U.PORTO

MESTRADO INTEGRADO EM MEDICINA

Study of the Prevalences and Associations in a Lung Cancer Cohort from Portuguese Population

Inês Maria Rosas da Silva Neves

M

2023

INSTITUTO DE **CIÊNCIAS BIOMÉDICAS ABEL SALAZAR**

Study of the Prevalences and Associations in a Lung Cancer Cohort from Portuguese Population

Inês Maria Rosas da Silva Neves
up201704832@edu.icbas.up.pt

Mestrado Integrado em Medicina

Instituto de Ciências Biomédicas Abel Salazar
Universidade do Porto

Orientadora: Doutora Tânia Maria Pereira Lopes
Investigadora Auxiliar no INESC TEC

Coorientadora: Professora Doutora Maria Manuela França Oliveira
Assistente Graduada em Radiologia
Diretora do Serviço de Radiologia do Centro Hospitalar Universitário de Santo António

Porto, junho de 2023

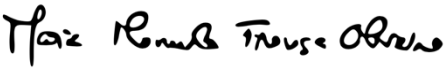
Estudante

(Inês Maria Rosas da Silva Neves)

Orientadora

(Tânia Maria Pereira Lopes)

Coorientadora



(Maria Manuela França Oliveira)

Dedicatória

Gostaria de dedicar este trabalho à minha família pelo apoio incondicional que me deram que me permitiram desfrutar destes seis anos de estudos e ao Delfim por ser o suporte que facilita o meu dia-a-dia, quem me ensinou que para ser feliz nem tudo tenho de controlar. A todos os meus amigos e colegas pelos risos, ajuda e por serem os primeiros a me celebrarem.

A todos estes estou grata, de um modo impossível dizer por palavras.

Agradecimentos

Um agradecimento especial à Doutora Tânia Pereira, minha orientadora, por me ter guiado durante este percurso, sempre com a maior simpatia e disponibilidade. Não há forma de agradecer a transparência e amigabilidade oferecida, sempre num diálogo aberto com muita partilha de conhecimento e sabedoria em diversas áreas. Em conjunto, não poderia deixar de referir também o Professor Doutor Hélder Oliveira o qual do mesmo modo sempre incentivou o pensamento crítico. A ambos, agradeço a incrível oportunidade que foi o estágio extracurricular no grupo Grupo Visual Computing and Machine Intelligence do INESCTEC, onde muito aprendi e que culmina então nesta dissertação de mestrado.

À Professora Doutora Manuela França, minha coorientadora, agradeço também pela disponibilidade e colaboração oferecidas, igualmente num ambiente em que sobressai a simpatia e facilidade de comunicação, que assim facilitou a realização desta dissertação de mestrado.

À Professora Doutora Carolina Lemos, à Drª Cláudia Freitas e ao Doutor Venceslau Hespanhol, que, não sendo orientadores, se disponibilizaram a ajudar-me e levantaram sempre questões pertinentes, elevando o rigor deste trabalho.

Resumo

O Cancro do Pulmão (CP) é a segunda neoplasia mais frequente a nível mundial e a principal cauda de mortes relacionadas com o cancro. De facto, a taxa de sobrevivência a 5 anos estabilizou há várias décadas, o que reflete as dificuldades relacionadas com o diagnóstico, estadiamento e tratamento desta neoplasia. Por este motivo, a patofisiologia e as vias moleculares presentes no CP têm sido francamente estudadas, com a identificação de alvos moleculares específicos, quer oncogenes como o EGFR, o KRAS ou o ALK, quer *checkpoints* imunes como o PD-1/PD-L1 e o CTLA--4. Ainda, as características do nódulo na tomografia computorizada e a sua caracterização têm sido vistos como novos indicadores para a suspeita e diagnóstico precoce do CP. Ademais, o recurso a métodos de Inteligência Artificial baseados nas características clínicas e genéticas está a ser posto em prática com o objetivo de evitar a utilização de técnicas invasivas e de antecipar o diagnóstico desta neoplasia. Contudo, as relações entre as propriedades clínicas, radiológicas e genéticas ainda não são claras, pelo que podem beneficiar da sua análise estatística.

Deste modo, com o objetivo de explorar as prevalências de perfis clínicos e radiológicos dos nódulos e as suas associações com os principais biomarcadores, foi realizado um estudo retrospetivo com um total de 257 doentes com CP do Centro Hospitalar Universitário de São João, selecionados entre 2018 e 2021. As informações clínicas e anatomopatológicas foram atualizada em 2022 e foram realizadas anotações radiológicas para 169 dos 257 doentes selecionados.

Os resultados evidenciam um dataset em que as características mais frequentes era ser do sexo masculino, ter hábitos tabágicos e história de ocupação exposicional. Ainda, 62,9% dos doentes apresentava mutações, sendo o mais comum nos oncogenes EGFR e KRAS. Simultaneamente, 58,8% dos doentes tinha expressão positiva do PD-L1. Os padrões radiológicos mais frequentes dos nódulos era a atenuação sólida, com margens lobuladas ou espiculadas e a presenã de enfisema. Quanto aos resultados de associação, observou-se que as mutações em oncogenes estavam mais relacionadas com particluaridades extra-nódulo, enquanto as características do nódulo se associavam maioritariamente a atributos clínicos do CP. Exemplificando, a mutação do EGFR demonstrou associação com o sexo feminino e fumadores com baixa carga tabágica; o contrário foi evidenciado com a mutação do KRAS que se relacionou com o sexo masculino e fortes hábitos tabágicos. Padrões radiológicos como atenuação sólida ou em vidro despolido, o envolvimento pleural, a presença de calcificações ou necrose nodulares revelaram ainda associação com maior mortalidade.

