Lung Cancer: Genotype Prediction in Computer-Aided Decision Systems

Subjects: Engineering, Biomedical

Contributor: Tania Pereira, Francisco Silva, Joana Sousa, Inês Neves, Joana Morgado, Mafalda Malafaia, Cláudia Freitas, José Luis Costa, Venceslau Hespanhol, António Cunha, Hélder Oliveira

Genotype studies are the fundamental keys in the development of personalized medicine in lung cancer and they enable the progress of targeted therapies. Furthermore, gene analysis allows to identify biomarkers that can be used for early cancer detection, predict the prognosis and the response to the treatment plans, and monitor disease progression.

Keywords: Computer-aided Decision ; Genotype Prediction ; Lung Cancer ; Personalized Medicine

1. Genotype Prediction

Genotype studies are the fundamental keys in the development of personalized medicine in lung cancer and they enable the progress of targeted therapies. Furthermore, gene analysis allows to identify biomarkers that can be used for early cancer detection, predict the prognosis and the response to the treatment plans, and monitor disease progression. The most recent ambition for computer-aided decision systems (CAD) has been to correlate the phenotype captured by the radiological images and determine the associated genotype. Recent studies have focused on predicting the *EGFR* mutation status using CT imaging since targeted therapies for this gene already exist.

In total, twenty studies were found after employing the query (“Gene Mutation Status”) AND (“Prediction”) AND (“Lung Cancer”) in the research databases IEEE Xplore and PubMed, and excluding the ones that were not based on CT scans. These studies included semantic, radiomic, and deep features, which were the inputs of statistical, machine learning, or deep learning models. All of these studies were from 2017 to 2021, which shows how novel the investigation of this area is.

2. Centered on Nodule

Thus far, twelve studies were found, dedicated to explore gene mutation status prediction by CT scan analysis taking into account features related to the nodule. **Table 1** provides detailed information on each work dedicated to genotype prediction using nodule features.

Table 1. Overview of published studies regarding predictive models for gene mutation status based on nodule features (2017–2021).

Authors	Year	Dataset	Methods	Performance Results (%)
Zou et al. [1]	2017	Private (209 patients)	Multivariable analyses	*EGFR:* AUC = 73.7
Cheng et al. [2]	2017	Private (2146 patients)	Weighted mean difference, inverse variance	*EGFR:* OR = 49.0
Li et al. [3]	2018	Private (1010 patients)	Random forest/CNNs	*EGFR:* AUC = 83.4
Koyasu et al. [4]	2019	NSCLC-radiogenomics	XGBoost/random forest	*EGFR:* AUC = 65.9
Wang et al. [5]	2019	Private (844 patients)	CNNs	*EGFR:* AUC = 85.0
Zhao et al. [6]	2019	TCIA and private (879 patients)	3D DenseNet	*EGFR:* AUC = 75.8

Authors	Year	Dataset	Methods	Performance Results (%)
Moreno et al. [7]	2021	NSCLC-radiogenomics	SCAV with ML/CNN	*EGFR:* AUC = 82.0 (CNN) *KRAS:* AUC = 73.9 (CNN)
Zhang et al. [8]	2021	Private (914 patients)	Machine learning (SVM/RF/MLP) Deep learning (SE-CNN/CNN/1D-CNN/AlexNet/Fine-tuned VG16/Fine-tuned VGG19)	*EGFR:* AUC = 91.0 (SE-CNN) AUC = 83.6 (SVM)
Le et al. [9]	2021	NSCLC-radiogenomics	LR / KNN / RF / XGBoost	*EGFR:* ACC = 77.8 *KRAS:* ACC = 83.3
Cheng et al. [10]	2021	Private (670 patients)	Pre-trained 3D DenseNet	*EGFR:* AUC = 76.0 ACC = 72.5 F-score = 71.3
Zhang et al. [11]	2021	Private (134 patients)	Logistic regression	*EGFR:* AUC = 78.0 *KRAS:* AUC = 81.0 *ERBB2:* AUC = 87.0 *TP53:* AUC = 84.0
Han et al. [12]	2021	Private (827 patients)	Logistic Regression	*EGFR:* AUC = 75.8 *ALK:* AUC = 73.9

ACC: Accuracy; AUC: area under the ROC curve; KNN: K-nearest neighbors; LR: logistic regression; MLP: multilayer perceptron; OR: odds ratio; RF: random forest; SCAV: selective class average voting; SE-CNN: squeeze and-excitation convolutional neural network; SVM: support vector machine; XGBoost: extreme gradient boosting.

