12-2016

Predicting Malignant Nodules from Screening CT Scans

Samuel Hawkins
University of South Florida

Hua Wang
Key Laboratory of Cancer Prevention and Therapy

Ying Liu
Key Laboratory of Cancer Prevention and Therapy

Alberto Garcia
H. Lee Moffitt Cancer Center and Research Institute

Olya Stringfield
H. Lee Moffitt Cancer Center and Research Institute

See next page for additional authors

HHS Public Access

Author manuscript

J Thorac Oncol. Author manuscript; available in PMC 2017 December 01.

Published in final edited form as:

J Thorac Oncol. 2016 December ; 11(12): 2120–2128. doi:10.1016/j.jtho.2016.07.002.

Predicting malignant nodules from screening CTs

Samuel Hawkins[1], **Hua Wang**[2,3], **Ying Liu**[2,3], **Alberto Garcia**[3], **Olya Stringfield**[3], **Henry Krewer**[1], **Qian Li**[2,3], **Dmitry Cherezov**[1], **Robert A. Gatenby**[4], **Yoganand Balagurunathan**[3], **Dmitry Goldgof**[1], **Matthew B. Schabath**[5], **Lawrence Hall**[1], and **Robert J. Gillies**[3,4]

[1]Dept. Computer Sciences and Engineering, University of South Florida

[2]Department of Radiology, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center of Cancer, Key Laboratory of Cancer Prevention and Therapy

[3]Department of Cancer Imaging and Metabolism, H. Lee Moffitt Cancer Center and Research Institute

[4]Department of Radiology, H. Lee Moffitt Cancer Center and Research Institute

[5]Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute

Abstract

PURPOSE—Determine if quantitative analyses (“radiomics”) of low dose CT lung cancer screening images at baseline can predict subsequent emergence of cancer.

PATIENTS AND METHODS—Public data from the National Lung Screening Trial (ACRIN 6684) were assembled into two cohorts of 104 and 92 patients with screen detected lung cancer (SDLC), then matched to cohorts of 208 and 196 screening subjects with benign pulmonary nodules (bPN). Image features were extracted from each nodule and used to predict the subsequent emergence of cancer.

*Corresponding Author: Robert Gillies, Ph.D, H. Lee Moffitt Cancer Center and Research Institute, 12902 Magnolia Drive, Tampa, FL 33612. Robert.Gillies@Moffitt.org.

DISCLOSURES

Samuel Hawkins reports grants from Moffitt Cancer Center, during the conduct of the study.
Hua Wang reports grants from NIH/NCI U01 CA143062, grants from Florida 2KT01, during the conduct of the study;
Ying Liu has nothing to disclose.
Alberto Garcia has nothing to disclose.
Olya Stringfield has nothing to disclose.
Henry Krewer has nothing to disclose.
Qian Li has nothing to disclose.
Dmitry Cherezov has nothing to disclose.
Robert A. Gatenby has nothing to disclose.
Yoganand Balagurunathan has nothing to disclose.
Dmitry Goldgof has nothing to disclose.
Matthew B. Schabath has nothing to disclose.
Lawrence Hall reports grants from National Cancer Institute, NIH, during the conduct of the study.
Robert J. Gillies reports grants from National Cancer Institute, grants from State of Florida Dept. Health, during the conduct of the study; non-financial support and other from Health Myne, Inc, outside the submitted work;

RESULTS—The best models used 23 stable features in a Random Forest classifier, and could predict nodules that will become cancerous 1 and 2 years hence with accuracies of 80% (AUC 0.83) and 79% (AUC 0.75), respectively. Radiomics outperformed Lung-RADS and volume. McWilliams' risk assessment model was commensurate.

CONCLUSION—Radiomics of lung cancer screening CTs at baseline can be used to assess risk for development of cancer.

Keywords

Screening; Radiomics; Lung Cancer

Introduction

Lung cancer is the leading cause of cancer-related death in the US and worldwide[1,2]. Because of the large number of affected individuals, improvements in diagnosis at an early, potentially curable stage would have a major impact on human health. The National Lung Screening Trial (NLST) compared low-dose computed tomography (CT) to standard chest radiography (CXR) across three annual screens. There were 309 deaths per 100,000 person-years in the CXR group and 247 deaths per 100,000 person-years in the low dose CT group representing a 20% relative reduction in lung cancer mortality within the CT arm compared to CXR[3]. Of the CT-detected lung cancers, 58% had a prior nodule-positive screen that was not determined to be lung cancer (i.e., nodule positive/cancer negative)[4]. An important issue that arose from these studies was the high detection of 4 to 12 mm diameter indeterminate pulmonary nodules (IPNs) that were "suspicious", but not diagnosed as cancer. Of the IPNs, 96.4% were not diagnosed as, or did not develop into, cancers during the screening period or follow-up. Hence, only 3.6% of IPNs were nascent cancers[5]. Overdiagnosis as "suspicious" is harmful because of patient anxiety, and a subsequent work-up or treatment of these cancers can incur unnecessary costs and morbidity for a condition that may pose no threat if not otherwise treated.[5,6]

The present work tests the hypothesis that quantitative image features ("radiomics") can accurately predict whether an IPN at baseline, T0, will subsequently present as a clinically relevant cancer at the first, T1, or second follow-up, T2. Radiomics treats medical images as data that can be mined for information. To investigate the hypothesis, we created demographically-matched cohorts of CT screening subjects with IPNs that did, or did not subsequently develop into cancers. Two hundred and nineteen 3-D image features describing size, shape, location and texture were extracted from segmented volumes and prioritized to develop predictive classifier models that could predict incidence lung cancer after the baseline screen.