Em conclusão, foram observadas associação promissoras entre características clínicas, rádiológicas e do *status* mutational oncogénico, que pode providenciar informação crucial para a

identificação de perfis clínicos. Estes, por sua vez, ajudarão os modelos de Inteligência Artificial a criar previsão mais robustas no CP.

Palavras-Chave: Cancro do Pulmão; Biomarcadores; Achados Radiológicos; Tomografia Computorizada; Perfis Clínicos

Abstract

Lung Cancer (LC) is the second most common neoplasm worldwide and still a leading cause of cancer-related deaths. Its minimally changed 5-year survival reflects the difficulties regarding diagnosis, staging and treatment of this disease. Hence, the pathophysiology and molecular pathways of LC have been deeply studied, with the identification of specific molecular targets, such as oncogenes as EGFR, KRAS, or ALK and immune checkpoints as PD-1/PD-L1 and CTLA-4. Additionally, computed tomography nodule findings and its characterization are being viewed as possible novel indicators for early suspicion or diagnosis of LC. Furthermore, the use of medical imaging and Artificial Intelligence-based methods on clinical and genetic information are currently in place with the aim to improve less invasive and earlier cancer diagnosis. Nonetheless, the relationship between clinical, imaging and genetic findings is not yet clear and can benefit from the assessment of their association in LC patients.

With the focus on exploring the prevalence of clinical characteristics and radiologic features of the nodules and their association with the main biomarkers, a retrospective study was performed for a total of 257 patients with LC gathered from the Centro Hospitalar Universitário de São João between 2018 and 2021. Clinical and anatomopathological results were compiled and updated in 2022 and radiological annotations were conducted for 169 of the selected patients.

The results exhibit a dataset where most of the patients were male, with present or past smoking habits and with an history of occupational exposure. Additionally, 62.9% had oncogene mutations, being EGFR and KRAS the most frequent and PD-L1 expression was present in 58.8% of the patients. The most frequent radiologic features were solid attenuation of the nodule, lobulated or spiculated margins and lung emphysema. Driver oncogenes were mainly associated with extra-nodule features while tumor characteristics were largely related to clinical elements of LC. In particular, EGFR was associated with female sex and light smoking versus KRAS that was related to males and smokers. Nodule features such as solid or ground glass attenuation, pleural involvment, presence of calcification and nodule necrosis demonstrated association with higher mortality.

In conclusion, promising associations between LC clinical, radiological and oncogenes mutational status are presented, which can provide crucial information for the identification of clinical profiles. Moreover, these findings can be helpful to guide AI models to create more robust and accurate predictions.

Keywords: Lung Cancer, Biomarkers, Radiological Findings, Computed Tomography, Clinical Profiles

Abreviation List

AI – Artificial Intelligence
CHUSJ – Centro Hospitalar Universitário de São João
COPD – Chronic Obstructive Pulmonary Disease
CT – Computed Tomography
DL – Deep Learning
LC – Lung Cancer
MRI – Magnetic Resonance Imaging
NSCLC – Non-small cell lung cancer
PET-CT – Positron Emission Tomography - Computed Tomography
SCLC – Small cell lung cancer

Index

1. Introduction

Lung Cancer (LC) is the second most common neoplasm worldwide with an estimated 2.21 million new cases in 2020, the first in men and third in women [1]. LC is a leading cause of cancer--related deaths, with an overall 5-year survival rate of 22% that has changed minimally in decades, which demonstrates the difficulties regarding diagnosis, staging and treatment of this disease[2-4]. Although various risk factors have been identified as contributors to the LC pathology mechanism, smoking has by far the strongest association [6,7]. LC occurs more frequently after 40 years of age, with an increasing incidence until the age of 80, which is explained by higher DNA damage and shortening telomeres [4,5]. Younger patients are more likely females, with comorbidities, in whom genetic factors seem to have a stronger contribution [4,5]. As for occupational exposure, the extent of its association with LC remains to be defined, even though the correlation with asbestos, air pollution from the use of unprocessed fossil fuels and radon radiation appears to be higher [4,8]. Previous lung disease, such as chronic bronchitis, emphysema, and tuberculosis have also been associated with an increased risk of lung cancer. In fact, Chronic Obstructive Pulmonary Disease (COPD) is an independent risk factor, mostly in smokers, with a five-time higher risk of LC, particularly in squamous cell carcinoma [4,5]. Although tobacco smoking remains the leading cause of lung cancer worldwide, nowadays 60% of new lung cancers occur in former or never smokers, especially in women[4,5].

Lung carcinomas can be divided into four main cell types: small cell lung cancer (SCLC), adenocarcinoma, squamous cell carcinoma, and large cell carcinoma, although it can be divided into two major histologic subtypes: non-small cell lung cancer (NSCLC) and SCLC. All histologic types of LC can develop in current and former smokers, although squamous and small cell carcinomas are most associated with heavy tobacco use, whereas adenocarcinomas are most common in light to never smokers and young adults[4,8].

