UNIVERSITA' VITA-SALUTE SAN RAFFAELE

CORSO DI DOTTORATO DI RICERCA INTERNAZIONALE
IN MEDICINA MOLECOLARE

Curriculum Oncologia di Base e Applicate

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

Supervisore: Prof. M. Reni Michele Reni

Co-supervisore: Prof.ssa L. Hendriks

Tesi di DOTTORATO di RICERCA di

FRANCESCA RITA OGLIARI

Matr. 021556

Ciclo di dottorato XXXVIII

SSD MED/06

Anno Accademico 2024/2025

CONSULTAZIONE TESI DI DOTTORATO DI RICERCA

Il/la sottoscritto	FRANCESCA RITA OGLIARI
Matricola	021556
Nata a	BUSTO ARSIZIO (VA)
Il	21/11/1989

autore della tesi di dottorato di ricerca dal titolo

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

AUTORIZZA la consultazione della tesi

Data 22/11/2025 Firma *Francesca Rita Ogliari*

DECLARATION

This thesis has been:

- composed by myself and has not been used in any previous application for a degree. Throughout the text I use both '*I*' and '*we*' interchangeably.

- written according to the editing guidelines approved by the University.

Permission to use images and other material covered by copyright has been sought and obtained.

All the results presented here were obtained by myself.

All sources of information are acknowledged by means of reference. In particular, Results presented in Section 3.1 (*Feasibility of ML models in a single-centre experience*) have already been published by the PhD candidate as first author in the paper "*Exploring machine learning tools in a retrospective case-study of patients with metastatic non-small cell lung cancer treated with first-line immunotherapy: A feasibility single-centre experience*" in *Lung Cancer* (doi: 10.1016/j.lungcan.2024.108075).

Acknowledgments

I would like to deeply thank the Artificial Intelligence Lab of Università Vita-Salute San Raffaele for all the analyses performed, especially for building the machine learning models, and for the enlightening discussions, which are the foundation of our strength – bringing together the clinical and data science worlds. In particular, I would like to thank Simone, Patrick, Donato, Daniele, and Alberto. In the Medical Oncology department, a special thanks to Michele, for his all-round support in data curation and project development, and to Alessandra and all the clinical lung cancer team.

I would also like to thank the people I worked with during my time in Maastricht, at the Maastro Clinic and in the Clinical Data Science group. In particular, I would like to mention Dirk de Ruysscher, Cristina Mitea, Mandy Jongbloed and Jarno Huijs, who helped me work with imaging, collect data, and always seek biological explanations.

I am also grateful to the Medical Oncology and Radiology departments of ASST Spedali Civili (Brescia), with whom we shared part of the journey in the search for methodological solutions to radiomics problems. In particular, I would like to thank Salvatore Grisanti and Marco Ravanelli, for their expertise and contribution to the project, and Francesca Piazza and Alessandro Monti, for their meticulous data curation.

Last but not least, my deepest gratitude goes to my supervision team. Thank you, Michele, for guiding me through every tough decision and setting the standards for each step of my research, and thank you, Lizza, for your invaluable mentorship that goes far beyond the scope of this project.

Dedication

Ad Adriano, e alla nostra famiglia.

Abstract (in italiano)

Il carcinoma polmonare non a piccole cellule (NSCLC) metastatico è spesso una malattia eterogenea con sopravvivenze molto variabili, per cui c'è grande necessità di strumenti prognostici e predittivi affidabili, soprattutto nel contesto di *real-world*. Lo studio AI-HOPE è uno studio ambispetttico e multicentrico, con lo scopo di raccogliere dati di pratica clinica e sviluppare modelli di *machine learning* (ML) in grado di predire la sopravvivenza dei pazienti con NSCLC metastatico trattati con immunoterapia di prima linea (con o senza chemioterapia).

Nello studio, i dati vengono raccolti da diverse fonti intra-ospedaliere, come *case-report forms* elettroniche, sistemi di laboratorio e archivi di imaging. Lo studio si basa su una piattaforma sicura e conforme ai requisiti di privacy (*S-RACE platform*) per la condivisione e l'analisi dei dati.

In una prima coorte interna, i modelli Auto-ML hanno raggiunto buone performance nella predizione della progressione precoce (AUC fino a 0.82), con biomarcatori di laboratorio (percentuale di neutrofili, rapporto neutrofili/linfociti, piastrine) ed espressione di PD-L1 come predittori più forti, seguiti da fattori clinici (performance status, uso di steroidi). Al contrario, i modelli che includono *features* di radiomica non hanno migliorato i risultati dei modelli di classificazione binaria, richiedendo inoltre una massiccia supervisione manuale. Approcci *unsupervised* (come il clustering basato esclusivamente su radiomica) sembrano utili per identificare gruppi di tumori con caratteristiche biologiche simili, supportando una loro possibile utilità in studi esplorativi futuri. Infine, diverse librerie di modelli time-to-event sono state testate sulla coorte multicentrica preliminare di 498 pazienti. I modelli RandomForest SurvivalAnalysis e XGBSE hanno superato i modelli transformer in termini di C-index quando applicati ai soli dati tabulari, ma tutti gli algoritmi si sono dimostrati sufficientemente robusti per ulteriori esperimenti con dataset più ampi e fonte dati eterogenee.

Questi risultati confermano la fattibilità di sviluppare modelli di ML riproducibili in un contesto ospedaliero di *real-world*. Una raccolta dati più ampia e prospettica è in corso per ulteriore sviluppo e validazione di modelli multimodali time-to-event.

Abstract (English language)

Metastatic non-small-cell lung cancer (NSCLC) represents a heterogeneous disease with markedly variable outcomes, and reliable prognostic and predictive tools are urgently needed, particularly in the real-world setting. The AI-HOPE study was designed as an ambispective, multicentre project to collect real-world data and develop machine-learning models capable of predicting clinical endpoints in patients with metastatic NSCLC receiving first-line immunotherapy with or without chemotherapy.

In this study, we acquired real-world data from multiple in-hospital sources, including structured electronic case report forms, laboratory systems, and imaging repositories. The study relied on a secure and privacy-compliant platform (S-RACE) for data sharing and analysis, which enabled automated data ingestion and ML model training.

In the single-centre cohort, Auto-ML supervised models achieved robust performance for early-progression prediction (AUC up to 0.82), with laboratory biomarkers (neutrophil percentage, neutrophil/lymphocyte ratio, platelets) and PD-L1 expression emerging as the strongest predictors, followed by clinical factors (performance status, steroid use). Conversely, radiomics-augmented models did not provide added predictive value in prognostic models, despite requiring substantial manual oversight. Unsupervised clustering based solely on radiomic features revealed biologically plausible patient groups, suggesting potential complementary value for future exploratory studies. Moreover, different libraries of time-to-event models were tested on the preliminary multicentre cohort of 498 patients. RandomForest SurvivalAnalysis and XGBSE-based models outperformed transformer-based models in terms of C-index when built on tabular data only, but all the algorithms proved robust enough for further development with larger sample size and multimodal views.

These findings confirm the feasibility of developing reproducible ML models in a real-world hospital environment, and highlight the challenges of imaging-based biomarkers. Larger and prospective data collection is ongoing for future validation of multimodal time-to-event models.

Table of contents

Chapter 1: Background

1.1 Metastatic non-small-cell lung cancer (NSCLC): the clinical scenario

1.1.1 Epidemiology of NSCLC

Lung cancer is the most commonly diagnosed cancer in 2022 (2.5 million new cases according to GLOBOCAN), and the first cause of cancer-related death worldwide (18.7%)[1]. The IARC (International Agency for Research on Cancer) states that lung cancer incidence will further increase in the next decades, reaching an incidence of 58.8% and a mortality rate of 64% in 2040[2]. The most important risk factor is tobacco smoking, which increases the risk of lung cancer 10 to 30 times compared to never-smokers[3]. The global decrease in tobacco use can partially explain the reduced incidence and mortality in men[4], but other inhalants such as e-cigarettes, air pollution, occupational exposure, and radon might explain the rise of lung cancer in never-tobacco-smokers[5]. Women with lung cancer, in particular, tend to be younger and more often never-smoker compared to men, which partially justifies the sex-based disparities in incidence, mortality and treatment-response observed in upper-middle income countries such as Europe and the United States (US)[5]. However, the median age at diagnosis is still over 70 years, and only less than 10% of cases occur in subjects under 55 years of age[6]. Familiarity and/or the presence of germline variants in lung-related oncogenes (such as TP53, ATM, EGFR) have to be considered additional risk factors[7], but solid literature evidence from large scale genetic screening for lung cancer is still lacking. On the other hand, screening programs have been developed in the last decades for heavy smokers (generally ≥20 pack-years and age ≥45 years)[8], proving the efficacy of a low-dose computed-tomography (CT) scan in reducing lung-cancer specific mortality[9]. Despite these results, active and enrolling nationwide lung cancer screening programs are not available in most countries, and the ongoing ones often rely on academic funding. In this context, the role of screening in never-smokers is still controversial[5]. In an international cohort, it has been demonstrated that air pollution (particulate matters measuring ≤2.5 μm) can promote lung cancer by unblocking cells with pre-existing oncogenic mutations in never-smokers[10], but public health initiatives addressing this problem are suboptimal at the current status. Therefore, because of a general lack of lung cancer screening programs, the relatively small population targeted by these programs, and the fact that lung cancer is usually asymptomatic when early-stage, approximately half of the patients are diagnosed with

advanced stage lung cancer[11], when local treatment such as radical surgery are not an option anymore, and systemic treatments are the gold standard to improve long-term survival.