Materials and Methods

NLST Study Population

Data and images from the NLST were accessed through the NCI Cancer Data Access System[7]. The study design and main findings of the NLST have been described previously[3]. Briefly, 53,454 current or former smokers between 55 and 74 years of age were enrolled at

33 U.S medical centers. The participants were randomly assigned to the LDCT-arm (26,722 subjects) or CXR-arm and asked to undergo a baseline and two annual follow-up screenings. CT images were downloaded and a trained radiologist (H.W.) identified the nodules of interest and ensured correct matching across annual scans.

Screen-detected lung cancer patients and nodule-positive/cancer-free participants

As described in Schabath et al.[8], we restructured the entire CT arm of the NLST according to screening histories. Based on the NLST protocol, a positive screen was defined as non-calcified nodule ≥ 4 mm in the axial plane or, less commonly, other abnormalities such as adenopathy or pleural effusion. Six different screen-detected lung cancer patient cohorts were defined based on specific sequences of screening results. For this analysis, we focused on two screen-detected lung cancer (SDLC) patient cohorts described in Figure 1. Both patient groups had baseline (T0) positive screens not associated with a lung cancer diagnosis. Individuals in the SDLC cohort 1 had a screen-detected lung cancer at the first follow-up screen (T1); SDLC 2 had positive screens at T0 and T1 and a screen-detected lung cancer at the second follow-up screen (T2), which was approximately two years after the baseline screen. Complications with segmentation as described in Supplemental Table 1 led to there being 85 patients each in SDLC1 and SDLC2.

To compare incidence lung cancer cases to cancer-free (controls) screening participants, we identified two cancer-free cohorts with benign pulmonary nodules [bPN] that were frequency matched 2:1 to the SDLCs on demographic characteristics and risk factors (i.e., age [± 5 years], sex, smoking status, and pack-years smoked [± 5 pack-years]). bPN cohorts -1 and -2 contained 208 and 184 subjects, respectively. Of these, 176 nodules from bPN cohort 1 and 152 nodules from bPN cohort 2 were successfully segmented and hence, available for subsequent radiomic feature extraction. Segmentation complications include calcification, or the nodule being attached to the pleural wall (Supplemental Table 1). Some of these challenges with spiculated and semi-solid nodules can be overcome when better segmentation algorithms are developed. This nested, matched study design minimizes the influence of confounders and risk factors between lung cancer patients and bPN subjects. Full demographic and clinical descriptors of these cohorts are provided in Table 1. The NLST database-specific patient I.D.'s are provided in Supplemental Table 2. At baseline, there was a trend to larger size in the cohort that eventually presented with cancer. The average ± SD of the longest diameters were: 8.06 ± 3.45 mm for bPN1 and 8.6 ± 3.85 mm for bPN2, and 12.07 ± 5.35 mm for SDLC1 and 12.086 ± 9.89 for SDLC2. Although these differences were significant, the multivariate approach increased predictive accuracy.

Target lung nodule identification

Two radiologists reviewed all CT images at both the lung window setting (width, 1500 HU; level, -600 HU) and mediastinal window setting (width, 350 HU; level, 40 HU). The identification of cancerous nodules in SDLC cohorts was based on the tables provided by the NLST with information about the location, size, and histology for those that were resected. Nodule location wasn't always available for bPN cohorts. In these cases, the head radiologist (Y.L.) identified the suspicious 4–12 mm diameter IPN using prior experience. For those cases with multiple lung nodules, any nodule with diameter of more than 4 mm in a lung

window setting was identified. *The locations of all nodules in this study have been made available in the TCIA database* (www.cancerimagingarchive.net). The largest nodule at time 0 was used for feature extraction.

Patients diagnosed with cancer at T1 or T2 were placed into separate cohorts based on their screening history (Figure 1) and their baseline T0 scans were analyzed. Supplemental Table 3, shows 270 prevalent cases of cancer at the first screen, and 196 SDLCs were identified following a prior positive scan, compared to 125 SDLCs following nodule-negative screens, and 44 interval cases diagnosed incidentally before the next screening[8].

Segmentation

Slice numbers of cancerous nodules were provided by NLST and reviewed by radiologists (L.Y., Q.L.), who provided additional anatomical locations for use during segmentation. Nodules were segmented in 3D with our single-click ensemble segmentation approach[9], running on a LuTA platform (Definiens, Munich Germany). NLST provided up to three reconstructions for each time point. The reconstruction chosen by scanner type is found in Supplemental Table 4. Using an automated segmentation algorithm reduces intra-observer variations; however relying on a radiologist to find the nodules means there is inter-observer variation.