The work-up leading to LC diagnosis depends on its initial presentation, either with respiratory, constitutional or extra-pulmonary symptoms, as a paraneoplasic syndrome, or as an incidental lesion during an imaging procedure [4,8]. Upon suspicion, a complete history and physical examination, and imaging tests are required, such as a chest Computed Tomography (CT) scan. Moreover, Positron Emission Tomography - Computed Tomography (PET-CT) and brain Magnetic Resonance Imaging (MRI) can also be useful for the diagnosis and staging of LC [4,8,9]. A tissue sample is required for diagnosis in all patients who are fit for treatment and it is usually obtained by transthoracic CT guided biopsy or bronchoscopy. As an invasive surgical technique, it is not free of complications [10-12]. On the other hand, liquid biopsy is a non-invasive and developing technique that

allows the detection of tumor-derived products in body fluids with a promising contribution to LC diagnosis, treatment, and disease monitoring. However, currently there is no established clinical indication for its use [10,11,13]. Furthermore, new therapies directed to specific molecular targets have been developed over the last decade, improving the prognosis of patients who are not candidates for surgical treatment [14,15]. In NSCLC, one-third of patients exhibit a druggable oncogenic driver mutation, another third exhibits tumor microenvironment excessive inflammation, and the final third of patients are treated with combination chemotherapy [16,17]. Some specific targets already known are oncogenes such as EGFR, KRAS, or ALK and immune checkpoints such as PD-1/ PD-L1 and CTLA-4 [15,18]. In fact, targeted therapies are viewed as the main driver for NSCLC recent mortality decrease [19]. Unfortunately, most patients appear at an advanced stage of the disease where the treatment options are limited [14]. Currently, there is still controversy surrounding LC screening implementation, although low-dose CT scans have been used in high-risk patients with the goal to find a feasible stratification plan for LC screening, but mostly to enhance early detection and decrease mortality [4,9,20,21]. Even more, CT lung findings such as subcentimeter pulmonary nodules and their imaging features are being viewed as possible novel indicators for early suspicion or LC diagnosis and disease prognosis. Solid nodules with spiculated or lobulated margins, pleural indentation and lung emphysema are some of the features with strongest statistical evidence associated with malignancy and poor survival [22-25].

There is an extensive need to understand how clinical and radiologic data are related to different oncogenes expression [26-30]. In fact, various studies have been published about this matter, particularly in breast, prostate, rectal, and LC [31-36]. The imaging data and clinical information will allow the characterization of LC patients and the identification of different clinical profiles but also empower radiogenomics Artificial Intelligence (AI) studies to create more accurate decision support systems to help clinicians. The use of medical imaging and deep learning (DL) has been rapidly evolving in recent years to achieve less invasive, earlier, and more accurate diagnoses. The complexity of the DL models seems to allow to capture of relevant patterns on the medical image to predict the oncogene associated with LC development [37-40]. However, the relationship between some of the main biomarkers, clinical data, and radiologic findings is still not clear and will benefit from the assessment of their relations in LC patients. In this study, information from a retrospective dataset will be combined to understand how the imaging characteristics of the nodule, lung, and health conditions in LC patients are related to the main biomarkers.

2. Material and Methods

2.2 Ethics Approval

The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of the Centro Hospitalar Universitário de São João (CHUSJ) (no. 290/18).

2.2 Study Population and Design

The current work describes and analyses a dataset collected at CHUSJ between 2018 and 2021, where 377 patients with confirmed or suspected LC were selected, being 2016 as the earliest date of LC diagnosis. Exclusion criteria were not having LC and poor clinical records, resulting in the exclusion of 120 patients. Clinical information was added and updated in 2022 for the remaining 257 patients and compiled according to the following variables: age, sex, smoking habits, smoking pack-year, occupational exposures, COPD, lung diseases, previous neoplasm, previous lung neoplasm, LC pathological diagnosis and LC disease stage. The presence of oncogenic driver mutations and PDL-1 status were acquired by consulting the anatomopathological results of tissue biopsy, only available for stage I to III patients.

Radiological annotations of the CT were conducted by three intern radiologists from CHUSJ, with 4 to 5 years of experience, for 169 of the selected patients. The CTs were analyzed in the different planes and annotations were performed in the axial plane. The radiological annotations included nodule location, axial location, attenuation, margins, shape, characteristic of nodule calcification, pleural involvment, nodule cavitation, lung air bronchogram, nodule cysts, nodule necrosis, presence of same lobe as primary nodule satelite lesion, presence of same lung but non--nodule lobe satelite lesion, contralateral lung satelite lesion, lung emphysema and bronchiectasis.

The selection of the features and patients to include in this study are represented in Figure 1.

2.3 Statistical Analysis

IBM SPSS statistical software (version 27) was used for statistical analyses. *P-value < 0.05* was considered statistically significant for the *t-test* and *chi-square test* that were used to compare continuous or categorical variables, as appropriated.

3. Results and Discussion

3.1 Descriptive Analysis

LC is a deadly disease that mainly affects older age males, largely due to their historically higher smoking habits, while women are a smaller proportion of patients that present at a younger age and whose disease pathophysiology is mainly associated with oncogene mutations, particularly in the adenocarcinoma group [4]. In our study, 68% of patients were male and 32% female, which goes in accordance with the literature. However, only in the squamous cell carcinoma group men had a higher median age than women. Also, the 5-year overall survival rate for all stages and types of LC on this dataset was 50%, higher than in a similar study performed in 2009 [5]. Analyzing by histopathological subsets, NSCLC showed 48.6% and SCLC 80% 5-year overall survival rates for all stages, higher than survival rates published recently by LC societies, probably reflecting today's earlier rate of diagnosis and treatment improvements but also disease stage at diagnosis in our group of patients [41].

Regarding exposure, 72.7% are current or former smokers, with a mean pack-year of 37.1. 37.9% had other occupational exposure, with construction workers, maids, or poultry farming being the most frequent.

