for Statistical Computing*. n.d. URL: http://www.R-project.org/ (Accessed 20 October 2016)
30. Ko JP, Suh J, Ibidapo O, Escalon JG, Li J, Pass H, *et al.* Lung Adenocarcinoma: Correlation of Quantitative CT Findings with Pathologic Findings. *Radiology* 2016; **280**: 931–9
31. Koenigkam Santos M, Muley T, Warth A, de Paula WD, Lederlin M, Schnabel PA, *et al.* Morphological computed tomography features of surgically resectable pulmonary squamous cell carcinomas: Impact on prognosis and comparison with adenocarcinomas. *Eur J Radiol* 2014; **83**: 1275–81. doi: https://doi.org/10.1016/j.ejrad.2014.04.019
32. Onn A, Choe DH, Herbst RS, Correa AM, Munden RF, Truong MT, *et al.* Tumor Cavitation in Stage I Non–Small Cell Lung Cancer: Epidermal Growth Factor Receptor Expression and Prediction of Poor Outcome. *Radiology* 2005; **237**: 342–7. doi: https://doi.org/10.1148/radiol.2371041650
33. Thibault G, Fertil B, Navarro CL, Pereira S, Cau P, Lévy N, *et al.* Texture Indexes and Gray Level Size Zone Matrix. Application to Cell Nuclei Classification. Presented at the Minsk, Belarus

34. Amadasun M, King R. Textural features corresponding to textural properties. *IEEE Trans Syst Man Cybern* 1989; **19**: 1264–74. doi: https://doi.org/10.1109/21.44046
35. Hatt M, Vallieres M, Visvikis D, Zwanenburg A. IBSI: an international community radiomics standardization initiative. *J Nucl Med* 2018; **59**: 287–287

7. Figures

Figure 1 Patient inclusion workflow in our study for training and validation datasets.

Figure 2 Performance curves of RF models on test data (A) and training data (B) show that RF models containing radiomic features (i.e., RF-rad and RF-all) yielded perfect discrimination (AUC 1) on training data (A), but very poor discrimination (AUC 0.52 and 0.56 respectively) on test data, similar to random guess (black line in A and B). RF-sem gave consistent good performance on training (B; AUC 0.78) as well as test data (B; AUC 0.82).

Figure 3 showing two cases of ADCA (A and B), and two of SCCA (C and D). All cases were assigned high probability of respective histologies by the RF-sem model (inset). Among other semantic features these tumours displayed features well known for ADCA, i.e., ground-glass component (arrow in A) and air bronchogram (arrow in B), and for SCCA, i.e., spiculation (arrow in C) and cavitation (arrow in D). Since spiculation was not strongly correlated with SCCA histopathology, the RF-sem model used absence of ADCA-specific features in C, although the overall confidence for SCCA (probability = 75%) was relatively lower.

8. Tables

Table 1. Nodule semantic features and their descriptions

Semantic feature	Description
Air-bronchogram	Presence of visible air-filled bronchi within the lesion. Measured as being present or absent.
Ground-glass component	Presence of hazy attenuation, higher than background, but not sufficiently high to obscure bronchial and vascular margins within the lesion [23].
Location	Central or Peripheral, based on whether the tumour was closer to the hilum than the nearest segmental bronchus or not.
Margins	Irregular, smooth, or lobulated. Lobulation was defined as the presence of at least 3 undulations with a height of more than 2 mm [23].
Pleural indentation	Retraction of pleura near the tumour margin [26].
Satellite nodules	Presence of smaller nodules in the immediate vicinity of the main lesion.
Spiculation	The presence of linear strands at least 2 mm thick extending from tumour margin into adjacent parenchyma [22,23].
Cavitation	Presence of a round lucency inside the lesion, usually within the centre of the lesion and larger than pseudo-cavitation; suggests necrosis [23].

Pseudo-cavitation	Presence of bubble-like areas of low attenuation within the nodule.

Table 2 Clinical and demographic features of patients in training dataset.