Features

There were 219 3-D image features extracted from the baseline scan. A challenge for high-dimensional feature data is over-fitting by having too many features and too few subjects. Hence, there is a need to prioritize features that: 1) aren't redundant, 2) have a large inter-subject biological range, and 3) are stable. In prior work, we studied stability of quantitative features under repeated ("coffee break") scans and found some of the stable features are prognostic and predictive[10,11].

Classifier Modeling

WEKA[12] was used to build and test classifiers. We compared J48, JRIP, Naïve Bayes, support vector machines (SVM), and Random Forest(s). J48 is a decision tree classifier[13]. The confidence factor for error-based pruning was set to 0.25. JRIP is a rule learner[14]. Naïve Bayes[15] is a probabilistic classifier. Support vector machines[16] project the data into a multi-dimensional space to separate classes with a hyper plane. We used libSVM as our implementation of support vector machines[17]. Both linear and radial basis function kernels were used in building a support vector machine. Cost and gamma parameters were tuned on training data with a grid search. The random forests classifier is an ensemble classifier that produces multiple decision trees. The number of decision trees used was 200[18]. When doing cross validation experiments, two filter feature selection methods were run per fold before classification: relief-f[19–21] and correlation-based feature subset selection (CFS)[22]. Relief-f used a ranker search method. CFS used a greedy stepwise search method.

RESULTS

We test the hypothesis that radiomic analyses of screening CTs at baseline can accurately predict which IPNs will subsequently develop into clinical cancers. The workflow of our study is presented in Supplemental Figure 1. According to NCCN[23] and ACR[24] guidelines, the method of choice to distinguish cancerous from benign nodules is to measure nodule growth following a subsequent screening session after 7–12 months: those with significant growth, 1.5mm or greater[24], are classified as cancerous. Figure 2 presents two nodules at baseline and after a subsequent 1-year follow-up screen. Notably, there was nothing obvious to distinguish the benign (upper) from cancerous (lower) IPNs at baseline. Hence, they were both characterized as IPNs in the T0 baseline screen. The radiomics features show a few of the most divergent measures, including relative volume of air spaces and mean attenuation. Notably, baseline volume was larger in the benign nodule in this case.

Feature Stability

We prioritized a set of features from the RIDER data set, which consisted of two non-enhanced CT scans of the same patients taken 15 minutes apart with the same scanner settings. From these analyses, 23 features (Supplemental Table 5) exhibited a concordance correlation coefficient (CCC) of ≥ 0.95[10,11]. The most stable feature category contained nodule size descriptors, where 84% of the features showed concordance ≥ 0.95. Texture features demonstrated lower levels of concordance due to their high dependence on the CT attenuations. Scanner parameter settings such as field of view, which affects pixel size, also affect textures. A histogram of the pixel sizes for each of our cohorts is provided in Supplemental Figure 2, showing that there was a large amount of variability in the data sets. While such variability may not adversely affect a radiologists' ability to provide qualitative assessment, it will likely affect the ability to extract quantitative radiomic data. Further, although the protocol specified a slice thickness of 2.0 mm, it can be seen in Supplemental Figure 3 that the majority were 2.5 mm and above, which also may impact the extraction of radiomic data. Nonstandardized acquisitions are one known limitation in large multi-center trials like the NLST. In radiomics, all these variations add to feature description noise and influence prediction accuracy. Re-interpolation of the data to a fixed voxel size is possible, but generates noise that cannot be compensated[25].

Classifier Models

As presented in Table 2 the best accuracy for predicting development of cancer one year hence (at T1) using baseline scans was 80.1% (AUC = 0.83; FPR = 9%) using a random forests classifier with RIDER prioritized features[26]. We used the Wilcoxon signed-rank test[27] on the results of thirty 10-fold cross validations using the best volume classifier and best all feature classifier. Significance was found at the 0.01 level for our full feature approach with the top classifiers compared to volume for both accuracy and AUC. However, we did try other tests such as the 5x2 fold cross validation followed by an F-test[28] , finding significance at the 0.05 level for only a subset of random seeds. We believe with more data our approach will always be statistically significantly better than volume. Supplemental Table 6 shows full results for cohort 1.

The best shows the top accuracy of 78.7% (AUC = 0.75; FPR = 11%) for predicting development of cancer 2 years hence (at T2). This accuracy was achieved with support vector machines using a radial basis function kernel with RIDER-prioritized features and feature selection with relief-f to find the 10 best features. Using Wilcoxon's signed-rank test[27] on 30 10-fold cross validations showed this result to be better than volume, which had an accuracy of 71.4%, at the 0.01 level. Full results for cohort 2 are in Supplemental Table 7. It is understandable that a prediction further into the future in cohort 2 is not as accurate as cohort 1.