Although COPD is an independent risk factor for LC [4], only 20% of patients presented with this clinical diagnosis. In addition, 9% of patients had previous tuberculosis infection. Other lung and extra-pulmonary diseases such as asthma, hypersensitivity pneumonitis, HIV infection, and chronic renal or hepatic disease were less frequent, affecting 3-7% of patients. Twenty one percent of patients have had a previous neoplasm, which was LC in 2.7%.

Adenocarcinoma and squamous cell carcinoma were the two major histologic types with 84.8% and 7.8% respectively. In the SCLC group, 66% of nodules were central, while 83% of NSCLC were peripheral. Furthermore, 46.1% of patients were diagnosed at stage IV, followed by stage I with 28.9%. The main characteristics of these two subtypes are represented in Table I.

Regarding the main oncogenes and imunobiomarkers, 62.9% had oncogenes mutations, being EGFR and KRAS the most frequent with 18.8% and 21.9% respectively, which was expected from the literature [14]. Figure 2 represents the oncogene expression distribution in the group of patients with gene mutations and by different clinical characteristics, such as mortality, sex and smoking habits. EGFR patients were mostly women with a low smoking pack-year, even though this group mean age was higher than expected since it was above the general median age of LC, as exemplified in Table II. One explanation is that in this sample, women were older than anticipated, although the difference in age between men and females was not statistically significant. KRAS

patients were mainly men, in a category where smoking was particularly prevalent, as expected [14,42]. Also as anticipated, patients with the ALK mutation were younger, more frequently male and light to never smokers [14,43]. Concerning PD-L1, although its molecular pathway, correlation to patient prognosis and potential as a biomarker have been deeply studied, its expression epidemiology is not as clear. However, it is estimated that PD-L1 overexpression occurs between 20 to 50% of NSCLC [44-46]. In comparison to the findings in this work, where 58.8% had positive PD-L1 expression, in one other study it was negative in the majority of patients [47]. However, the high number of patients in the present work who are male, smokers and/or without EGFR mutations, which are characteristics associated with the overexpression of this immune checkpoint, can be confounders to this matter.

Regarding radiologic features, we observed that attenuation, in which solid and ground glass features are the most suggestive of malignancy [23], 82.4% of nodules appeared solid while only 3.6% were observed as ground glass. Margins also are known to predict malignancy, with poorly defined nodule borders accounting for an increased risk of lung cancer [23]. In fact, 94.5% of nodules had non-smooth margins, either lobulated, spiculated, irregular, or poorly defined. Lobulated and spiculated borders were the most frequent, particularly in the SCLC and NSCLC groups, respectively, as represented in Table I. Pleural involvement is a worse prognosis indicator, reflected by higher staging. Although it is true that SCLC usually presents with large and invasive nodules and a higher stage at diagnosis [22,24], only a third of them had pleura involvement in this dataset. Conversely, in NSCLC 87.2% had either pleural retraction or attachment. It has been described that malignancy is higher when satellite lesions are present in the same lobe as the primary tumor [24]. In accordance, over 50% of nodules had satellite lesions on the same lobe, particularly in the SCLC group, while nodules in the same lung but different lobes were less frequent. In addition, we observed that most SCLC nodules also presented with contralateral ones, that is, at stage IV. Emphysema-predominant COPD phenotype is associated with an increased risk of malignancy and LC [48]. In fact, emphysema was a common characteristic in our group of patients, particularly compromising the whole lung. However, NSCLC also had other forms of emphysema distribution such as affecting only the nodule lobe, but also the lung excluding the nodule lobe.

In this dataset, nodule shape was uniformly distributed across different groups. In addition, cavitation, calcification, necrosis, air bronchogram, fibrosis and bronchiectasias were not frequent nodule features in our dataset, either in SCLC and NSCLC. Regarding this, small evidence has been gathered except for certain patterns of calcification, which are believed to be a benign feature [4,9,48].

3.2 Association Analysis

The association between oncogenes mutation status and radiologic annotations confirmed some of the literature results and their relations were represented in Table II as an * for a *p-value* that shows a statistically significant association (*p-value* < 0.05). Although with differences to what is found in the literature, no other work has studied both nodule and extra-tumoral lung features and their association with oncogenes mutation status. In fact, while one study found that EGFR mutation status is correlated with CT scans imaging tumor phenotypes [38], in this study EGFR had an association with external tumor features. As a matter of fact, in this work, only extra-tumoral lung features were associated with oncogenes mutation status: EGFR with female sex (p <0.001), never to light smoking (p <0.001) and negatively associated with emphysema (p 0.004), and the latter again related to MET mutation (p=0.016); KRAS with male sex (p=0.014) and smoking habits (p <0.001). Also, ALK was related to the absence of pleural findings (p=0.004). No other genes such as HER2, BRAF, PIK3CA, and RET demonstrated significant statistical findings although in the literature most of them are related to the female sex and never to light smoking patterns [48-52]. The lack of these results in this study may be related to the small number of patients within these less frequent mutations.

On the other hand, tumor-related features were frequently associated with clinical characteristics. For instance, higher attenuation nodules were associated with death (p=0.002) and occupational exposure (p=0.033). In an interesting result, although partially solid nodules are at lower risk of malignancy, this attenuation form showed an association with the history of previous lung neoplasm (p=0.018). One explanation can be that the invasive nature of solid nodules may enhance LC recurrence, initially as ground glass tumors that progressively increase their solid component, presenting then as partially solid nodules [53,54].