Lung cancer is usually classified in two main histological types: small-cell (SCLC) and non-small-cell lung cancer (NSCLC), with NSCLC being the most common one (85% of all cases)[12]. Within NSCLCs, there are at least three different subtypes: adenocarcinoma (glandular components), squamous cell carcinoma (derived from the squamous epithelium) and large cell carcinoma (less differentiated epithelium-derived cells), with adenocarcinoma being the most common one (50% of cases) and squamous cell carcinoma ranking second (20-30%)[5]. To note, correct histological classification is crucial for further treatment implications, as we know that, besides smoking status, especially the histological subtype has an impact on expected occurrence of actionable genomic alterations (AGA)[13].

Once there is a suspicion of lung cancer, radiological exams should be performed: the minimal amount of imaging recommended by international guidelines is a contrast-enhanced CT-scan of the chest and the upper abdomen. Additionally, in most disease stages, brain imaging is recommended to exclude asymptomatic central nervous system metastases (preferably magnetic resonance imaging (MRI), but often CT-scan due to limited resources)[14]. For all patients potentially eligible for curative intent treatment, an 18F-FDG-PET should be performed[14]. In addition, it is often also performed for patients with advanced disease thanks to its sensitivity to detect occult distant metastases, especially bone metastases.

As stated before, most of the patients are diagnosed with advanced disease after this diagnostic work-up. The staging system currently in use is the 9th edition of the American Joint Committee on Cancer (AJCC)–Union for International Cancer Control (UICC)[15], and it is composed by three parameters: T stage (tumor size), N stage (nodal involvement) and M stage (extent of metastases). The correct interpretation of these three components on baseline imaging scans allows clinicians to classify NSCLC into four different stages, from I (very early disease) to IV (metastatic disease), with clear repercussions on prognosis and treatment plans[15], see *Table 1* for details.

T	Category description	N0	N1	N2a	N2b	N3
T1a	≤1 cm, confined, no invasion	IA1	IIA	IIB	IIIA	IIIB
T1b	>1–2 cm	IA2	IIA	IIB	IIIA	IIIB
T1c	>2–3 cm	IA3	IIA	IIB	IIIA	IIIB
T2a	>3–4 cm OR main bronchus ≥2 cm from carina, visceral pleura invasion, partial atelectasis	IB	IIB	IIIA	IIIB	IIIB
T2b	>4–5 cm	IIA	IIB	IIIA	IIIB	IIIB
T3	>5–7 cm OR chest wall/phrenic nerve/parietal pleura invasion OR separate nodules same lobe	IIB	IIIA	IIIA	IIIB	IIIC
T4	>7 cm OR invasion mediastinum/diaphragm/heart/great vessels/trachea/esophagus/vertebra OR nodules in different ipsilateral lobe	IIIA	IIIA	IIIB	IIIB	IIIC
M1a	Separate nodule contralateral lung OR pleural/pericardial nodules OR malignant effusion	IVA	IVA	IVA	IVA	IVA
M1b	Single extrathoracic metastasis in one organ	IVA	IVA	IVA	IVA	IVA
M1c1	Multiple metastases in one organ system	IVB	IVB	IVB	IVB	IVB
M1c2	Multiple metastases in >1 organ system	IVB	IVB	IVB	IVB	IVB

Table 1. 9th edition of the AJCC Staging system, N categories as follows: N0 – No regional node metastasis, N1 – Ipsilateral peribronchial/hilar/intrapulmonary nodes, N2a – Single

ipsilateral mediastinal or subcarinal station, N2b – Multiple ipsilateral mediastinal/subcarinal stations, N3 – Contralateral mediastinal/hilar or supraclavicular/scalene nodes

Generally, the prognosis of patients with stage IV NSCLC is poor, ranging from 31% of patients alive 1 year after the diagnosis to less than 6% after 5 years in a real-world US cancer registry[16]. Systemic treatments, such as chemotherapy, immunotherapy or targeted therapies, are considered the gold standard for patients with stage IV NSCLC and should be administered to improve their survival, if considered clinically feasible. In the aforementioned cancer registry, it is estimated that around 1 patient out of 3 does not receive active treatment for metastatic disease, and this proportion increases to 38% if we consider only patients older than 65 years[16], mainly because of poor performance status (PS) at diagnosis.

The identification of AGA marked a fundamental change in the treatment of advanced NSCLC, because the administration of targeted therapies in this setting has led to a great improvement in patients' outcomes[17]. Nevertheless, only 30% of NSCLC harbor AGA[18], and the access to targeted therapies is not standardized across countries, with different availabilities in the US, in Europe and also among members of the European Union[5]. Tyrosine-kinase inhibitors targeting KRAS-mutations or MET exon 14 skipping, for instance, are available in Europe only from second-line only, expanding the number of patients that are offered non-targeted treatments in the first-line setting[5], which usually encompass immunotherapy ± chemotherapy according to other cancer-related biomarkers.

1.1.2 First-line treatment choice and rationale for immunotherapy

For non-AGA metastatic NSCLC (or for AGA without first-line approved therapy), the standard first-line treatment has considerably evolved in the past decade thanks to the introduction of immunotherapy. Immune-checkpoint inhibitors (ICI) are monoclonal antibodies that target inhibitory checkpoint molecules expressed on immune cells (mainly T-cells) or on antigen-presenting cells (APCs) and tumor cells[19], regulating their interaction. The most exploited target of this interaction is the PD-1/PD-L1 (programmed-death/programmed-death ligand) axis, which activates a downstream signal resulting in an inhibitory effect on cytotoxic T-cells[19]. The suppression of this pathway allows tumor cells to grow and spread in the micro-environment, eluding the immune surveillance and facilitating cancer progression[19]. Monoclonal antibodies targeting PD-1 or PD-L1 disrupt

this binding, reviving the cytotoxic functions of T-cells against tumor antigens, and slowing tumor growth and dissemination[19].

In the context of metastatic NSCLC, different ICIs have achieved significant survival benefits compared with standard chemotherapeutic agents, starting from second-line setting[20–23]. Due to these extraordinary results, ICIs have been moved to the first-line scenario and have been studied alone or in combination with chemotherapy, as monotherapy or as dual ICI combined with anti-CTLA4 blockade. Since the early development of these drugs, trial designs focused on potential biomarkers, able to maximize the expected benefit of ICIs in NSCLC. PD-L1 expression on tumor cells, for instance, has been described as a useful tool in this setting, with an expression on 50% or more of tumor cells identified as a good threshold for the efficacy of ICI monotherapy[24]. This tumor proportion score (TPS) is routinely available in clinical practice, and is evaluated by pathologists through a simple immunohistochemical (IHC) analysis of tissue biopsy samples, making it a generalizable and easy-to-access biomarker[25].