An alternative approach to cross validation is to use one cohort for training and the other for testing. Table 2 also shows the accuracy when training on SDLC and bPN T1 cohorts and testing on SDLC and bPN T2 cohorts. In this case, the best features were RIDER prioritized further sub-selected with Relief-f. These features were then used to build a random forests classifier and the classifier was applied to the previously unseen cohort 2, which yielded a top accuracy of 76.79% (AUC = 0.81; FPR = 18%). This relatively reduced accuracy is expected because the biology's of cancers presenting 1 or 2 years hence are likely different. Using bagging[29] to generate 30 training sets a Wilcoxon test[27] showed significance over using volume alone with the top classifiers for AUC at the 0.01 level. With volume alone, using a JRIP classifier yielded an accuracy of 72.15%. Full results are presented in Supplemental Table 8.

Solidity

There were 58 nodules that were ground glass in appearance, 41 nodules that were semi-solid, and 338 that were solid. Some nodules could not be scored. For a full break down see Supplemental Table 9. Across cohorts, 24 ground glass nodules became cancerous, 27 semi-solid nodules became cancerous, and 85 solid nodules became cancerous. One of the limitations of this study is that nodules that are not solid may take longer than the study period to present as cancer.

Risk Score

The Lung Imaging Reporting and Data System (Lung-RADS) was developed by the American College of Radiology to standardize the screening of CT lung cancer images[24] into categories from benign to cancer., We performed Lung-RADS categorization on T0 images from 58 pre-cancers and 127 benign nodules from SDLC-2 and bPN-2, respectively. Categories 3 and below were labeled as benign and categories 4A and 4B as malignant. The accuracy of lung-RADS in predicting the subsequent development of cancer was 71.4% (Table 3). Another risk score by McWilliams et al. (Brock University cancer prediction equation) utilizes age, sex, family history of cancer, presence of visually detected emphysema, nodule size, solidity, nodule location, number of nodules, and spiculation to generate a probability of cancer[30]. In the McWilliams et al. model, 5% risk is a low probability of developing cancer, intermediate is a 5% to 10% risk, and high is greater than 10%. We applied this model to the same cohort 2 data that were scored for Lung-RADS. We labeled the first two groups as non-cancer and greater than 10% as cancer. The accuracy of this model was 78.9% (Table 3). To extend our radiomics model, we generated a risk score by categorizing individuals based on their probability of belonging to the malignant or

benign group. In our case, we separated low, intermediate low, intermediate high and high risk as quartiles. The results using a random forests classifier on the same data set are shown in Table 3. As shown, the radiomics approach performed very well for extreme phenotypes, with accuracies of 92% and 93% for predicting high and low risk, respectively. Although results in the intermediate groups were more equivocal, at 63–68%, the overall accuracy of automatically extracted features was 80.0%, compared to McWilliams, 78.9%, and Lung-RADS 71.4%. We also compared the radiomics approach to using volume as the only feature, which had an accuracy of 71.8%. Using McNemar's test[31], the radiomics result is significantly better than Lung-RADS, two-tailed p=0.0177, and better than classification with the same models using volume as the only feature, two-tailed p=0.025, but not significantly better than McWilliams, two-tailed p=0.8383. However, the radiomics model has the added benefit to radiologists of being automated after the nodule has been found for segmentation for a given nodule. Figure 3 shows the ROC curves for the McWilliams approach, which has an AUC of 0.67, volume, which has an AUC of 0.74, and the radiomics scoring schema, which has an AUC of 0.87.

Discussion

The long-term vision for this work is to qualify the application of radiomic biomarkers to reduce over-diagnosis and over-treatment of screen- and incidentally-detected lung nodules. It can be envisioned that these radiomics risk scores can be used now to prescribe optimal time for follow up scans for definitive differential diagnosis. Hence, a subject with a low risk score could be scanned less often than one with a high score. The current results show that a subset of radiomic features extracted from indeterminate pulmonary nodules at a *baseline* CT screening scan can be used to predict the subsequent occurrence of cancer or non-cancer with an overall accuracy of 80%. Importantly, this approach has an accuracy >90% when predicting extreme benign and malignant phenotypes; classifications that include more than half of the subjects in this study. Currently, prediction of lung cancer risk in a screening setting is achieved using the Lung-RADS^{TM} system, which classifies risk of cancer from CT scans based on size, solidity and location. Lung-RADs was developed for lung cancer screening by the American College of Radiology, ACR[24]. Although Lung-RADS^{TM} was not used in the NLST for prospective structured reporting, it was recently evaluated in a large retrospective study[32], resulting in a decrease in FPR from 26.6% to 12.8%; hence a significant reduction in the overdiagnosis and overtreatment. Sensitivity of detecting a cancer was lower for Lung-RADS vs. the NLST (84.9% vs. 93.5%) and this did not appreciably improve upon subsequent follow-up scans. In comparison, the radiomics approach herein achieved FPRs of 9% and 11%, sensitivities of 58% and 60% and specificities of 91% and 89% at baseline to predict subsequent cancer 1 and 2 years hence, respectively. The most advanced molecular technique used serum miRNA to achieve a prediction sensitivity of 87% and specificity of 81%.[33] However, most of their sampling was done at the time of diagnosis (50 of 69), so its ability to predict across time is unknown. It should be noted that our case-control design has a 2:1 mixture of bPN to SDLC for training, which leads to a lower FPR, however the training mixture could be changed and so this measurement is reported.