Spiculated and poorly defined margins displayed an association with existing occupational exposure (p=0.033) and PD-L1 positive expression (p=0.012), while irregular edges with history of a previous lung neoplasm (p=0.038). Lobulated borders in particular showed an association with higher alcohol intake (p=0.033). Studies suggesting alcohol to be an important lung carcinogen have their limitations, either in sample size or by having tobacco use as a confounder, although a recent study observed this association, particularly in the adenocarcinoma histological type [55]. The presenting results can be considered another confirmatory finding regarding this matter.

Tumor necrosis and pleural involvement were linked to advanced disease stages (p <0.001 and p=0.001, respectively) and poor survival (p <0.001 and 0.019, respectively). Furthermore, the latter was also associated to nodules with peripheral and central calcification (p=0.020) and male sex (p=0.016).

In addition, features such as same lobe satellite lesions and tumor cavitation were associated with higher PD-L1 expression ($p < 0.013$ and 0.040, respectively) and the presence of contralateral nodules with higher mortality ($p = 0.038$). Emphysema was also more frequent in males ($p < 0.001$) and in patients with smoking habits ($p < 0.001$), while fibrosis was related to occupational exposures such as working in poultry ($p = 0.034$) [4].

The study of the imaging findings and their relation with the main biomarkers can be useful to guide the development of better AI-based models. For instance, certain biomarkers, such as EGFR mutations, can be associated with specific imaging features on CT scans, which can aid in the accurate identification and classification of lung tumors. By analyzing large datasets of CT images and corresponding biomarker information, AI algorithms can be trained to accurately predict the presence of specific mutations or other molecular abnormalities in lung tumors, which can help guide personalized treatment decisions and improve patient outcomes. The current study tried to identify the main relations between the biomarkers and radiological findings in order to help AI development that can integrate the main causal relations from the imaging findings and disease genesis and progression, and create more robust and accurate predictions. Additionally, the use of AI-based radiogenomic models can also enable the discovery of new imaging features that are predictive of specific biomarkers or disease subtypes, which can aid in the development of new targeted therapies and improve overall patient care.

3.3 Limitations

The major limitation of this study is the small sample size and single-center nature, which may not represent all the variability that can be found in the population of patients with LC. This limitation was particularly noticeable in the number of patients with radiologic annotations. Nevertheless, most of the variables show findings in accordance with what is known in the literature, validating this dataset. As a retrospective study and acknowledging the usual deficits during medical interviews, the statistical analyses of some variables might be undervalued. Occupational exposure and smoking pack year are examples of variables expected to be the most affected. To further validate these results and diminish any confounding factors, a multivariate analysis with a larger sample would be necessary.

4. Conclusions

In conclusion, it is presented a sample that represents the majority of LC patients worldwide, with differences reflecting the Portuguese population and its medical stand-of-care. Furthermore, promising associations were observed between LC clinical, radiological and oncogenes mutational status that can provide crucial information for the identification of clinical profiles. In the future, the aim is to overcome the presented limitations by involving more than one national medical center, thereby increasing the sample size, which will allow further validation and more robust statistical analysis.

5. References

1. Ferlay, J., Colombet, M., Soerjomataram, I. *et al.* Cancer statistics for the year 2020: An overview. *Int. J. Cancer.* 2021. 149(4), 778-789.
2. Kim, H.C., Kim, S.H., Kim, T.J. *et al*. Five-year overall survival and prognostic factors in patients with lung cancer: results from the Korean Association of Lung Cancer Registry (KALC-R) 2015. *Cancer Res Treat: Official Journal of Korean Cancer Association*. 2022. 55(1), 103-111.
3. Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. Cancer statistics, 2022. *CA Cancer J Clin*. 2022. 72(1), 7-33.
4. Horn, L., Lovly, C.M., Johnson, D. *Neoplasms of the Lung*. *In*: McGraw Hill Educational. Harrison's Principles of Internal Medicine. 20th Edition. *McGraw Hill Educational*. 2018.
5. de Groot, P.M., Wu, C.C., Carter,B.W., Munden, R.F. The epidemiology of lung cancer. Transl Lung Cancer Res. 2018. 7(3), 220-233.
6. Durham, A.L., Adcock, I.M. The relationship between COPD and lung cancer. *Lung cancer*. 2015. 90(2), 121-127.
7. Sousa, J., Pereira, T., Neves, I., Silva, F., & Oliveira, H. P. The Influence of a Coherent Annotation and Synthetic Addition of Lung Nodules for Lung Segmentation in CT Scans. *Sensors*. 2022. 22(9), 3443.
8. Kim, J., Lee, H., Huang, B. W. Lung Cancer: Diagnosis, Treatment Principles, and Screening. *Am Fam Physician*. 2022. 105(5), 487-494.
9. Postmus, P.E., Kerr, K.M., Oudkerk, M. *et al*. Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol*. 2017. 28(Suppl. 4), 1-21.
10. Freitas, C., Sousa, C., Machado, F. et al. The role of liquid biopsy in early diagnosis of lung cancer. *Front Oncol*. 2021. 11, 634316.
11. Ganesha, D.V., Naik, R., Mufti, S.S., Varayathu, H. Molecular Therapeutics of Non-Small Cell Lung Cancer (NSCLC) and Challenges in Repeat Tissue Biopsy. *Advances in lung cancer*. 2021. 10(03), 21-39.
12. Richardson, C.M., Pointon, K.S., Manhire, A.R., Macfarlane, J.T. Percutaneous lung biopsies: a survey of UK practice based on 5444 biopsies. *Br J Radiol*. 2022. 75(897), 731-735.
13. Rijavec, E., Coco, S., Genova, C., Rossi, G., Longo, L., Grossi, F. Liquid biopsy in non-small cell lung cancer: highlights and challenges. *Cancers*. 2019. 12(1), 17.
14. Hirsch, F.R., Scagliotti, G.V., Mulshine, J.L. *et al*. Lung cancer: current therapies and new targeted treatments. *Lancet*. 2017. 389(10066), 299-311.
15. Moya-Horno, I., Viteri, S., Karachaliou, N., Rosell, R. Combination of immunotherapy with targeted therapies in advanced non-small cell lung cancer (NSCLC). *Ther Adv Med Oncol*. 2018. 10, 1-12.
16. Pakkala, S., Ramalingam, S.S. Personalized therapy for lung cancer: striking a moving target. *JCI insight*. 2018. 3(15):e120858.
17. Silva, F., Pereira, T., Neves, I. *et al*. Towards machine learning-aided lung cancer clinical routines: Approaches and open challenges. *J. Pers. Med*. 2022. 12(3), 480.
18. Osmani, L., Askin, F., Gabrielson, E., Li, Q.K. Current WHO guidelines and the critical role of immunohistochemical markers in the subclassification of non-small cell lung carcinoma (NSCLC): Moving from targeted therapy to immunotherapy. *In Seminars in cancer biology, Elservier*. 2018. 52, 103-109.
19. Howlader, N., Forjaz, G., Mooradian, M.J. *et al*. The effect of advances in lung-cancer treatment on population mortality. N Engl J Med. 2020. 383(7), 640-649.
20. Lebrett, M.B., Crosbie, E.J., Smith, M.J., Woodward, E.R., Evans, D.G., Crosbie, P.A. Targeting lung cancer screening to individuals at greatest risk: the role of genetic factors. *J Med Genet*. 2021. 58(4), 217-226.