EGFR is the most relevant oncogene due to the frequency of occurrence and the target therapies available for clinical use. For these reasons, several CADs have been developed for the detection of the mutation status of this gene. Correlations between CT morphological features and the presence of EGFR mutations were studied and showed that the EGFR mutation tended to exist in tumors with part-solid GGO [2]. Approaches based on ML methods were extensively used and showed promising results [1][4]. A different approach was used to predict EGFR mutation status and to extract high-level deep features [3][5][6][8]; a CNN showed the best classification performance with an AUC of 0.85. Few other works were dedicated to identify the mutation statuses of other oncogenes, including KRAS [7][9], ALK [12], or even other genes (ERBB2 receptor tyrosine kinase 2 (ERBB2) and tumor protein 53 (TP53)) [11]. Those predictions were performed using ML-based approaches and considering radiomic features [7][9][11][12]. The best performance results obtained achieved an AUC of 0.81 for KRAS, 0.87 for ERBB2, and 0.84 for TP53 [11].

3. More Comprehensive Approaches

Thus far, seven studies were found that took into account at least one feature related to the structure or disease external to the nodule. **Table 2** presents an overview of each work that used a more comprehensive approach for genotype prediction.

Table 2. Overview of published studies regarding predictive models for gene mutation status based on nodule and extra nodule features (2017–2021).

Authors	Year	Dataset	Methods	Performance Results (%)
Gevaert et al. [13]	2017	Private (186 patients)	Decision Tree	*EGFR:* AUC = 89.0
Cao et al. [14]	2018	Private (156 patients)	Principal component analysis	*EGFR:* TPR = 72.3 TNR = 78.5
Rizzo et al. [15]	2019	Private (122 patients)	Univariate analysis	*EGFR:* AUC = 82.0 *KRAS:* AUC = 67.0
Pinheiro et al. [16]	2019	NSCLC-radiogenomics	Gradient tree boosting	*EGFR:* AUC = 74.6

Authors	Year	Dataset	Methods	Performance Results (%)
Xiong et al. [17]	2019	Private (1010 patients)	ResNet 101	*EGFR:* AUC = 83.8
Silva et al. [18]	2021	LIDC-IDRI NSCLC-radiogenomics	Convolutional autoencoder	*EGFR:* AUC = 68.0
Morgado et al. [19]	2021	NSCLC-radiogenomics	LR, Elastic Net, Linear SVM, RBG SVM, RF, and XGBoost	*EGFR:* AUC = 73.7 (Linear SVM) AUC = 73.3 (Elastic Net) AUC = 72.5 (LR)

AUC: area under the ROC curve; LR: logistic regression; RF: random forest; SVM: support vector machine; TNR: true negative rate; TPR: true positive rate.

A comprehensive approach is based on the combination of information from nodule features, other lung structures, and a possible fusion with clinical data. The use of all this knowledge allows a deep characterization of the pathophysiological changes that occurred, which could benefit the prediction of the mutational status of the oncogenes. Part of the models developed on a more comprehensive analysis employed semantic imaging data annotated by thoracic radiologists that captured extensive regions on the lung and patient conditions, instead of focusing only on the nodule region. These approaches were based on radiological qualitative features [13][14][15]. On the other hand, the features from the CT images can be objective and automatically extracted, such as radiomic or high-level deep features [17][18][19]. Additionally, both types of features (semantic features and the automatically extracted) can be used together by the learning models [16].

The studies that used semantic features combined with the simplest classification models allowed the assessment of the most relevant lung and nodule features for the mutation status prediction. The wild-type status for *EGFR* was predicted by the appearance of emphysema and airway abnormality while the presence of any ground glass component indicates *EGFR* mutations [13]. Moreover, gender, smoking history, emphysema, and diameter in the mediastinal, TDR, and GGO showed statistical differences between the wild-type group and mutated group of *EGFR* [14]. The connection between *EGFR* mutation and internal air bronchogram, pleural retraction, emphysema, and lack of smoking was found [15]. The mixing of nodule-related features with features from other lung structures showed to benefit the *EGFR* mutational status prediction [16].

The *KRAS* mutation status prediction showed non-consensual results even in these more comprehensive studies, and in some studies, this oncogene status was not connected with image features [13][16].

4. Discussion and Future Work: Genotype Prediction

Radiogenomic approaches used to classify the mutation status of oncogenes for lung cancer patients have shown that there are radiomic signatures in CT images that can be used to distinguish mutant from wild-type statuses. Previous studies have also demonstrated that radiological features, corresponding to descriptive features more familiar to radiologists, may be associated with tumor biology. Subsequent studies further demonstrated that the combination of radiomic features and the inclusion of clinical information strengthens the robustness of predictive models. Furthermore, recent studies that have taken into consideration features from a larger region of analysis that contained other structures from the lung appear to have more accurate predictive performances compared to traditional nodule-based approaches. Since lung cancer development is related to multiple physiological changes not restricted to the nodule region, it is expected that the studies that employ comprehensive approaches and consider extra-tumor features from the lung with the tumor obtain a significant increase in predictive performance. It is crucial to highlight these results and further investigate the importance of holistic lung cancer characterization studies, as many complex combinations of morphological, molecular, and genetic alterations occur during lung cancer development that, when taken into account, would allow the development of more accurate predictive models.