These results must be tempered by acknowledging the limitations to the current study, and areas for improvement exist. The biggest limitations to the current study were cohort sizes, non-standardization of image acquisition, and the lack of clinical or molecular data. Although there were 26,722 subjects in the LDCT arm of the NLST, only 206 of these subjects developed screen-detected lung cancers (SDLC) following a nodule-positive screen. Hence, with these relatively small numbers it is difficult to accommodate co-variates of patient characteristics. We controlled for these by demographic matching the cohorts under study, but this did not allow for analysis of the individual subjects with greater granularity. At baseline, there were a total of 6,921 NLST participants who had nodule-positive/cancer-negative screens, with 6,715 having IPNs that did not develop into lung cancer. Hence, the ratio of non-cancer to eventual-cancer of an IPN at baseline is ~32:1. In our cohort analyses, we compared non-cancer to eventual-cancer at a ratio of 2:1 and hence, there was a false discovery bias emanating from the proportionalities in our study population. While the nested cohort design limits confounding factors, it may also limit extrapolation to the larger NLST population. Regarding non-standard imaging, although exams in the NLST were supposed to be reconstructed to a slice thickness of 2.0 mm, the actual thickness varied from 1 to 5 mm; fields of view (FOV) varied significantly between and within patients (Supplemental Figures 2, 3); and reconstruction kernels are not comparable between manufacturers (Supplemental Table 4). These issues limit the potential power of radiomics. A further limitation is the time required to curate the database, identify the lesions and extract the features. In theory, these could be reduced if the data curation and nodule identification occurred at the time of the primary radiology read, the so-called “Radiology Reading Room of the Future”[25]. Nonetheless, even with these caveats, radiomic-based classifier models and risk assessments exhibited significant power to identify those patients with IPNs at baseline who are most or least, likely to develop cancer. Moving forward, features are being qualified based on their sensitivity to reconstruction kernels and overly sensitive features can be removed during dimensionality reduction. With very large data sets, these can be parsed as co-variates. While it will be preferable to acquire all images with standardized fields of view and reconstruction matrixes, this is proving to be impractical. To accommodate inter-subject differences, pixel sizes can be regularized by interpolation.

To rectify these deficits large databases will be needed. An important opportunity will be the ACR based Lung Cancer Screening Registry (LCSR) to capture screening metadata. Therein will be an opportunity to develop a federated, living database of images and radiomic data so that co-variates and evolving acquisition standards can be accommodated.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding: This work was supported by USPHS Research Grants U01 CA143062 (RJ Gillies, PI) and U24 CA180927 (B Rosen, PI), Cancer Center Support Grant P30 CA076292 (T. Sellers, PI) and the State of Florida Dept. of Health grants 2KT01 and 4KB17.

References

1. American Cancer Society. Cancer Facts & Figures 2015. Atlanta: American Cancer Society; 2015.
2. American Cancer Society. Global Cancer Facts & Figures. 2. Atlanta: American Cancer Society; 2011.
3. Aberle DR, Adams AM, Berg CD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. The New England journal of medicine. 2011; 365:395–409. [PubMed: 21714641]
4. Gopal M, Abdullah SE, Grady JJ, et al. Screening for lung cancer with low-dose computed tomography: a systematic review and meta-analysis of the baseline findings of randomized controlled trials. Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer. 2010; 5:1233–9.
5. Manser R, Lethaby A, Irving LB, et al. Screening for lung cancer. The Cochrane database of systematic reviews. 2013
6. Patz EF Jr, Pinsky P, Gatsonis C, et al. Overdiagnosis in low-dose computed tomography screening for lung cancer. JAMA internal medicine. 2014; 174:269–74. [PubMed: 24322569]
7. National Cancer Institute. Cancer Data Access System. 2016.
8. Schabath MB, Massion PP, Thompson ZJ, et al. Survival of patients with incident lung cancer following screening by computed tomography in the National Lung Screening Trial [abstract]. Cancer Research. 2014; 74:3250.
9. Gu Y, Kumar V, Hall LO, et al. Automated Delineation of Lung Tumors from CT Images Using a Single Click Ensemble Segmentation Approach. Pattern Recognit. 2013; 46:692–702. [PubMed: 23459617]
10. Balagurunathan Y, Kumar V, Gu Y, et al. Test-Retest Reproducibility Analysis of Lung CT Image Features. J Digit Imaging. 2014
11. Balagurunathan Y, Gu Y, Wang H, et al. Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. Translational Oncology. 2014; 7:72–87. [PubMed: 24772210]
12. Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update. SIGKDD Explor.Newsl. 2009; 11:10–18.
13. Quinlan, JR. C4.5: programs for machine learning. San Francisco, CA: Morgan Kaufmann Publishers Inc; 1993.
14. Cohen, WW. Fast Effective Rule Induction. Proceedings of the Twelfth International Conference on Machine Learning; Morgan Kaufmann; 1995. p. 115-123.
15. John, G., Langley, P. Estimating Continuous Distributions in Bayesian Classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence; Morgan Kaufmann; 1995. p. 338-345.
16. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995; 20:273–297.
17. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011; 2:27:1–27:27.
18. Banfield RE, Hall LO, Bowyer KW, et al. A comparison of decision tree ensemble creation techniques. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2007; 29:173–180.
19. Kira, K., Rendell, LA. A Practical Approach to Feature Selection. Ninth International Workshop on Machine Learning; Morgan Kaufmann; 1992. p. 249-256.
20. Kononenko, I. Estimating Attributes: Analysis and Extensions of RELIEF. European Conference on Machine Learning; Springer; 1994. p. 171-182.
21. Robnik-Sikonja, M., Kononenko, I. An adaptation of Relief for attribute estimation in regression. Fourteenth International Conference on Machine Learning; Morgan Kaufmann; 1997. p. 296-304.
22. Hall, MA. Correlation-based Feature Selection for Machine Learning, Computer Science. Hamilton, New Zealand: The Unversity of Waikato; 1999.
23. National Comprehensive Cancer Network. 2016
24. American College of Radiology. Lung CT Screening Reporting and Data System (Lung-RADS™). 2014.

25. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2015:151169.
26. RIDER. The Reference Image Database to Evaluate Therapy Response. 2006.
27. Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bulletin. 1945; 1:80–83.
28. Alpaydm E. Combined 5× 2 cv F test for comparing supervised classification learning algorithms. Neural computation. 1999; 11:1885–1892. [PubMed: 10578036]
29. Breiman L. Bagging predictors. Machine learning. 1996; 24:123–140.
30. McWilliams A, Tammemagi MC, Mayo JR, et al. Probability of Cancer in Pulmonary Nodules Detected on First Screening CT. New England Journal of Medicine. 2013; 369:910–919. [PubMed: 24004118]
31. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. 1947
32. 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; 162:485–91. [PubMed: 25664444]
33. Sozzi G, Boeri M, Rossi M, et al. Clinical utility of a plasma-based miRNA signature classifier within computed tomography lung cancer screening: a correlative MILD trial study. Journal of Clinical Oncology. 2014; 32:768–773. [PubMed: 24419137]

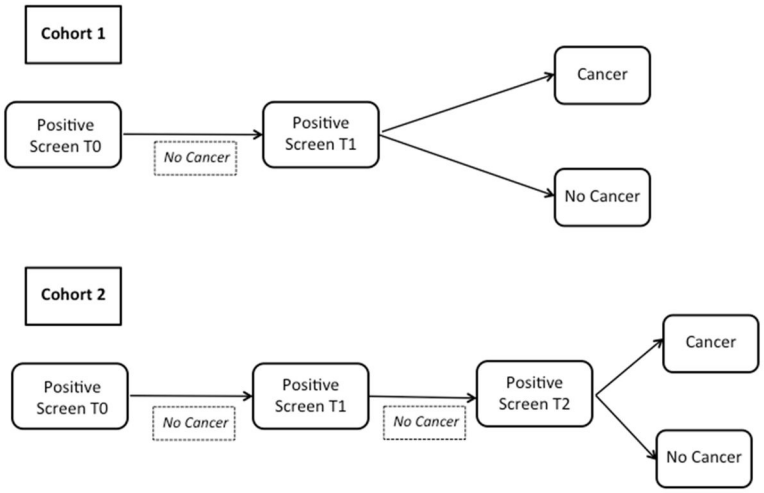

Figure 1. Flowchart of cohorts

Both Cohorts 1 and 2 had a nodule-positive/cancer negative screen at time 0. Cohort 1 had a nodule positive screen at Time 1, of which 104 were diagnosed with a screen-detected lung cancer, SDLC. These were demographically matched to subjects with benign pulmonary nodules, bPN, and the same screening history. 208 bPN-1 were identified, and of these, 176 were successfully segmented. Cohort 2 had a nodule-positive/cancer negative screen at time 1, followed by a nodule- positive screen at Time 2, of which 92 had SDLC. These were demographically matched to 184 bPN subjects, of which 152 were successfully segmented. Segmentation errors are presented in Supplemental Table 1.