21. Ten Haaf, K., van der Aalst, C.M., de Koning, H.J., Kaaks, R., Tammemägi, M.C. Personalising lung cancer screening: An overview of risk-stratification opportunities and challenges. *Int. J. Cancer*. 2021. 149(2), 250-263.
22. Hu, B., Ren, W., Feng, Z. *et al*. Correlation between CT imaging characteristics and pathological diagnosis for subcentimeter pulmonary nodules. *Thorac Cancer*. 2022. 13(7), 1067-1075.
23. Liu, Y., Wang, H., Li, Q. *et al*. Radiologic features of small pulmonary nodules and lung cancer risk in the National Lung Screening Trial: a nested case-control study. *Radiology*. 2018. 286(1), 298-306.
24. Beigelman-Aubry, C., Dunet, V., Brun, A.L. CT imaging in pre-therapeutic assessment of lung cancer. *Diagn. Interv. Imaging*. 2016. 97(10), 973-989.
25. Lee, G., Lee, H.Y., Park, H. *et al*. Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: state of the art. *Eur. J. Radiol.* 2017. 86, 297-307.
26. Rutman, A. M., Kuo, M.D. Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging. *Eur. J. Radiol.* 2009. 70(2), 232-241.
27. Katabathina, V.S., Marji, H., Khanna, L. *et al*. Decoding genes: Current update on radiogenomics of select abdominal malignancies. *RadioGraphics*. 2020. 40(6), 1600-1626.
28. Lo Gullo, R., Daimiel, I., Morris, E. A., Pinker, K. Combining molecular and imaging metrics in cancer: radiogenomics. *Insights Imaging*. 2020 11, 1-17.
29. Shui, L., Ren, H., Yang, X. *et al.* The era of radiogenomics in precision medicine: an emerging approach to support diagnosis, treatment decisions, and prognostication in oncology. *Front Oncol*. 2021. 10, 570465.
30. Li, S., Zhou, B. A review of radiomics and genomics applications in cancers: the way towards precision medicine. *Radiat. Oncol*. 2022. 17(1), 1-10.
31. Zhu, Y., Li, H., Guo, W. *et al*. Deciphering genomic underpinnings of quantitative MRI-based radiomic phenotypes of invasive breast carcinoma. *Sci. Rep*. 2015. 5(1), 17787.
32. Li, M., Zhang, L., Tang, W. *et al*. Dual-energy spectral CT characteristics in surgically resected lung adenocarcinoma: comparison between Kirsten rat sarcoma viral oncogene mutations and epidermal growth factor receptor mutations. *Cancer Imaging*. 2019. 19(1), 1-11.
33. Singal, G., Miller, P.G., Agarwala, V. *et al*. Association of patient characteristics and tumor genomics with clinical outcomes among patients with non–small cell lung cancer using a clinicogenomic database. *JAMA*. 2019. 321(14), 1391-1399.
34. Jo, S. J., & Kim, S. H. Association between oncogenic RAS mutation and radiologic-pathologic findings in patients with primary rectal cancer. *Quant Imaging Med Surg*. 2018. 9(2), 238.
35. Yip, S.S., Kim, J., Coroller, T.P. *et al*. Associations between somatic mutations and metabolic imaging phenotypes in non–small cell lung cancer. *J. Nucl. Med*. 2017. 58(4), 569-576.
36. Moez, E.K., Pyne, S., Dinu, I. Association between bivariate expression of key oncogenes and metabolic phenotypes of patients with prostate cancer. *Comput. Biol. Med*. 2018. 103, 55-63.
37. Tripathi, S., Moyer, E.J., Augustin, A.I. *et al*. Radgennets: Deep learning-based radiogenomics model for gene mutation prediction in lung cancer. *Inform. Med. Unlocked.* 2022. 33, 101062.
38. Pinheiro, G., Pereira, T., Dias, C. *et al*. Identifying relationships between imaging phenotypes and lung cancer-related mutation status: EGFR and KRAS. *Sci. Rep*. 2020. 10(1), 1-9.
39. Shao, J., Ma, J., Zhang, S. *et al*. Radiogenomic System for Non-Invasive Identification of Multiple Actionable Mutations and PD-L1 Expression in Non-Small Cell Lung Cancer Based on CT Images. *Cancers*. 2022. 14(19), 4823.
40. Wong, C. W., & Chaudhry, A. Radiogenomics of lung cancer. *J Thorac Dis*. 2020. 12(9), 5104.
41. Lung Cancer Early Detection, Diagnosis and Staging. *American Cancer Society*; disponível em: https://www.cancer.org/cancer/types/lung-cancer/detection-diagnosis-staging/survival-rates.html. Consultado pela última vez a 20/04/2023.
42. Salgia, R., Pharaon, R., Mambetsariev, I., Nam, A., Sattler, M. The improbable targeted therapy: KRAS as an emerging target in non-small cell lung cancer (NSCLC). *Cell Rep. Med*. 2021. 2(1):100186 1-15.

