TITLE

AI-HOPE Lung cancer: building a predictive tool for metastatic lung cancer

Code

AI-HOPE

Version: 0.5

Date: 06/12/2024

Coordinating centre: IRCCS Ospedale San Raffaele, S.R.L., via Olgettina 60 - 20132, Milano (MI)

Principal investigator: Dott.ssa Francesca Rita Ogliari

U.O. Oncologia Medica, IRCCS Ospedale San Raffaele, Milano

1 RATIONALE

In the last decade, the treatment of metastatic lung cancer has been continuously evolving, thanks to the approval of innovative therapies such as immunotherapy and molecular-targeted therapy. Based on molecular selection criteria and pathological features (such as the immunohistochemical determination of PD-L1 expression), it is now possible to differentiate subpopulations of patients likely to benefit from a specific treatment: the PD-L1 value is used as a predictor of response to first-line immunotherapy (with a cut-off of 50%), while the identification of some target genes (EGFR, ALK, ROS1, KRAS, BRAF, but also RET, MET, NTRK) guides the therapeutic choice towards gene specific tyrosine kinase inhibitors (1). Despite the development of targeted therapies and the tools available for an accurate patient selection, there still is a considerable proportion of patients who do not benefit from the treatment identified as the best first-line available, resulting in early disease progression. This percentage varies widely depending on the treatment, ranging from 3-10% with most targeted therapies such as anti-EGFR or anti-ALK (2, 3) to 30-40% in real-life studies of first-line mono-immunotherapy (4). In the best-case scenario, early progression imposes a reconsideration of the treatment regimen with a change in therapeutic strategy, but sometimes it can cause a severe deterioration of the performance status, causing early death (5). Several studies have suggested factors that may play a predictive role in response to immunotherapy, especially focusing on molecular characterization and tumor microenvironment, such as the coexistence of multiple genetic alterations (STK11), the proportion of tumor-infiltrating lymphocytes and the burden of somatic mutations in cancer cells (or Tumor Mutation Burden) (6). Last but not least, morphological evaluation using radiomics techniques applied to imaging has shown promising applications in lung oncology, as has metabolic activity studied through PET-FDG (7).

2 OBJECTIVES

The aim of this project is to identify predictive and/or prognostic variables in a group of patients with metastatic lung cancer treated with current standard of care, differentiating them into two cohorts based on the treatment received:

- COHORT A: Stage IV NSCLC patients and high PD-L1 expression (≥50% of tumour cells), treated with first-line mono-immunotherapy

- COHORT B: Stage IV NSCLC patients and PD-L1 expression on less than 50% of tumour cells, treated with a first-line combination of chemo- and immunotherapy (platinum, pemetrexed and immunotherapy for non-squamous histology, carboplatin, paclitaxel and immunotherapy for squamous histology).

Primary Objective: In non-oncogene addicted metastatic NSCLC patients, according to their own first-line standard-of-care, describe the variables (or associations of variables) capable of predicting:

Q1) Early progressions (progression or death within 3 months from treatment start and/or at the first restaging).

Q2) Moderate-severe toxicities (with special focus on pulmonary toxicities).

Q3) Long-term survivors (patients with an overall survival at least 1.5 times longer that the mOS reported in clinical trials, with particular interest in patients progression-free at the end of the 2-year treatment).

3 STUDY DESIGN

The study is structured as an observational, non-interventional, prospective-retrospective, multicenter study.

4 STUDY DURATION

Patients with a diagnosis of metastatic NSCLC treated with the drugs of interest (according to Cohorts) from national approval to 31 December 2023.

5 ELEGIBILITY CRITERIA

Inclusion criteria:

1. Age > 18 years
2. Histological or cytological diagnosis of non-small-cell lung cancer (NB: mixed histologies are includible)
3. Patients that received at least 1 cycle of first-line treatment (according to cohort)

Exclusion criteria:

1. Age < 18 years
2. Histological or cytological diagnosis of small-cell lung cancer
3. Patients lost to follow-up (no follow-up data available)
1. Patients treated with drugs of interest in later lines (>1)

6 SAMPLE SIZE AND STATISTICAL ANALYSIS PLAN

The study aims to develop Machine Learning (ML) algorithms for predictive modeling, emphasizing that determining the exact sample size a priori is challenging due to the exploratory nature of the research. There is a recognized correlation between larger sample sizes and improved algorithm performance, yet no universal method exists to determine the necessary sample size precisely. Multiple techniques and varying hypotheses about the learning approach and statistical model complicate this determination. Post-hoc analysis, such as estimating the learning curve, helps verify the adequacy of the sample size. Empirically accepted rules of thumb suggest a sample size containing at least 10 events per outcome, contingent upon completing a preliminary exploratory phase to identify key factors. Additionally, literature outlines classical statistical methodologies for calculating sample sizes for different outcomes, providing a foundation for ensuring model precision and reliability. The study plans to recruit around 1500 patients, ensuring homogeneity in management and clinical pathways to avoid biases. A K-fold cross-validation approach will be employed for model validation, optimizing predictive performance. Based on the sample, expected SE and confidence intervals are calculated, with ongoing revisions during the data exploration phase to refine sample size and model accuracy. Adjustments will be made to the model's hyperparameters or additional subjects recruited if performance criteria are not met.