Clinical feature	ADCA	SCCA
Age in years, mean (range, SD)	69 (40.2-84.75, 10.2)	70.8(52.35-85.54,8.1)
Sex (M : F)	32 : 32	24 : 18
Smokers	65.6% (n=42)	71.4%(n=30)
T1a	10	7
T1b	12	6
T2a	27	15
T2b	3	5
T3	10	8
T4	2	1
N0	50	35
N1	3	3
N2	11	3
N3	0	1
M0	64	40
M1	0	2

SD=standard deviation

Table 3 Frequencies of semantic features according to tumour type.

	Semantic feature		Tumour type		Fisher's exact test	Interobserver agreement
			ADCA (n=64)	SCCA (n=42)		Weighted-κ (95% CI)
1.	Air-bronchogram	Absent	31 (48.44%)	36 (85.71%)	<0.0001	0.34 (0.16 to 0.52)
		Present	33 (51.56%)	6 (14.29%)		
2.	Airway thickening	Absent	31 (48.44%)	15 (35.71%)	0.2	0.44 (0.25 to 0.63)
		Present	30 (46.88%)	20 (47.62%)		
3.	Emphysema	Absent	24 (37.5%)	10 (23.81%)	0.2	0.78 (0.69 to 0.86)
		Present	20 (31.25%)	16 (38.1%)		
4.	Ground-glass component	Absent	50 (78.13%)	42 (100%)	0.0006	0.74 (0.54 to 0.94)
		Present	14 (21.88%)	0 (0%)		

5.	Location	Central third	20 (31.25%)	10 (23.81%)	0.5	0.35 (0.16 to 0.55)
		Peripheral two-thirds	44 (68.75%)	32 (76.19%)		
6.	Margins	Irregular	35 (54.69%)	22 (52.38%)	0.9	0.2 (0.04 to 0.35)
		Lobulated	27 (42.19%)	18 (42.86%)		
		Smooth	2 (3.13%)	2 (4.76%)		
7.	Pleural indentation	Absent	18 (28.13%)	10 (23.81%)	0.65	0.44 (0.24 to 0.63)
		Present	46 (71.88%)	32 (76.19%)		
8.	Satellite nodules	Absent	50 (78.13%)	41 (97.62%)	0.004	0.74 (0.55 to 0.92)
		Present	14 (21.88%)	1 (2.38%)		
9.	Spiculation	Absent	38 (59.38%)	23 (54.76%)	0.69	0.27 (0.11 to 0.42)

		Present	26 (40.63%)	19 (45.24%)		
10.	Cavitation	Absent	63 (98.44%)	34 (80.95%)	0.002	0.78 (0.57 to 0.99)
		Present	1 (1.56%)	8 (19.05%)		
11.	Pseudo-cavitation	Absent	51 (79.69%)	39 (92.86%)	0.09	0.23 (0.01 to 0.45)
		Present	13 (20.31%)	3 (7.14%)		

IQR = interquartile range, SD = standard deviation

TABLE 4. Variable importance determined by random forests classifier using MDA. A high MDA score of a variable corresponds to greater predictive power.

Variable	MDA
Semantic features	
Air bronchogram	0.039
Ground-glass component	0.023
Cavitation	0.019
Satellite nodules	0.015
Airway thickening	0.008
Pleural indentation	0.006
Emphysema	0.004
Pseudo-cavitation	0.002
Location	-0.002[a]
Spiculation	-0.005
Margin	-0.011
Radiomic features	
db1 LLL GLSZM Short Zone Low Intensity Emphasis	0.005
db1 HLH Coefficient of Variation	0.004
db1 LLL NGTDM Coarseness	0.003
db1 HHH GLCM Cluster Shade	0.003

db1 HHH NGTDM Coarseness	0.003
db1 HHH GLCM Correlation	0.003
NGTDM Contrast	0.003
Maximum intensity	0.003
db1 HHL Coefficient of Variation	0.002

[a]Negative MDA means the variable did not perform better than random chance. MDA=Mean decrease in accuracy. Note: Only the top 10 radiomic features are given here. For full table, please see supplemental file.

List of supplemental material

Supplemental data.docx. This document details image post-processing steps (including a summary of derived radiomic features) and model development procedure.