For treatment-naïve NSCLC, first-line strategy evaluating ICIs monotherapy have targeted mainly PD-L1 high tumors (TPS ≥50%), with limited results in the overall PD-L1 positive (TPS≥1%) population[26,27]. In the KEYNOTE-024 trial, patients with metastatic NSCLC and a TPS ≥50% were randomly assigned to pembrolizumab 200 mg every 3 weeks (anti-PD-1 antibody) or investigator's choice of platinum-based chemotherapy[27]. The primary endpoint was progression-free survival (PFS), estimated as the time between treatment initiation and progression of disease or death from any cause. In this trial, median PFS was 10.3 months (95% CI, 6.7 to not reached) in the pembrolizumab group versus 6.0 months (95% CI, 4.2 to 6.2) in the chemotherapy group (hazard ratio for disease progression or death, 0.50; 95% CI, 0.37 to 0.68; $p<0.001$)[27]. With a longer follow-up, a significant improvement in overall survival (OS) was achieved: median OS was 26.3 months (95% CI, 18.3 to 40.4) for pembrolizumab and 13.4 months (9.4-18.3) for chemotherapy (hazard ratio, 0.62; 95% CI, 0.48 to 0.81), resulting in a 5-year OS rate of 31.9% for the pembrolizumab group[28]. These impressive results in the first-line setting for this patient population were then replicated by other ICIs targeting the same pathway, such as atezolizumab (anti-PD-L1 antibody) in the IMpower110 trial[29,30] and cemiplimab (anti PD-1 antibody) in the EMPOWER-Lung1

trial[31,32]. This evidence led international guidelines to recommend the use of ICI monotherapy in patients with metastatic NSCLC and a TPS ≥50% for up to two years of treatment (for pembrolizumab and cemiplimab) or until disease progression (for atezolizumab)[33]. As far as safety is concerned, treatment-related adverse events of any grade were not increased with ICI compared to chemotherapy, but immune-related phenomena emerged as events of special interest. Unblocking the interaction between PD-1 and PD-L1 enhances the possibility of autoimmunity against host antigens[19], which can increase the chance of T-cell response against healthy organs such as lung parenchyma, skin, colic mucosa, thyroid gland etc. In clinical trials, the incidence of severe immune-related adverse events is under 10% (mostly skin reactions, pneumonitis and colitis)[27], but in real-world practice the cumulative incidence of new autoimmune diseases among patients receiving immunotherapy was over 13% at only 6 months after treatment initiation[34].

Other clinical trials evaluated the combination of chemotherapy ± ICI in first-line setting, independently from TPS. The KEYNOTE-189 and -407 trials randomized patients between chemotherapy and chemotherapy plus pembrolizumab for non-squamous and squamous histology, respectively[35,36]. These two trials share multiple similarities, with the main difference being the chemotherapy backbone that is histology-driven (platinum plus pemetrexed for non-squamous histology, carboplatin and either paclitaxel or nanoparticle albumin-bound [nab]–paclitaxel for squamous histology). In KEYNOTE-189, the estimated rate of OS at 12 months was 69.2% (95% CI, 64.1 to 73.8) in the pembrolizumab-combination group versus 49.4% (95% CI, 42.1 to 56.2) in the placebo combination group (hazard ratio for death, 0.49; 95% CI, 0.38 to 0.64; p<0.001), while median PFS was 8.8 months (95% CI, 7.6 to 9.2) versus 4.9 months (95% CI, 4.7 to 5.5) for pembrolizumab arm and control arm, respectively (hazard ratio for disease progression or death, 0.52; 95% CI, 0.43 to 0.64; P<0.001)[35]. With a longer follow-up, the reduction in risk of death was 40% for chemotherapy plus pembrolizumab compared to chemotherapy alone (hazard ratio [95% CI] 0.60 [0.50 to 0.72]), resulting in a 5-year OS rate of 19.4% versus 11.3% of the chemotherapy arm[37]. In the KEYNOTE-407, the median OS was 15.9 months (95% CI, 13.2 to not reached) in the pembrolizumab-combination group and 11.3 months (95% CI, 9.5 to 14.8) in the placebo-combination group (hazard ratio for death, 0.64; 95% CI, 0.49 to 0.85; p<0.001), and the median PFS

was 6.4 months (95% CI, 6.2 to 8.3) in the pembrolizumab-combination group and 4.8 months (95% CI, 4.3 to 5.7) in the placebo-combination group (hazard ratio for disease progression or death, 0.56; 95% CI, 0.45 to 0.70; P<0.001)[36]. At 5-year follow-up, OS and PFS were consistently improved with pembrolizumab plus chemotherapy versus placebo plus chemotherapy (hazard ratio [95% CI], 0.71 [0.59 to 0.85] and 0.62 [0.52 to 0.74]), reaching 5-year OS rates of 18.4% versus 9.7%, respectively[38]. Interestingly, the benefit in OS and PFS was consistent across all subgroups, irrespectively of TPS[35,36]. Therefore, the guidelines of the European Society of Medical Oncology (ESMO) suggest to consider chemotherapy plus ICI combinations also for metastatic NSCLC with TPS ≥50% if a more rapid tumor shrinkage is required[33], but not all European regulatory agencies have adopted this indication for the reimbursement of ICI-based combinations. In Italy, for instance, chemotherapy plus immunotherapy is available only for patients with metastatic NSCLC and a TPS between 0 and 49%, limiting clinicians in tailoring treatment choices for patients with higher TPS. As for ICI monotherapy, other regimens including ICI and chemotherapy have been developed in these recent years, such as the Checkmate-9LA regimen (nivolumab plus ipilimumab plus 2 cycles of platinum-based chemotherapy[39]) and the EMPOWER-Lung3 regimen (cemiplimab plus platinum-based chemotherapy[40]), which are now approved for reimbursement in Europe[33].

1.1.3 The two tails of immunotherapy: patterns of response and resistance

A key advance of ICIs is their ability to induce long-lasting responses in a consistent proportion of patients with advanced cancers and, in particular, metastatic NSCLC. Beyond impressive 5-year OS rates (up to 30% in patients with metastatic NSCLC and TPS≥50%[28]), there is an interesting plateau of around 10% of patients being alive and progression-free at that same cut-off[28,37,38]. These durable responses, which may persist even after treatment interruption, were barely impossible with chemotherapy alone, and have posed significant challenges in this patient population. Managing long-term response is crucial in clinical practice, because long-term survivors, and more prominently long-term responders, have entirely new needs compared to the overall population of patients with metastatic NSCLC, and need tailored approaches[41]. There is no universal consensus on "durable response" definition, but it can generally be described as a patient having a PFS exceeding three times the median PFS of the whole population of patients treated with the same drug(s) in the same trial, and it is usually more common

in patients treated in the first-line metastatic setting[42]. In a European cohort of patients treated with first-line pembrolizumab for metastatic NSCLC with a TPS≥50%, hierarchical organization indicated good PS, younger age and higher PD-L1 expression as the most important predictors of long-term survival[43], but the early identification of this subgroup is still a matter of research.

In clinical trials, the definition of response upon ICI-based regimens is evaluated with strict scientific criteria. The RECIST (Response Evaluation Criteria in Solid Tumors) criteria are standardized guidelines used to measure tumor burden and evaluate response to treatment in clinical trials; they rely mainly on changes in the longest diameter of target lesions assessed by CT-imaging[44]. These criteria dissect the possible evolution of cancer lesions as summarized in *Table 2*.

Response category	Definition (RECIST 1.1)
Complete Response (CR)	Disappearance of all target lesions
Partial Response (PR)	≥30% decrease in the sum of diameters of target lesions (reference: baseline)
Stable Disease (SD)	Neither sufficient shrinkage to qualify for PR nor sufficient increase to qualify for PD
Progressive Disease (PD)	≥20% increase in the sum of diameters of target lesions (reference: smallest sum recorded) or appearance of new lesions

Table 2. RECIST criteria version 1.1

However, tumors treated with ICIs tend to respond differently compared with other standard agents, and specific RECIST criteria have been developed by the scientific community to address these disparities[45]. The iRECIST criteria introduced new concepts in the evaluation of response, such as pseudo-progression and the consideration of clinical status[45]. Despite their publication more than 5 years ago, they are still not routinely used neither in research not in clinical practice, and they probably need further validation across cancer types[46]. Nevertheless, pseudo-progression is now a well-known pattern of

response, usually defined as an objective response after having an initial disease progression[47]. It has been explained as a transient immune-cell infiltrate in the tumor causing a temporary increase in tumor size[48], without concurrent clinical deterioration. Across different cancer types, pseudo-progression is indeed quite rare (less than 10%)[47], and should therefore be taken into account as a possible event only in selected clinical cases.