Statistical analysis on the study population will be conducted by internal professionals at UniSR. The initial analyses to assess the most significant and clinically relevant variables will be performed with univariate analysis (Wilcoxon and T-Test for continuous variables, Chisquare and Fisher test for categorical variables),

dividing the population based on the question of interest. For multivariate analyses, Machine Learning (ML) approaches such as AutoML (Azure Automated Machine Learning) will be used, which employs various algorithms in combination to find the model that best fits the specific clinical question (expressed as a precision score from 0 to 1, where 1 corresponds to an error rate of 0%). This analytical approach, also known as "black box", identifies the best model for the specific dataset but does not explain the role of each variable within the model itself. To better interpret the emerged model for predictive purposes, other technologies such as SHAP and Decision Tree will be used to explain the weight of each variable and the hierarchy among relevant variables and represent them graphically. Throughout the dataset, these analyses will be conducted for exploratory and descriptive purposes, and the data will be represented with probability graphs and Kaplan-Meier survival curves, where applicable

6 STUDY PROCEDURES

All study procedures will be conducted per clinical practice, no additional procedures will be performed.

7 DATA COLLECTION

For the enrolled patients, the following will be collected:

1. **Clinical data** (demographic information, comorbidities, concomitant medications, performance status, symptoms) before the first therapy cycle;
2. **Clinical course data** (survival, toxicity, additional treatments) at the latest available follow-up visit;
3. **Laboratory data** (complete blood count, liver and kidney function, acute-phase indices) before the first therapy cycle and at three months and/or during the first instrumental reassessment;
4. **Radiological data** (size of the primary tumor, lymph node involvement, sites and number of metastases, etc.) and **radiomic data** (where applicable) before the first therapy cycle and at three months and/or during the first instrumental reassessment;
5. **Nuclear medicine data** acquired with FDG-PET before the first therapy cycle and, if available, at three months and/or during the first instrumental reassessment;
6. **Pathological and molecular biology data** (histology, PD-L1, analysis of EGFR, ALK, ROS1, BRAF, KRAS, RET, MET, NTRK mutational status) before the first therapy cycle.

8 DATA STORAGE

Data will be uploaded on the eCRF platform (RedCap) by authorized researchers and/or data managers. The access to RedCap is subject to a personal user account, which is requested by the PI and protected by a non-transferable password. No data is available for consultation or modification without personal authorized user credentials

9 INFORMED CONSENT, PRIVACY AND ANONYMIZATION

This study will be conducted in accordance with the ethical principles derived from the Declaration of Helsinki and the current regulations on Observational Studies.

With reference to the Retrospective Cohort: The principal investigators at the clinical centers handling already collected data will make every reasonable effort to recontact patients to provide them with appropriate information and obtain their consent for data processing. If, due to specific reasons, informing the individuals proves impossible or entails a disproportionate effort, or if it risks making the achievement of the study's objectives impossible or severely compromised, the investigators will document the impossibility of recontacting the individuals. Only after obtaining the favorable opinion of the competent territorial ethics committee, they will be allowed to process the data (see Ospedale San Raffaele website), adopting all appropriate measures to protect the rights, freedoms, and legitimate interests of the individuals as outlined by the data controller.

With reference to the Prospective Cohort: All patients enrolled in the prospective phase must sign an informed consent form for the use of their data for the purposes of this study.

Data will be collected as pseudo-anonymized for all the centres that have the status of "Istituto di Ricovero e Cura a Carattere Scientifico" (IRCCS) or as fully anonymized for other centres and accordingly to local procedures for data management.

10 DATA PROPERTY

The data and results generated by the study are the property of IRCCS Ospedale San Raffaele.
The study results will be made public in anonymized form through congress presentations and scientific publications. Under no circumstances will data be published that could allow patient identification.
The sponsor/Principal Investigator of the study is responsible for preparing an annual report on the clinical study to be submitted to the Ethics Committee and for preparing a final report on the clinical study. Once the data have been fully analyzed, they will be shared, in anonymized form, with all researchers involved in the study.

REFERENCES

1. Metastatic Non-Small-Cell Lung Cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up, Ann Oncol (2018) 29 (suppl 4): iv192–iv237, https://www.esmo.org/guidelines/lung-and-chest-Tumours/clinical-practice-living-guidelines-metastatic-non-small-cell-lung-cancer

2. Osimertinib in Untreated EGFR-Mutated Advanced Non–Small-Cell Lung Cancer, Soria et al, N Engl J Med 2018; 378:113-125

3. Alectinib versus Crizotinib in Untreated ALK-Positive Non–Small-Cell Lung Cancer, Peters et al, N Engl J Med 2017; 377:829-838

4. First-line pembrolizumab for non–small cell lung cancer patients with PD-L1 ≥50% in a multicenter real-life cohort: The PEMBREIZH study, Amrane et al, Cancer Medicine April 2020; 2309-2316

5. Clarification of Definitions of Hyperprogressive Disease During Immunotherapy for Non-Small Cell Lung Cancer, Kas et al, JAMA Oncol 2020 Jul 1;6(7):1039-1046

6. Predictors of Response, Progression-Free Survival, and Overall Survival in Patients With Lung Cancer Treated With Immune Checkpoint Inhibitors, Memmott et al, JTO July 2021, Pages 1086-1098

7. 18F-FDG Pet Parameters and Radiomics Features Analysis in Advanced NSCLC Treated with Immunotherapy as Predictors of Therapy Response and Survival, Polverari et al, Cancers (Basel) 2020 May 5;12(5):1163