The value of image analysis to reveal biological information will not completely replace the need for tissue biopsy or liquid biopsy. However, image-driven studies can provide additional information that is complementary to biopsies. For example, if the biopsy result of a tumor shows *EGFR*-wild type, the result may include false negatives because of intra-tumor heterogeneity. At this time, the learning model can be seen as an alternative validation tool, as CT imaging provides biological information that can describe the genotype and phenotype of the whole tumor and project the biological

information onto each pixel of images to reflect intra-tumor heterogeneity. If it predicts the tumor to be *EGFR*-mutant, clinicians may need to re-biopsy tissues. In addition, predicting mutation status by CT imaging helps people to choose the most suspicious tumor for biopsy if multiple tumors are present in a patient. Finally, the predictive model requires only routinely used CT imaging, which is a non-invasive technique and easy to acquire throughout the course of treatment. The CT scan can be performed multiple times along the treatment plan, allowing the assessment of the treatment response of the patient. Multiple assessments throughout the treatment plan may not be possible to perform by biopsy due to its invasive nature. Therefore, it is worthwhile to develop an image analysis to complement the tissue biopsy and liquid biopsy for more precise systemic treatment and local therapy.

The radiogenomics field presents a small number of publications that are strongly limited by the small sizes of the available databases, which are hardly a good representation of the population affected by lung cancer. In addition, there is a larger number of benign nodules compared to malignant ones in the available public databases, which hinders the ability to extract useful features related to malignant cases only. Furthermore, performance comparisons between models trained and tested with different data do not allow clear and objective conclusions, and image acquisition protocols and performance validation methods (i.e., cross-validation) differ from study to study. Still, direct quantitative comparisons on prediction results are crucial for a clearer understanding of the research evolution, increasing the need for a large and heterogeneous cohort of patients affected by lung cancer, as well as methods capable of coping with data heterogeneity. Accordingly, the sharing of image data among different clinical institutions, but under an uniform protocol to avoid any inconsistency during data record, is valuable to obtain an unique reliable dataset.

Before translation into clinical practice, multisite trials are also needed to validate the results obtained in training cohorts on separate independent groups of patients. Since a model fitting is optimal in the training set used to build the model itself, it is crucial to validate the model in a large external cohort of patients to obtain more reliable fitting estimates. External validation will determine the transportability of the model in different locations consisting of plausibly similar individuals.

Studying the variability amongst radiologists in multi-institutional cohorts is required in the near future to further study the robustness of the annotation of semantic features. Moreover, explainable AI is a field that should be further explored in radiogenomics studies, as it is important not only to consider black-box models but also interpretable models whose predictive decisions can be understood by human observers.

References

1. Zou, J.; Lv, T.; Zhu, S.; Lu, Z.; Shen, Q.; Xia, L.; Wu, J.; Song, Y.; Liu, H. Computed tomography and clinical features associated with epidermal growth factor receptor mutation status in stage I/II lung adenocarcinoma. Thorac. Cancer 2017, 8, 260–270.
2. Cheng, Z.; Shan, F.; Yang, Y.; Shi, Y.; Zhang, Z. CT characteristics of non-small cell lung cancer with epidermal growth factor receptor mutation: A systematic review and meta-analysis. BMC Med. Imaging 2017, 17, 5.
3. Li, X.Y.; Xiong, J.F.; Jia, T.Y.; Shen, T.L.; Hou, R.P.; Zhao, J.; Fu, X.L. Detection of epithelial growth factor receptor (EGFR) mutations on CT images of patients with lung adenocarcinoma using radiomics and/or multi-level residual convolutionary neural networks. J. Thorac. Dis. 2018, 10, 6624–6635.
4. Koyasu, S.; Nishio, M.; Isoda, H.; Nakamoto, Y.; Togashi, K. Usefulness of gradient tree boosting for predicting histological subtype and EGFR mutation status of non-small cell lung cancer on 18F FDG-PET/CT. Ann. Nucl. Med. 2019, 34, 49–57.
5. Wang, S.; Shi, J.; Ye, Z.; Dong, D.; Yu, D.; Zhou, M.; Liu, Y.; Gevaert, O.; Wang, K.; Zhu, Y.; et al. Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. Eur. Respir. J. 2019, 53, 1800986.
6. Zhao, W.; Yang, J.; Ni, B.; Bi, D.; Sun, Y.; Xu, M.; Zhu, X.; Li, C.; Jin, L.; Gao, P.; et al. Toward automatic prediction of EGFR mutation status in pulmonary adenocarcinoma with 3D deep learning. Cancer Med. 2019, 8, 3532–3543.
7. Moreno, S.; Bonfante, M.; Zurek, E.; Cherezov, D.; Goldgof, D.; Hall, L.; Schabath, M. A Radiogenomics Ensemble to Predict EGFR and KRAS Mutations in NSCLC. Tomography 2021, 7, 14.
8. Zhang, B.; Qi, S.; Pan, X.; Li, C.; Yao, Y.; Qian, W.; Guan, Y. Deep CNN Model Using CT Radiomics Feature Mapping Recognizes EGFR Gene Mutation Status of Lung Adenocarcinoma. Front. Oncol. 2021, 10, 1.
9. Le, N.Q.K.; Kha, Q.H.; Nguyen, V.H.; Chen, Y.C.; Cheng, S.J.; Chen, C.Y. Machine Learning-Based Radiomics Signatures for EGFR and KRAS Mutations Prediction in Non-Small-Cell Lung Cancer. Int. J. Mol. Sci. 2021, 22, 9254.