Despite the potential for durable responses, the downside of ICI is the occurrence of resistance, which can be further classified into primary and acquired resistances[49]. Patients with primary resistance fail to respond to ICI (experiencing progressive disease at first restaging or short-lasting stable disease): current evidence suggests that both tumor-intrinsic and tumor-extrinsic factors contribute to this mechanism, i.e. the absence of antigenic proteins or their presentation, T-cell exclusion, upregulation of inhibitory immune-checkpoints or immunosuppressive cell infiltration such as tumor-associated macrophages[49]. From a clinical perspective, these patients may face not only disease progression, but also clinical deterioration and rapid tumor growth, as in case of hyperprogression (HP). The concept of HP was first reported in retrospective studies reporting a faster tumor growth after ICI initiation[50]. Different studies used different methods to assess this uncontrolled growth, but HP was overall associated with worse OS, even if never fully biologically explained[47]. In clinical trials with ICI monotherapy, the rate of early progression and deaths is around 15%, but in real-world data this percentage rises up to 45%[51].

On the other hand, acquired resistances develop when patients progress during ICI after an initial tumor response[52]. The most common mechanisms for relapse include loss of T-cell function and development of escape mutation variants in tumor cells[49], but there are different definitions of acquired resistance in literature. In the context of NSCLC, Schoenfeld et al stated that it should meet the following criteria: patients treated with ICI, experiencing objective response upon ICI, and facing progressive disease within 6 months of last ICI dose[53]. While it is clear that adding chemotherapy to ICI reduces the risk of primary resistances and halves the early mortality rate[51], its role in preventing acquired resistances is controversial[54].

In conclusion, patients with metastatic NSCLC treated with ICI may experience different shades of response, clinical benefit and survival gain, therefore the early identification of solid predictive and prognostic biomarkers is an unmet need in this population.

1.1.4 Prognostic and predictive tools for metastatic NSCLC

Many studies have explored the role of different biomarkers in predicting response to ICIs. Prognostic factors are usually defined as predictors of the natural history of the disease (mirroring disease aggressiveness), while predictive factors should help understanding which patients derive the most benefit from a specific treatment[55]. Nevertheless, variables emerging from clinical trials are not fully reproducible in clinical practice, due to strict eligibility criteria, and are therefore of scarce utility in real-world setting[56], highlighting the need of more reliable tools.

As discussed before, PD-L1 TPS is the only validated biomarker available nowadays for treatment allocation in this setting[33], but its predictive power is suboptimal: it may vary consistently between primary and metastatic tumor samples, and among different anatomic sites of the same patient[57]. There is retrospective evidence that patients with a very high TPS (90-100%) tend to have higher response rate and longer median PFS and OS compared to patients with a TPS 50-89% if treated with ICI monotherapy[58], but these findings still have minimal impact on patients' selection in absence of alternative therapeutic strategies.

Therefore, the scientific community is working towards the implementation of other factors (or sum of factors) that can help clinicians in risk estimation and subjects' selection. A list of the most relevant known variables follows:

- ECOG PS: patients with a PS of 0-1 have a better survival (PFS and OS) compared to patients with a PS≥2, above all if the deterioration of PS is due to cancer burden and not to co-occurring comorbidities[59,60];
- Weight loss: patients with weight loss of more than 5% have poor OS[61];
- Sex: women have a reduced efficacy of ICI monotherapy compared to men[62];
- Body-mass index (BMI): patients that are considered overweight in the BMI classification have longer survival (PFS and OS) if treated with ICI[63], while cancer-induced cachexia is a well-known adverse prognostic factor[64];

- Sites of metastases: pleural effusion is a negative predictor for PFS upon ICI[65], while liver metastases are associated with worse outcomes irrespective of TPS[66], as well as brain[67] and bone metastases[68];
- Concomitant medications: both systemic corticosteroids[69], proton-pump inhibitors[70] and antibiotics[71] are associated with reduced therapeutic effects of ICI therapy;
- Blood-based biomarkers: baseline high neutrophil-to-lymphocyte ratio (NLR)[72], high lactate dehydrogenase (LDH) levels[73] and elevated C-reactive protein (CRP)[74] are associated with poorer outcomes during ICI therapy;
- Co-occurring mutations: KRAS mutations alone, which are the most common driver mutations in Western countries, do not have a clear impact on ICI efficacy[75], but they may play a negative prognostic role if co-occurring with STK11 or KEAP1 mutations[76] or a positive role if co-occurring with TP53[77] as it happens with SMARCA4 mutations[78].

Some of these variables have been summarized in ready-to-use scores, such as the Lung Immune Prognostic Index (LIPI) (baseline NLR ≥3 and LDH ≥normal values correlate with worse survival[79]) or the Lung Immuno-oncology Prognostic Score (LIPS-3) (including baseline NLR, PS ≥2 and pretreatment steroids as poor prognostic factors[80]), but none of them have entered routine clinical practice. While other next-generation biomarkers emerge from clinical trials (Tumor Mutational Burden [TMB] or circulating tumor DNA dynamics)[57], prospective evidence of robustness of well-known factors is still lacking, and clinicians tend to rely on their clinical expertise only for prognosis estimation.

1.2 Machine learning for predictive modelling

1.2.1 Common machine learning algorithms for clinical outcome prediction

Machine learning (ML) is a branch of artificial intelligence (AI), whose aim is to identify patterns among data and optimize predictions across different domains[81]. In the oncological field, ML can encompass a broad range of tasks and methods, both as clinical decision support tool and as an exploratory research basis[82].

The algorithms that are most often used for predictive modelling are *supervised* models, which means that ML models learn from data that have already been labelled for the outcome of interest (i.e. the ML model knows which patients have to be considered responders or non-responders to a specific treatment)[82]. There are many supervised models that differ from one another for their internal mechanism and their overall complexity. Linear models, for example, tend to map the independent variables to the outcome via a linear equation, which makes them easy to interpret also for end-users[82]. Decision tree (DT) models are more complex models that rely on differently ranked features (splits) that sub-divide groups until the last observation (leaf), which actually is the final affiliation subgroup in terms of prediction[82]. Every split is chosen to minimize a loss function, and DTs tend to be more useful in case of categorical variables, since they do not capture very well the continuous relationships between features and outcomes[82]. In the same framework as DTs, ensemble methods are considered more powerful, since they build many DTs to generate predictions[82]. Some examples of this category are random forests or gradient boosted machines (XGBoost), which train each tree independently (using a random subset of data or iteratively) and have an error-correcting approach to improve their performances[82].

However, the multiple layers that constitute the internal trees are difficult to understand, and this lack of transparency hinders their application in the clinical setting, where interpretability is crucial for end-users[82]. In this case, algorithms that explicit feature importance are often paired with ensemble models in a post-processing phase to improve their usability[83]. Explainable AI (XAI) has gained attention in these last years because of its potential ability to interpret the reasoning within the models, and improve users' trust[84]. The most used XAI algorithm in the biomedical setting is SHAP (SHapley Additive exPlanations), which can be seamlessly integrated into supervised ML models[85].

It is a feature-based interpretability method that provides outcomes that are easy to read and to interpret in terms of ranking (most relevant features on the top) and behavior (different colors mean different impacts, i.e. low values are blue and they shift the model towards the event, high values are red and shift the model to the other direction)[85]. Therefore, in the recent years, many supervised ML models have been developed for outcome prediction, and they have often been matched with a XAI algorithm to improve their transparency for end-users.

On the other hand, *unsupervised* models work well as exploratory tools. In these approaches, data are not labelled (the outcomes are not known by the model), and ML models are free to detect patterns within the dataset that are totally unbiased from human supervision[82]. For example, clustering algorithms (such as *K*-means and hierarchical clustering) are able to part the data into *K* clusters to maximize similarity within clusters and separation among clusters[82]. Despite their potential role in exploratory data analysis (after clustering, statistical tests could be performed to study different means or variances of features among clusters)[82], unsupervised models do not inherently consider interpretation, limiting their applicability in clinical settings as discussed above.

In the context of ML, deep learning (DL) has emerged as a subfield built on neural networks (NN), which perform very well in the presence of big and raw data (such as free text or raw images)[82]. The power of these tools lays in their hidden layers, where complex interactions between features and outcomes are captured, but this results in a complete lack of interpretability that limits their utility[82]. Therefore, DL models need a higher computational power and have very specific applications in the biomedical setting, such as digital pathology or imaging processing.

In conclusion, different ML models can play a role in outcome prediction in oncology, both for classification purposes (benign/malign, responder/non-responder) and for risk prediction, but well-performing algorithms cannot transcend a good organization of input data, which are the foundation of their predictive ability.