43. Golding, B., Luu, A., Jones, R., Viloria-Petit, A.M. The function and therapeutic targeting of anaplastic lymphoma kinase (ALK) in non-small cell lung cancer (NSCLC). *Mol. Cancer.* 2018. 17, 52.
44. Calles, A., Liao, X., Sholl, L.M. *et al*. Expression of PD-1 and its ligands, PD-L1 and PD-L2, in smokers and never smokers with KRAS-mutant lung cancer. *J Thorac Oncol*. 2015. 10(12), 1726-1735.
45. Yang, C.Y., Lin, M.W., Chang, Y.L., Wu, C.T., Yang, P.C. Programmed cell death-ligand 1 expression in surgically resected stage I pulmonary adenocarcinoma and its correlation with driver mutations and clinical outcomes. *Eur. J. Cancer*. 2014. 50(7), 1361-1369.
46.Velcheti, V., Schalper, K.A., Carvajal, D.E. *et al*. Programmed death ligand-1 expression in non-small cell lung cancer. *Lab Invest*. 2014. 94(1), 107-116.
47. Takamochi, K., Hara, K., Hayashi, T. *et al*. Programmed death-ligand 1 expression and its associations with clinicopathological features, prognosis, and driver oncogene alterations in surgically resected lung adenocarcinoma. *Lung Cancer*. 2021. 161, 163-170.
48. MacMahon, H., Naidich, D.P., Goo, J.M. *et al*. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. *Radiology*. 2017. 284(1), 228-243.
49. Pillai, R.N., Behera, M., Berry, L.D. *et al*. HER2 mutations in lung adenocarcinomas: A report from the Lung Cancer Mutation Consortium. *Cancer*. 2017. 123(21), 4099-4105.
50. Cui, G., Liu, D., Li, W. *et al*. A meta-analysis of the association between BRAF mutation and nonsmall cell lung cancer. *Medicine*. 2017. 96(14):e6552.
51. Morris, T. A., Khoo, C., Solomon, B.J. Targeting ROS1 rearrangements in non-small cell lung cancer: crizotinib and newer generation tyrosine kinase inhibitors. *Drugs*. 2019. 79, 1277-1286.
52. Wang, Y., Xu, Y., Wang, X. *et al*. RET fusion in advanced non-small-cell lung cancer and response to cabozantinib: a case report. *Medicine*. 2019. 98(3):e14120.
53. Aoki, T. Growth of pure ground-glass lung nodule detected at computed tomography. *J Thorac Dis*. 2015. 7(9), E326-E328.
54. Zhang, Z., Zhou, L., Yang, F., Li, X. The natural growth history of persistent pulmonary subsolid nodules: Radiology, genetics, and clinical management. *Front. Oncol*. 2022. 12: 1011712.
55. Troche, J.R., Mayne, S.T., Freedman, N.D., Shebl, F.M., Abnet, C.C. The association between alcohol consumption and lung carcinoma by histological subtype. *Am. J. Epidemiol*. 2016. 183(2), 110-121.

6. Figure List

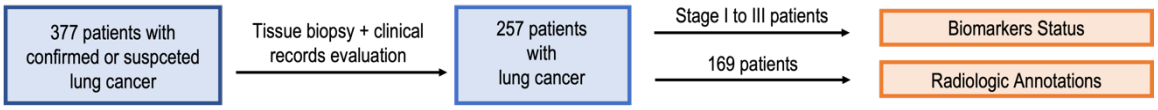


Figure 1. Patient and feature selection from de original dataset.