10. Cheng, J.; Liu, J.; Jiang, M.; Yue, H.; Wu, L.; Wang, J. Prediction of Egfr Mutation Status in Lung Adenocarcinoma Using Multi-Source Feature Representations. ICASSP 2021, 1350–1354.
11. Zhang, T.; Xu, Z.; Liu, G.; Jiang, B.; de Bock, G.H.; Groen, H.J.; Vliegenthart, R.; Xie, X. Simultaneous identification of egfr, kras, erbb2, and tp53 mutations in patients with non-small cell lung cancer by machine learning-derived three-dimensional radiomics. Cancers 2021, 13, 1814.
12. Han, X.; Fan, J.; Li, Y.; Cao, Y.; Gu, J.; Jia, X.; Wang, Y.; Shi, H. Value of CT features for predicting EGFR mutations and ALK positivity in patients with lung adenocarcinoma. Sci. Rep. 2021, 11, 5679.
13. Gevaert, O.; Echegaray, S.; Khuong, A.; Hoang, C.D.; Shrager, J.B.; Jensen, K.C.; Berry, G.J.; Guo, H.H.; Lau, C.; Plevritis, S.K.; et al. Predictive radiogenomics modeling of EGFR mutation status in lung cancer. Sci. Rep. 2017, 7, 41674.
14. Cao, Y.; Xu, H.; Liao, M.; Qu, Y.; Xu, L.; Zhu, D.; Wang, B.; Tian, S. Associations between clinical data and computed tomography features in patients with epidermal growth factor receptor mutations in lung adenocarcinoma. Int. J. Clin. Oncol. 2018, 23, 249–257.
15. Rizzo, S.; Raimondi, S.; de Jong, E.E.; van Elmpt, W.; De Piano, F.; Petrella, F.; Bagnardi, V.; Jochems, A.; Bellomi, M.; Dingemans, A.M.; et al. Genomics of non-small cell lung cancer (NSCLC): Association between CT-based imaging features and EGFR and K-RAS mutations in 122 patients—An external validation. Eur. J. Radiol. 2019, 110, 148–155.
16. Pinheiro, G.; Pereira, T.; Dias, C.; Freitas, C.; Hespanhol, V.; Costa, J.L.; Cunha, A.; Oliveira, H.P. Identifying relationships between imaging phenotypes and lung cancer-related mutation status: EGFR and KRAS. Sci. Rep. 2020, 10, 3625.
17. Xiong, J.; Li, X.; Lu, L.; Schwartz, L.H.; Fu, X.; Zhao, J.; Zhao, B. Implementation Strategy of a CNN Model Affects the Performance of CT Assessment of EGFR Mutation Status in Lung Cancer Patients. IEEE Access 2019, 7, 64583–64591.
18. Silva, F.; Pereira, T.; Morgado, J.; Frade, J.; Mendes, J.; Freitas, C.; Negrão, E.; De Lima, B.F.; Da Silva, M.C.; Madureira, A.J.; et al. EGFR Assessment in Lung Cancer CT Images: Analysis of Local and Holistic Regions of Interest Using Deep Unsupervised Transfer Learning. IEEE Access 2021, 9, 58667–58676.
19. Morgado, J.; Pereira, T.; Silva, F.; Freitas, C.; Negrão, E.; de Lima, B.F.; da Silva, M.C.; Madureira, A.J.; Ramos, I.; Hespanhol, V.; et al. Machine Learning and Feature Selection Methods for EGFR Mutation Status Prediction in Lung Cancer. Appl. Sci. 2021, 11, 3273.