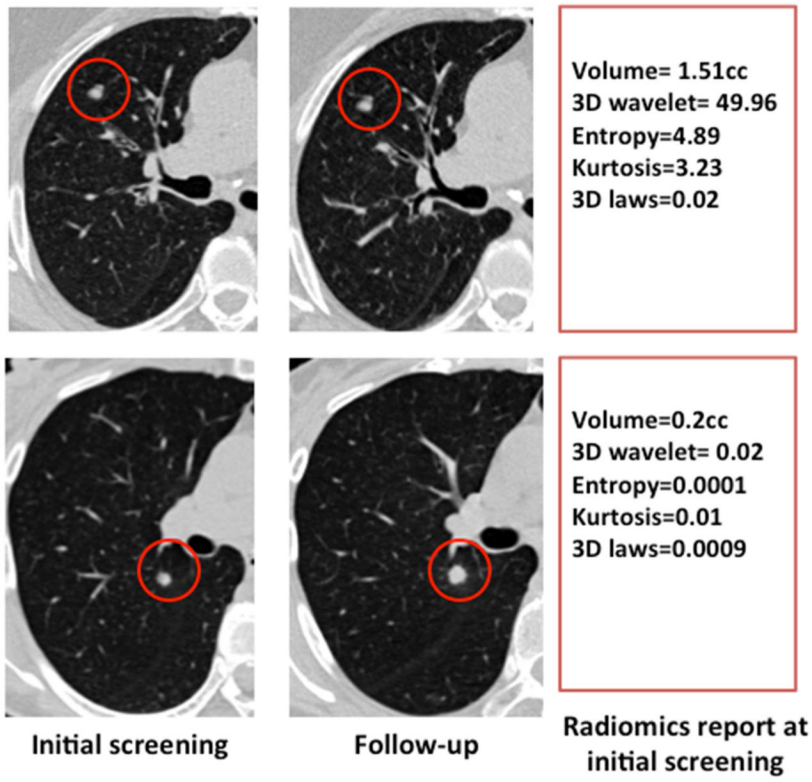

Figure 2. Images from SDLC and bPN at T0 and T1

The top images are from a patient with a benign pulmonary nodule, bPN, in cohort 1. The bottom images are from a patient with a screen-detected lung cancer, SDLC group, in cohort 1. The T0 scans appear similar to the eye, and growth can clearly be seen on the T1 SDLC scan, relative to no growth of the T1 bPN scan. Select radiomic features from the T0 scans that discriminated the groups are shown in the text boxes.

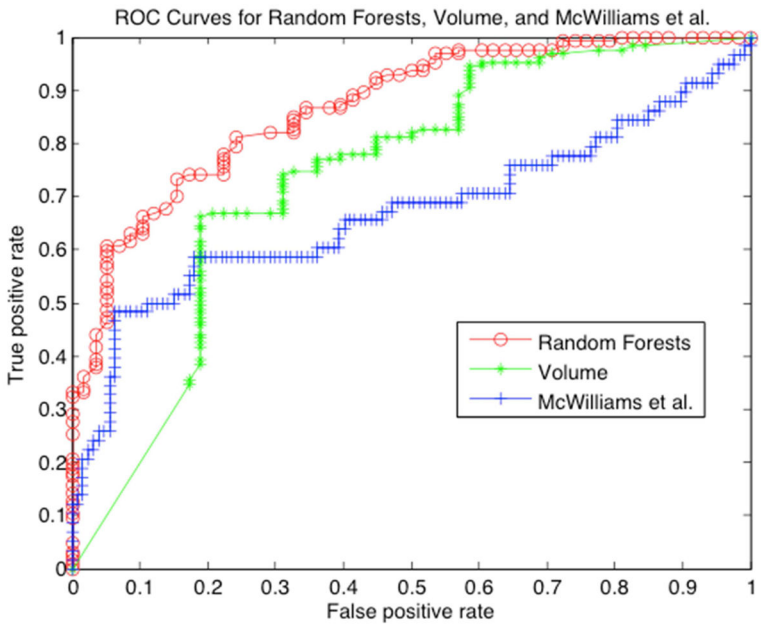

Figure 3. Binary classifier prediction

Receiver operator characteristics (ROC) curves of risk scores for McWilliams, our Random Forests based approach, and volume are shown (see text for details). McWilliams had an area under the ROC of 0.67 and volume had an AUC of 0.74, whereas the radiomics classifier using Random Forests had an AUC of 0.87.

Table 1

Demographics and Clinical Characteristics of NLST Cohort Cases

Characteristic	Lung Cancer Cases (N = 170)	Nodule-Positive Controls (N = 328)	P-value[1]
Age, mean (SD)	63.7 (5.11)	63.5 (5.1)	0.66
Sex, N (%)			
Male	94 (55.3)	192 (58.5)	
Female	76 (44.7)	136 (41.5)	0.28
Race, N (%)			
White	161 (94.7)	315 (96.0)	
Black, Asian, Other	9 (5.3)	13 (4.0)	0.49
Ethnicity, N (%)			
Hispanic or Latino	0 (0.0)	2 (0.6)	
Neither Hispanic/Latino and	170 (100.0)	326 (99.4)	0.55
Unknown			
Smoking, N (%)			
Current	89 (53.4)	175 (53.4)	
Former	81 (47.6)	153 (46.6)	0.85
Pack-Years Smoked, mean (SD)			
Current smokers	63.2 (25.8)	62.0 (21.3)	0.69
Former smokers	64.5 (27.6)	63.7 (26.8)	0.83
Self-Reported History of COPD, N (%)			
Yes	13 (7.6)	19 (5.8)	
No	157 (92.4)	309 (94.2)	0.44
FH of Lung Cancer, N (%)			
Yes	41 (24.1)	56 (17.1)	
No	129 (75.9)	272 (82.9)	0.07
Stage, N (%)			
I	117 (68.8)	--	
II	12 (7.1)	--	
III	21 (12.3)	--	
IV	18 (10.6)	--	
Carcinoid, Unknown	2 (1.2)	--	--
Histology, N (%)			
Adenocarcinoma	108 (63.5)	--	
Squamous cell carcinoma	38 (22.4)	--	
Other, NOS, Unknown	24 (14.1)	--	--

Abbreviations: COPD = chronic obstructive pulmonary disease; FH = Family history;

[1] P-values calculated using Fisher's exact test for categorical variables, Student's t-test for continuous variables

Table 2

Performance of Radiomic based prediction classifier models

The one-year prediction of cohort 1 obtained an accuracy of 80.12% The two-year prediction of cohort 2 obtained and accuracy of 78.78%. Training on cohort one and testing on cohort 2 obtained and accuracy of 76.79%. The top accuracy is bolded. The top accuracy using only volume is also listed.