1.2.2 Real-world data in healthcare: the challenge of standardization

In healthcare, data are heterogeneous and vary widely, both in their structure and in their informative content. Moreover, they are often stored in different locations within the same hospital, with software hosting clinical data (as free text or collected in a case report

form [CRF]) and others storing patient's images or laboratory tests[86]. In this context, data fragmentation is one of the biggest issues when collecting data for research, above all for multicentric studies[86]. In the Italian healthcare system, for example, there is no centralized registry for cancers, there are different electronic health records (EHR) used among different hospitals, and ethical approvals depend on local committees that do not consider or interact with each other[87]. These major challenges for data collection impact the research capacities of hospitals, reducing the authorized studies, limiting patients' opportunity and making high-quality studies nearly impossible for the huge amount of manual work needed[87]. Moreover, most of the available data are unstructured data (lacking a pre-defined data model, such as free text mixed with numbers or dates), which causes problems for data sharing and interchange, and have a high proportion of missing values[86]. Many projects are attempting to reduce heterogeneity in this setting by mapping institutional datasets to common data models (i.e. Observational Medical Outcomes Partnership [OMOP]), but their adoption is slow and hindered by multiple challenges at a local level[88].

Despite these limitations, real-world data (RWD) are considered important sources of information for patients' outcomes outside of conventional clinical trials[89]. RWD studies can cover bigger sample sizes, without strict eligibility criteria applied for sponsored trials, are usually faster and cheaper to perform, and sometimes are the only way to study rare diseases[89]. Moreover, they are considered "big data" for their quantity, and are therefore considered good candidates for training ML models for prediction, despite their intrinsic "imperfectness"[90].

Multiple approaches can mitigate the weaknesses of RWD in the phase of data preparation and data extraction. For clinical variables, they are often collected from EHR in free text format, and each report may use a different terminology to denote the same entity[91]. Natural Language Processing (NLP), a branch of AI for text processing, has significantly transformed healthcare data collection in these recent years[92]. Text Analytics for health, for instance, has been developed by Microsoft© to identify and label medical information in unstructured text via ML algorithms[93]. This kind of tools can define the entity of words (i.e. diagnosis, medication name, symptom), extract relations within words (event with time), associate entities with their ontological code (supported by the Unified Medical Language System Metathesaurus) and contextualize with certainty or

conditionality according to the presence of modifiers[93]. Despite performing very-well in most of the contexts, EHRs are often biased by the presence of typos, misspellings and abbreviations that restrict NLP applicability, and these tools are often underperforming if used in language other than English, because of the scarcity of training data[94]. As a result, clinical data collection relies mostly on manual effort, at least for all those centers (or departments) that do not routinely structure CRFs. In some hospital software, for example, data can be produced in a semi-structured format, where free text is usually limited within pre-specified tabs. In this case, NLP tools work better, but still are not fully independent from manual check[95]. On the other hand, other data sources are easier to navigate because they automatically produce tabular data, as laboratory information systems, and are often accompanied by metadata that identify the reference ranges and have measurement units that are more easily interchangeable worldwide[96].

1.2.3 Privacy-compliant platforms: the S-RACE pipeline

In the current context, RWD sharing is allowed under strict conditions, which have been the subject of multiple in-depth analyses by the main regulatory bodies responsible for data management and privacy. Within the data science community, data stewardship is defined as the long-term sustainable care for research data that lays its basis on the FAIR principles[97]. These principles state that data should be Findable (persistently identifiable), Accessible (clear conditions to use them), Interoperable (integrating resources with minimal effort) and Reusable (well-described by metadata for further re-use)[97]. Every researcher is responsible for his own data, and every research institution should monitor the compliance to FAIR principles and to General Data Protection Regulation (GDPR), as defined by international guidelines such as the European Regulation 2016/679 on personal data protection[97]. If we put AI into this context, its responsible integration into healthcare demands additional strict criteria to ensure safety and ethics, as supported by multiple European initiatives[98]. In 2023, ISO/IEC 24001, the world's first international standard for AI management system, has been published, and it discussed AI developments in the current regulatory framework[99]. In 2024, a comprehensive document has summarized the European standards for risk mitigation, quality data and human responsibility in medical AI, aligning with ISO 42001[99]. Data safety and ethical use are key concepts for AI application in healthcare, and research

institutions working with clinical data have started building their own infrastructures to accomplish the mission of integrating AI in real-world hospitals.

At Università Vita-Salute San Raffaele, a joint collaboration with Microsoft© and Porini srl (part of the DGS group) supported the development of the S-RACE (San Raffaele Ai CEnter) platform. This platform was ideated as a secure and trustworthy cloud-based solution for RWD, designed to integrate a pipeline ranging from data ingestion to ML models deployment[100]. The platform architecture is summarized in *Figure 1*.

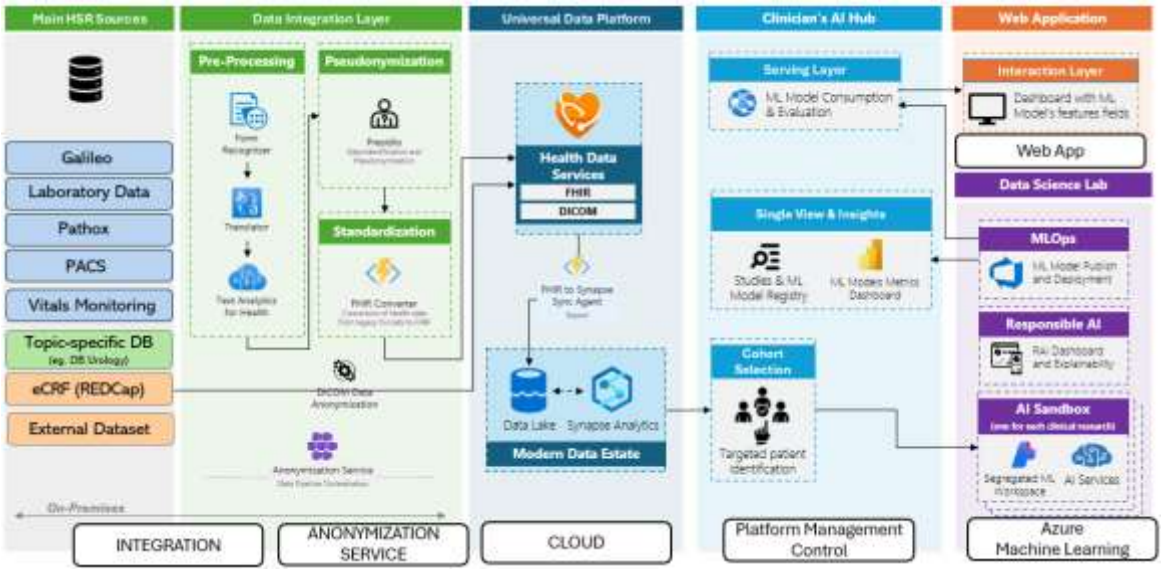

Figure 1. S-RACE platform architecture

Access to the platform is granted by an administrator, and different professional profiles (clinicians, study coordinators, or data scientists) have access to different rooms within the same environment. Once the study is created (and approved by the local ethical committee), the principal investigator (PI) can define his study cohort on-cloud, entering inclusion and exclusion criteria in the search engine or manually adding patients using their hospital-IDs. The PI has also to specify which local data sources to import, assuming that every hospital software (and its output) is tracked in the platform, minimizing the need for manual collection. An anonymization engine processes data on-premises, converting them into pseudonymized datasets before uploading. This setup addresses key privacy risks while enabling a hybrid cloud model that combines cloud scalability with private infrastructure security. After cohort definition, clinicians and data scientists can

visualize data in their own rooms (Clinicians' AI Hub or Data science Lab), which offer different frameworks for data visualization, preliminary analysis, and model building.

The S-RACE platform and its specifics for healthcare data management have been thoroughly described in a manuscript by Traverso et al, which has been accepted for publication by *npj Digital Medicine*, and its detailed reporting goes beyond the scopes of this study. It was briefly introduced with the specific purpose of describing the regulatory and technological context in which AI-HOPE study was conducted.