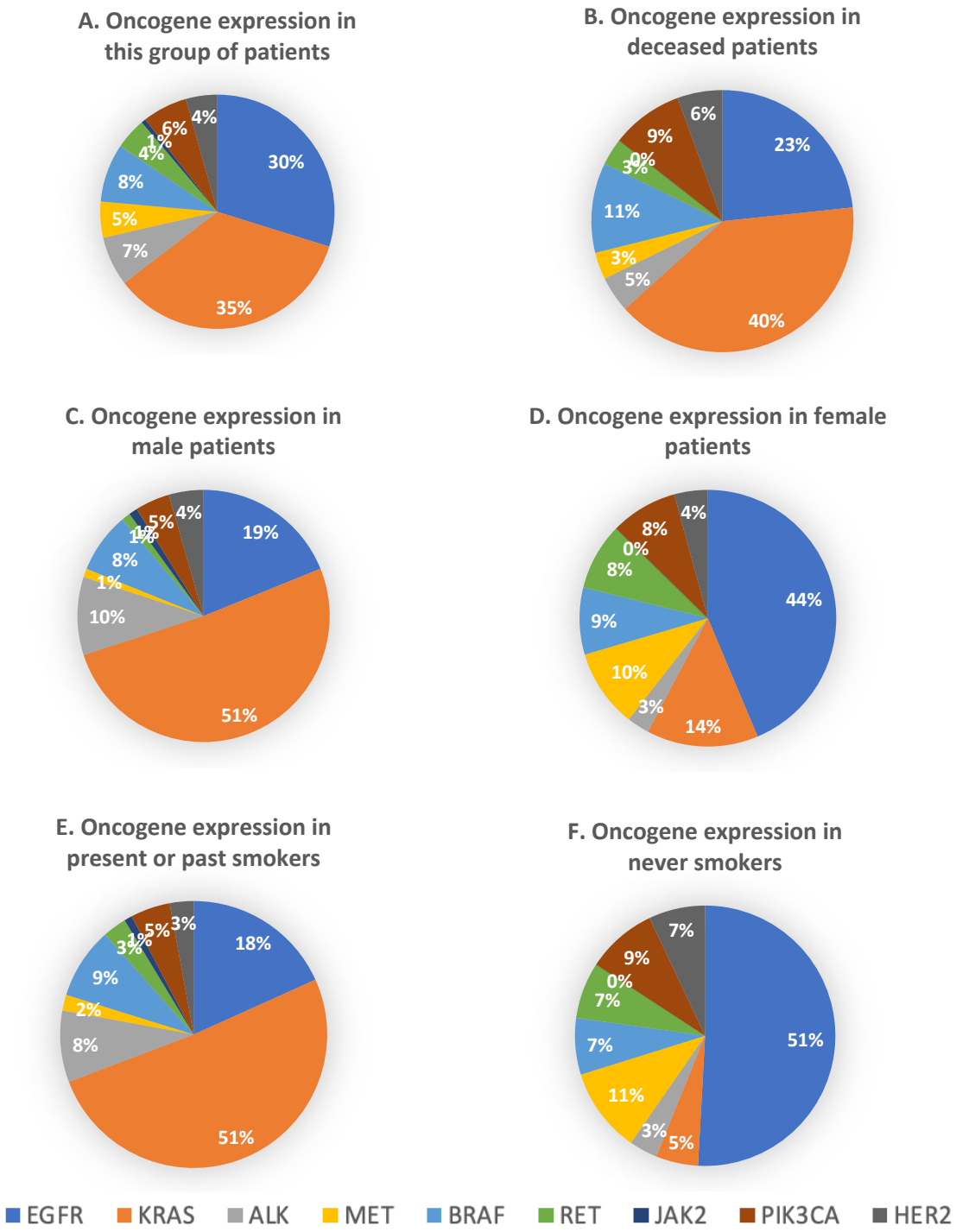


Figure 2. Oncogene expression in the population of patients with gene mutations by different clinical characteristics, expressed in relative frequencies.

7. Table List

Table I. Main clinical and radiologic characteristics by SCLC and NSCLC subtype, expressed in relative frequencies.

		SCLC	NSCLC
sex	male female	100	67,7 32,3
age	median	70	68
smoking habits	yes pack-year	100 51,6	72,1 31,67
overall 5-year survival	%	80	48,6
axial location	central peripheral	66,7 33,3	16 84
attenuation	solid partially solid ground glass	100	82,1 14,2 3,7
margins	irregular poorly defined spiculated lobulated smooth	 33,3 66,7	8,6 8,6 50 27,2 5,6
pleura involment	attachment retraction none	33,3 66,7	32,1 45,1 22,8
same lobe satelite lesion		66,7	50
calcification	peripheral central none	 100	3,7 0,6 100

Table II. Associations between clinical and radiologic features, oncogenes and PD-L1 expression. Results are expressed in absolute frequencies and * demonstrates statistically significant associations, where *p-value* was < 0.05.

		All	EGFR	KRAS	ALK	PD-L1 positive	Death
sex	male	175	17	46 *	9	103	98 *
	female	81	31 *	10	2	47	32
age	median	68,01	69,63	68,02	63,27	67,58	67,94
smoking	yes	186	19 *	53 *	9	115	96
	pack-year	37,12					
emphysema	nodule lobe only	2	1	0	0	1	0
	entire lung	83	6	24	4	53	55
	lung except nodule lobe	14	4	5	0	9	8
	absent	75	20 *	12	3	47	41
axial location	central	28	2	9	1	16	15
	peripheral	137	26	30	6	89	81
same lobe satelite lesion		84	17	26 *	2	47 *	53
pleural findings	attachment	53	6	13	0	34	31 *
	retraction	73	16	16	2	48	50 *
	none	39	6	10	5	23	15
attenuation	solid	136	21	33	5	86	85 *
	partially solid	23	6	4	1	15	6
	ground glass	6	1	2	1	4	5 *
calcification	peripheral	6	1	3	0	5	6 *
	central	1	0	1	0	1	1 *
	none	158	27	35	7	99	90
necrosis		29	4	7	0	21	26 *
margins	irregular	14	3	3	2	10 *	5
	poorly defined	14	2	5	0	11 *	10
	spiculated	82	13	21	3	54 *	53
	lobulated	46	8	8	1	23	23
	smooth	9	2	2	1	7	5