Cross Validation	Classifier	Feature Subset	Feature Selection	Number of Features	Accuracy	AUC	FPR	TPR	TNR
10x10-Fold on Cohort 1	J48	RIDER Stable	CFS 10	10	76.13%	0.70	0.11	0.5	0.82
	JRIP	RIDER Stable	CFS 5	5	77.82%	0.72	0.1	0.53	0.9
	NB	RIDER Stable	CFS 10	10	79.47%	0.79	0.08	0.53	0.92
	Random Forest	**RIDER Stable**	**None**	**23**	**80.12%**	**0.83**	**0.09**	**0.58**	**0.91**
	SVM- Linear kernel	RIDER Stable	CFS 10	10	79.4%	0.72	0.08	0.52	0.92
	J48	Volume	None	1	75.56%	0.72	0.18	0.62	0.82
10x10- Fold on Cohort 2	J48	All	CFS 5	5	76.89%	0.72	0.09	0.52	0.91
	JRIP	Rider Stable	None	23	76.18%	0.72	0.12		
	NB	All	CFS 10	10	72.88%	0.73	0.08	0.38	0.92
	Random Forest	RIDER Stable	None	23	77.83%	0.83	0.13	0.62	0.87
	SVM-RBF kernel	**RIDER Stable**	**RF 10**	**10**	**78.78%**	**0.75**	**0.11**	**0.6**	**0.89**
	J48	Volume	None	1	71.4%	0.66	0.07	0.32	0.93
Train on Cohort 1 and Test on Cohort 2	J48	NLST Stable	None	37	74.68%	0.62	0.05	0.38	0.95
	JRIP	All	None	219	72.57%	0.66	0.09	0.4	0.91
	NB	NLST Stable	CFS 10	10	73.00%	0.63	0.04	0.32	0.96
	Random Forest	**RIDER Stable**	**RF 10**	**10**	**76.79%**	**0.81**	**0.18**	**0.67**	**0.82**
	SVM-RBF kernel	RIDER Stable	RF 10	10	75.53%	0.73	0.19	0.66	0.81
	JRIP	Volume	None	1	72.15%	0.63	0.05	0.32	0.95

FS – Feature selection, RIDER Stable– is the intersection of the stable features from manual and ensemble segmentations with CCC > 0.95 on the Rider data set, FPR – false positive rate, TPR – True Positive Rate (Sensitivity), TNR – True Negative Rate (Specificity), Top accuracies are in bold

Table 3

Risk Scores for 3 approaches from baseline NLST for predicting subsequent cancer

Lung-RADS obtained a total accuracy of 71.4%. The McWilliams approach obtained an accuracy of 78.9%. The Radiomic approach obtained an accuracy of 80.0%. The Volume Only approach obtained an accuracy of 71.8%.

Model	Category	# Malignant	# Benign	Total	Accuracy	TPR	TNR
Overall	**Total**	**58**	**127**	**185**			
Lung- RADS (71.4%)	**2**	32	99	131	75.6%		
	3	13	20	33	60.6%		
	4A	10	7	17	58.8%		
	4B	3	1	4	75.0%		
	Total	58	127	185	**71.4%**	**22.4 %**	**93.7%**
McWilliams (78.9%)	**Low**	24	98	122	79.7%		
	Intermediate	7	21	28	75.0%		
	High	27	8	35	77.1%		
	Total	58	127	185	**78.9%**	**46.5 %**	**93.7%**
Radiomics (80.0%)	**Low**	6	80	86	93.0%		
	Intermediate-Low	22	38	60	63.3%		
	Intermediate-High	17	8	25	68.0%		
	High	13	1	14	92.9%		
	Total	58	127	185	**80.0%**	**51.7%**	**92.9%**
Volume (71.8%)	**Low**	17	85	102	83.3%		
	Intermediate-Low	13	20	33	60.6%		
	Intermediate-High	5	15	20	25.0%		
	High	23	7	30	76.6%		
	Total	58	127	185	71.8%	48.2%	82.7%

Lung RADS categories include: Benign Appearance and Behavior (2), Probably Benign (3), and Suspicious (4A, 4B). McWilliams categories include: Low, Intermediate, and High Radiomics categories include: Low, Intermediate-Low, Intermediate-High, and High Volume Only categories include: Low, Intermediate-Low, Intermediate-High, and High TPR – True Positive Rate (Sensitivity), TNR – True Negative Rate (Specificity) Top accuracies per method are in bold