1.3 Integration of multimodal data towards precision oncology

1.3.1 Approach to imaging: radiomics and its relevance in NSCLC

A very promising data source that has not been discussed so far is imaging, in particular computed-tomography (CT) scans that are routinely performed in clinical practice for patients with a cancer diagnosis. In lung cancer diagnosis and assessment, CT-based information is integrated with histological and molecular findings to inform treatment planning, and is the basis for clinical TNM staging[15]. In clinical practice, aside from size measurements, most imaging features are described *qualitatively*. Radiomics, on the other hand, extracts a set of high-dimensional data exploiting *quantitative* analysis on medical images, usually from a pre-specified region of interest (ROI) that is manually contoured on an imaging scan[101]. Radiomics assumes that biomedical images contain disease-related details beyond human perception, and, by applying mathematical algorithms, it extracts measurable textural features[101]. NSCLC provides an ideal context for radiomics due to the high contrast between the tumor and surrounding lung tissue on CT-scans, and the availability of data driven by high incidence rates and periodical imaging assessments[102]. Radiomic feature extraction requires strict protocols for imaging acquisition, in order to minimize inter-operator bias and optimize reproducibility. A standard process for radiomic features extraction is summarized in *Figure 2*.

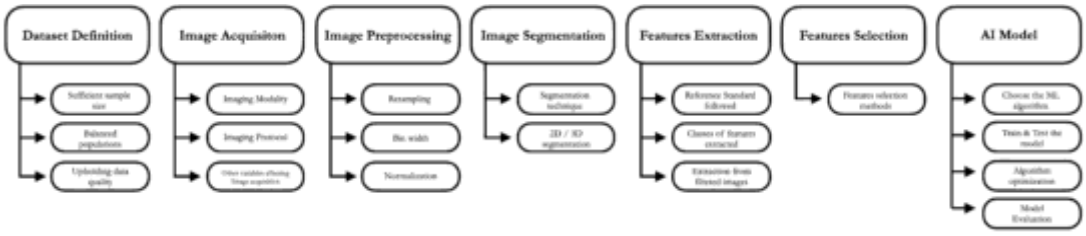

Figure 2. Diagram of the main steps in a radiomics study (extracted from "Insights into radiomics: a comprehensive review for beginners" by Mariotti et al, published in Clinical and Translational Oncology in 2025[101])

Overall, every study considering radiomics should follow these steps[101]:

1- Standardized acquisition protocols to minimize variability between scanners and settings;
2- Preprocessing of images for intensity discretization and normalization;
3- Definition of the ROI via manual, semi-automatic or automatic approaches;
4- Feature extraction with standard packages (i.e. PyRadiomics);

5- Feature selection (manual or ML-based);

6- Inclusion in models (statistical or ML).

Radiomics involves various classes of features, starting from first-order features (like energy and entropy) to more complex higher-order features. The Image Biomarker Standardization Initiative (IBSI) was formed to provide a standardized list of radiomic features that have demonstrated reproducibility, see *Figure 3* for details.

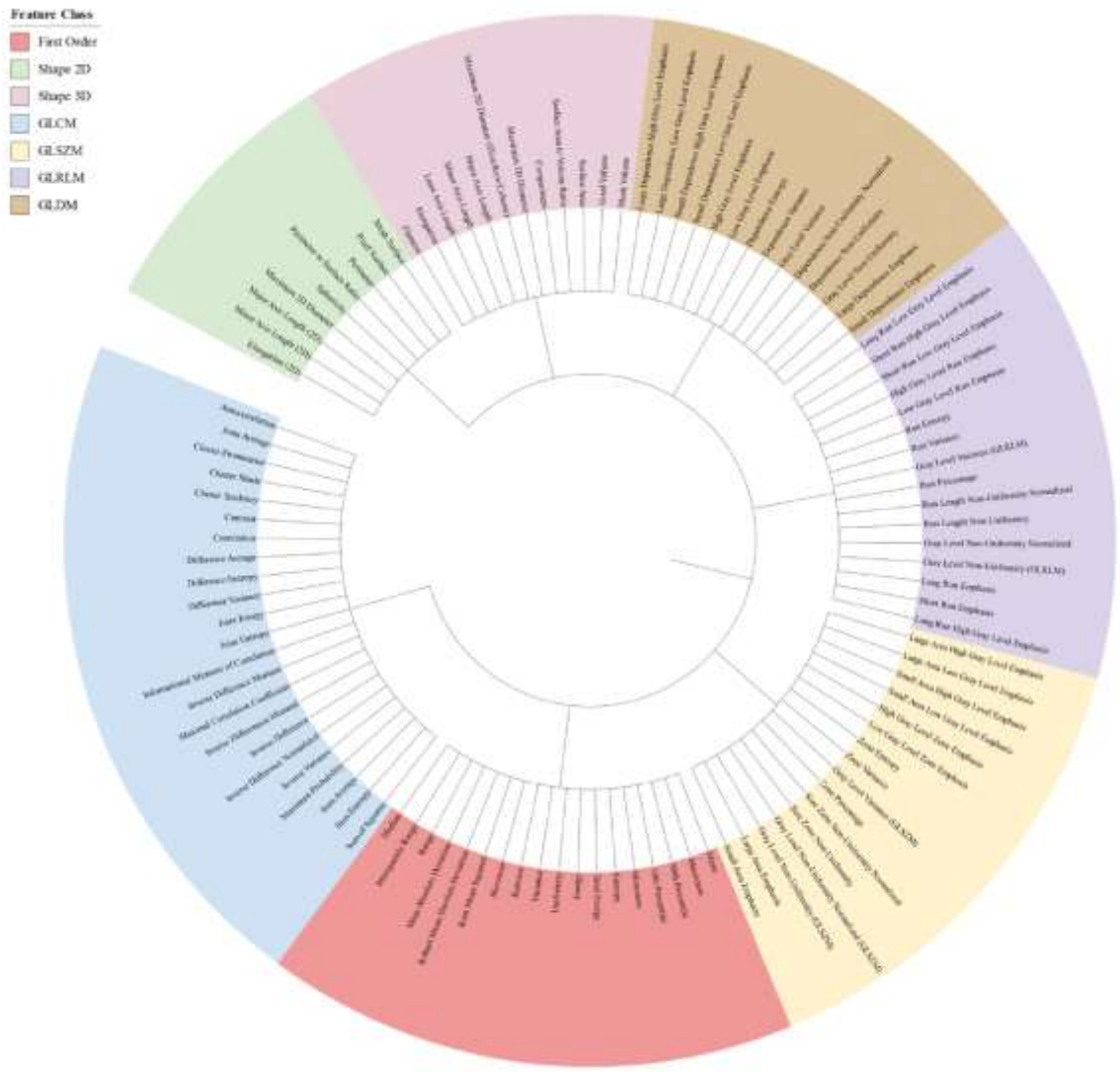

Figure 3. Phylogenetic tree illustrating common radiomic features grouped by their class (extracted from "Radiomics and artificial intelligence for precision medicine in lung cancer treatment" by Chen et al, published in Seminars in Cancer Biology in 2023[103])

In summary, *shape features* capture the geometric properties of tumors, such as diameter, volume, and ratios like surface-to-volume, *intensity-based features (first-order)* describe the distribution of pixel intensities within the region of interest (ROI), including

metrics like mean, variance, and skewness, and *texture features* quantify the relationships between voxels within an image, derived from the gray-level co-occurrence matrix (GLCM), which measures pixel pair frequencies[104].

Since its discovery in preliminary studies showing correlations between radiomic features and outcomes in patients with lung cancer[105], it has been widely studied with the aim of overcoming reproducibility issues and assess its clinical relevance. For early-stage NSCLC, radiomics has been used to predict pathologic response upon neoadjuvant treatments[106,107] or to correlate with T-cell infiltration on surgical samples[108]. Analyzed images are both baseline scans or post-treatment scans, and are not limited to CTs, but often involve 18F-FDG-PET[109], and the peritumoral area[110]. Radiomic features have also shown relevance in predicting distant metastases in locally advanced disease, such as stage III NSCLC[111], but larger studies reached conflicting results[112], limiting the applicability of radiomic-based models in real-world practice.

In the metastatic scenario, solid evidence on radiomics is indeed scarcer. A pioneering study of Trebeschi et al in 2019 showed a correlation between radiomic biomarkers and response to immunotherapy in 2023 patients with metastatic melanoma and NSCLC[113]. Interestingly, in the NSCLC cohort of this study, all metastatic lesions (larger than 5mm) were manually contoured for feature extraction and radiomic signatures were associated with pathways involved in mitosis on a gene set enrichment analysis, supporting a biological explanation of results[113]. This methodological approach (manual contouring, tissue samples matching) set a very high standard for radiomic studies, which is not easy to mirror on a retrospective real-world basis. Specifically, other studies tried to exploit radiomic features to predict immune response in NSCLC[114,115], but emerging results from different ML-approaches have proven inconsistent among different experiences, and not always superior to routine clinical biomarkers.

1.3.2 Multimodal ML models in metastatic NSCLC

Most of the ML models that have been published so far for outcome prediction in metastatic NSCLC consider standard tabular data such as clinical, laboratory and histopathological information, which are considered standardized, solid and easy-to-access data. In literature, however, there are experiences trying to merge also radiomic data into predictive ML models. For instance, Yolchuyeva et al built time-to-event models

for OS and PFS for patients with metastatic NSCLC treated with first-line immunotherapy, achieving mediocre performances (C-index of 0.60 for PFS and 0.65 for OS in validation cohorts) and including only radiomic features[116]. Beyond the proof-of-concept value, many methodological issues can arise from such an approach, regarding sample size (less than 150 patients), pre-processing procedures, and the inconsistency of a whole non-clinical model. While ML by itself works well with any other kind of "omics" data (transcriptomics, genomics, proteomics[117]), the main limitation for clinical applicability of all "omics" data is that this set of information is usually derived from clinical trials only, and not routinely available as RWD. Moreover, ML models require large datasets for training, not only considering the number of variables per patient but, more importantly, in terms of the overall number of cases[118].

Therefore, the most reliable models for outcome prediction still rely on common clinical, pathological and genomic features. Published models range from very simple logistic-regression models, such as LORIS[119], to more complex survival modelling[120]. The strength of these approaches, usually developed in the US, are many: availability of nation-wide EHRs and cancer registries, very big sample size, simple computational models, and transversal design. When paired with available biological variables (such as PD-L1 TPS or TMB), results were consistent with expected results in a pan-cancer perspective[119]. Moreover, with such large datasets, multiple model testing is possible, encompassing both survival modelling (time-to-event algorithms like Cox regressions) and classification models (response yes/no, stable disease yes/no at pre-specified timepoints) in the same framework[121]. Another more complex DL-model, published by Saad et al, was trained on a retrospective cohort of 2300 patients with metastatic NSCLC (clinical and genomic data) to assess a new score (A-STEP, Attention-based Scoring for Treatment Effect Prediction) for chemotherapy use in combination with immunotherapy in first-line setting[122]. The resulting model includes peculiarities of different models, from scoring functions to attention mechanisms, and is implemented with SHAP in a post-processing phase to improve its explainability[122].

Although AI technology is advancing at an unprecedented pace, its translation into clinical value and bedside adoption remains limited, and it is estimated that less than 2% of AI models actually overcome the prototyping phase[123]. The gap between research environments and clinical applications is gigantic: academically developed ML models

neglect factors that influence their adoption, such as implementation strategies and user interfaces for clinical settings, that are not adequately captured through performance metrics alone[124]. In order to fulfil this clinical need, overcoming the aforementioned limitations, in a totally different setting (COVID-19 pandemic), Palmisano et al (from Università Vita-Salute San Raffaele) developed an AI-platform with both computer and mobile phone interfaces for prognosis estimation (AI-SCoRE[125]). This model was retrospectively trained and prospectively validated, it was able to identify risk classes and included only five variables: two demographic data, one clinical data and two imaging-derived features extracted by a chest CT-scan[125]. A summary of AI-SCoRE computation and results is shown in *Figure 4*.

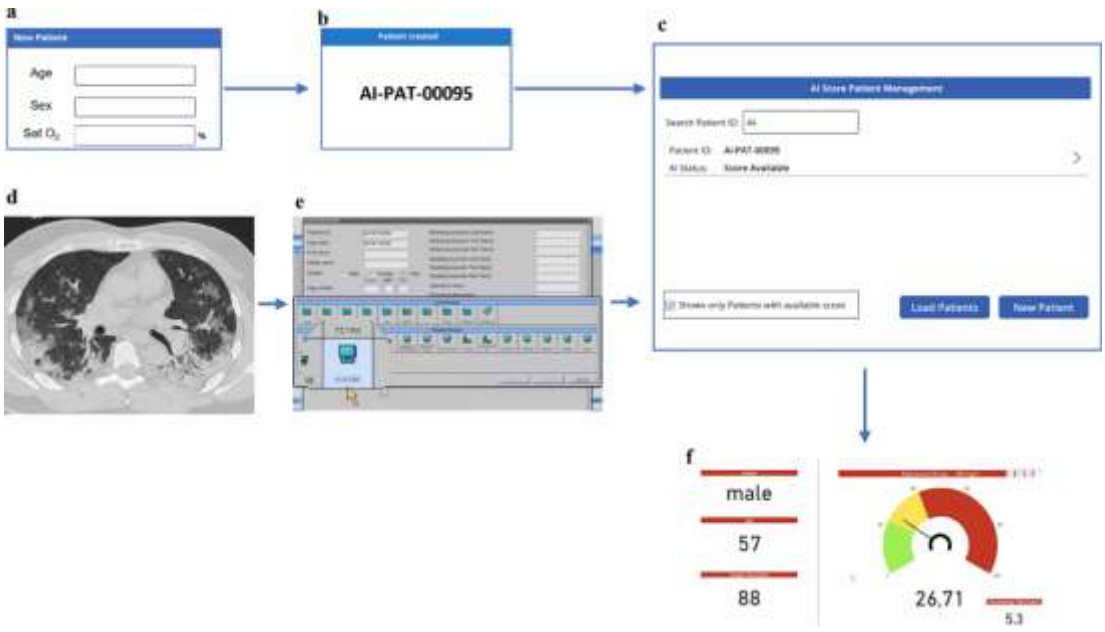

Figure 4. AI-SCoRE processing steps, a) clinical variables, b) anonymization code, c) pop-up message with risk score, d+e) CT-scan uploading, f) patient's risk class classifier (extracted from "AI-SCoRE (artificial intelligence-SARS CoV2 risk evaluation): a fast, objective and fully automated platform to predict the outcome in COVID-19 patients" by Palmisano et al, published in La radiologia medica in 2022[125])

Its transparency, requiring no additional XAI algorithms, its user-friendly interface and its processing speed are a clear example of a clinical decision support tool that can transform patients' care on a real-world basis, and that laid the basis for the development of other AI-based projects with prompt applicability.

1.4 Study objective

The AI-HOPE study was designed to collect RWD of patients with metastatic NSCLC treated with first-line immunotherapy (with or without chemotherapy) and build ML models that can predict outcomes (see section 2.3 “Outcome definition” for details).

In this thesis, only the following results will be presented: the pilot experiments of data acquisition and analysis from different data sources (mainly *in-hospital*), the development of multimodal binary models and of simple time-to-event models, which will serve as the backbone for more complex multimodal time-to-event models (primary endpoint).

The study is still ongoing, and planning to achieve the final estimated sample size of 1000 patients by the end of 2026 for the retrospective cohort. Prospective validation will follow.

Chapter 2: Materials and methods

2.1 Study design and cohort description

2.1.1 Study type

The AI-HOPE study is an ambispective observational study, including both a retrospective and a prospective cohort for data collection. The retrospective cohort collects data of patients treated with the drugs of interest (immunotherapy with or without chemotherapy) from the first approval by the regulatory agency (i.e. Agenzia Italiana del Farmaco, for Italian centers, College for Evaluation of Drugs [College ter Beoordeling van Geneesmiddelen] and Healthcare Institute Netherlands [Zorginstituut Nederland] for Dutch centers) to the start of the prospective phase. The prospective cohort will collect data of patients from 01 Jan 2026 to 31 Dec 2027. Foreign centers (outside of Italy) are involved in the retrospective phase only.

The list of participating centers follows:

1. IRCCS Ospedale San Raffaele (Milan, Italy)
2. Fondazione IRCCS San Gerardo dei Tintori (Monza, Italy)
3. ASST Grande Ospedale Metropolitano Niguarda (Milan, Italy)
4. A.O.U. San Luigi Gonzaga (Orbassano, Italy)
5. Azienda Ospedaliera Spedali Civili di Brescia (Italy)
6. Policlinico di Milano Ospedale Maggiore (Italy)
7. Ospedale di Circolo e Fondazione Macchi (Varese, Italy)
8. Ospedale Sant'Anna (San Fermo della Battaglia, Italy)
9. IRCCS MultiMedica (Sesto San Giovanni, Italy)
10. Ospedale Civile di Legnano (Italy)
11. Azienda Ospedaliera Universitaria Integrata di Verona (Italy)
12. IRCCS Ospedale Policlinico San Martino (Genova, Italy)
13. Humanitas Research Hospital (Rozzano, Italy)
14. Maastricht University Medical Center+ (The Netherlands)
15. Maxima MC (Veldhoven, The Netherlands)
16. Catharina Ziekenhuis (Eindhoven, The Netherlands)
17. VieCuri Medisch Centrum (Venlo, The Netherlands)
18. Zuyderland MC (Heerlen, The Netherlands)

19. University Medical Center (UMC) Groningen (The Netherlands)
20. Kepler University Hospital Linz (Austria)
21. Hospital da Luz (Lisbon, Portugal)

2.1.2 Inclusion and exclusion criteria

The main eligibility criteria are as follows:

- Age ≥18 years
- Diagnosis of non-small-cell lung cancer (histological or cytological)
- Diagnosis of advanced or recurrent disease not amenable for radical local treatments, or metastatic disease
- Treatment with first-line immunotherapy in monotherapy (Cohort A) or immunotherapy plus chemotherapy (Cohort B) according to international guidelines[33] and per local use
- Availability of follow-up data (at least survival status), for retrospective cohort only

2.1.3 Type of data

For all patients the following data are collected:

- Clinical (age, sex, smoking status, comorbidities, medications, performance status, symptoms)
- Laboratory (white blood cells, neutrophil count, lymphocyte count, hemoglobin, platelets, hepatic and renal function, lactate dehydrogenase levels, albumin, C-reactive protein, electrolytes)
- Pathological (cancer histology, PD-L1 TPS, genomic tests if available)
- Cancer-related follow-up (date of diagnosis, TNM staging, sites of metastases, local and systemic treatments, toxicities, progression, vital status)

2.1.4 Ethical approval and informed consent

The AI-HOPE study was approved by Ospedale San Raffaele ethical committee (Comitato Etico Territoriale [CET] Lombardia 1) on 24/01/2024 and registered to

clinicaltrials.gov as requested per local legislation (NCT06788366). All the participating centers that did not consider CET Lombardia 1 approval valid for their institution submitted the protocol to their local ethical committee and obtained approval before participation. Data transfer agreements were set up between the participating centers and Ospedale San Raffaele.

Informed consent is waived for the retrospective cohort as per local use (via Data Protection Impact Assessment written by Data Protection Office, for Ospedale San Raffaele, and in a similar way for the other centers). Written informed consent is mandatory for patients to be included in the prospective phase.

2.2 Data collection

2.2.1 Clinical data collection

Clinical data are collected on an electronic CRF (eCRF, REDCap), specifically designed for the study. Due to lack of approved and validated NLP tools for EHR, data collection is manually performed by study personnel.

For all the centers that already had a collected dataset encompassing the most relevant features (i.e. PREBREM dataset for Dutch centers), in line with the FAIR principles[97], we performed a variable mapping between pre-existing datasets and AI-HOPE dataset, to optimize and speed-up the data collection process.

2.2.2 Laboratory data collection

For the internal cohort, laboratory data are collected directly from Ospedale San Raffaele laboratory software after a thorough mapping of the variables of interest and automatically uploaded on the S-RACE platform (with anonymized patient ID). For collaborating centers, laboratory data are collected on the same eCRF or mapped when re-usable from pre-existing datasets. Rules are applied in order to consider the variations of measurement units within the observational period (i.e, hemoglobin from g/dL to g/L) and among different countries (i.e. hemoglobin from g/L in Italy to mmol/L in the Netherlands). The timeframe for laboratory data collection is between -30 days and +5 day from treatment initiation; if multiple data are available for the same variable, the closest one to date is collected.

2.2.3 Radiological data and radiomics pipeline

2.2.3.1 Imaging modality and acquisition protocols

Baseline CT-scans (and 18-FDG-PET scans, when available) are collected for all patients that have not received radical treatments on the primary tumor (i.e. excluding patients with disease recurrence after chemo-radiation or after surgery for early-stage disease). The timeframe for imaging collection is between -40 days and +10 days from treatment initiation; if multiple scans are available, the closest one to date is collected.

Collected CT-scans have to be contrast-enhanced, with a minimum slice-thickness of 3 mm, a reconstruction kernel optimized for lung parenchyma and including at least the thoracic region; 18-FDG-PET scans have to be performed according to international guidelines[126].

All the images are pseudo-anonymized locally, removing the privacy-sensitive metadata (i.e. DICOM header) and substituting it with the anonymized REDCap ID. All the images are uploaded on a Orthanc node (DICOM server), where they are available for visualization and segmentation. The DICOM server is then synchronized with the Azure ML Studio environment.

2.2.3.2 Tumor segmentation and radiomic feature extraction

Radiomic-target lesions are defined as primary lung tumors having a minimum largest diameter of 5 mm measured on the axial slice. On CT-scans, radiomic-target lesions are automatically segmented using the open-source software *TotalSegmentator*[127] and pushed automatically to Orthanc in DICOMSEG format. The segmentations are reviewed by an expert radiologist with at least 3-year experience in thoracic imaging. Each segmentation is ranked by radiologists with a 3-point satisfaction score to assess *TotalSegmentator* performance in our cohort (1=good contouring, 2=acceptable, but not as good as manually performed, 3=unacceptable, manual correction mandatory). If needed (score 3), radiologists can edit the segmentations directly on the platform using the Orthanc plugin OHIF viewer[128].

On 18-FDG-PET scans, radiomic-target lesions are automatically segmented using a threshold-based approach (40% of the Standardised Uptake Value, SUV, as reported in literature[129]) with the open-source software LIFEx[130] or with syngo.via software [131].

Radiomic features (shape, intensity and texture features) are extracted from each ROI using the open-source software PyRadiomics (version 3.1.0[132]) processing segmented images. After filtering and harmonization, when required, extracted radiomic features are stored in the platform as tabular data.

2.2.3.3 Feature reproducibility and stability assessment

The robustness of extracted features is evaluated using two publicly available datasets: RIDER (test-retest, repeatability analysis), and INTEROBSERVER (multiple contours, reproducibility analysis with respect to contouring differences) available on the TCIA (The Cancer Image Archive)[133].

The stability of radiomic features is evaluated using the ICC (Intraclass Correlation Coefficient), determining the optimal threshold using data driven methods which allows to balance both prognostic value and stability of the features[134].

2.3 Outcome definition

2.3.1 Primary and secondary endpoints

The co-primary outcomes are OS and PFS in the overall cohort and in subgroups based on different first-line treatments. The endpoint OS is defined as time from treatment initiation (recorded in eCRF) to date of death (if deceased) or date of last follow-up (if alive). The endpoint PFS is defined as time from treatment initiation (recorded in eCRF) to date of progression (recorded in eCRF) or date of death (if no progression has occurred before date of death).

The secondary binary outcomes are

- Early progressors, defined as patients experiencing progressive disease as best response upon first-line treatment (according to investigator's judgement) or deceasing before response evaluation;
- Grade 3 or 4 toxicities occurring upon first-line treatment as defined in CTCAE[135], with a special interest on immune-related pneumonitis;
- Long-term survivors, defined as patients achieving an overall survival of 30 months (1.5 times x median OS in clinical trials[28,37,38]) or more, and long-term responders, defined as patients achieving a PFS of 24 months or more.

2.3.2 Definition of event, censoring, and follow-up

Events are recorded in the eCRF based on the investigator's judgment. No central review of CT scans is planned.

Patients not experiencing the event of interest at last follow-up date are censored for the specific event beyond that date (administrative censoring).

Availability of follow-up status is a key inclusion criterion of the retrospective phase, and all patients have to be tracked for their vital status (alive or dead, with relative last follow-up date or date of death) at the time of the data cutoff. For patients enrolled in the prospective phase, follow-up per clinical practice is recorded until death or consent withdrawal.

2.4 Statistical analysis

2.4.1 Exploratory statistical analyses

Baseline characteristics of the variables collected in the dataset are described as frequencies (for categorical variables), median values (and interquartile range) or mean values (and standard deviation), when appropriate. Chi-squared test or Mann-Whitney U-test have been applied for group comparison, when appropriate.

OS and PFS have been estimated using the Kaplan-Meier method and compared between groups with Log-rank test, when appropriate. Median follow-up has been estimated with reverse Kaplan-Meier method[136].

Multivariate analyses for OS and PFS have been performed using the Cox proportional hazard (PH) regression model. Before testing,

- All non-binary categorical covariates have been binarized as follows: age under or over 70 years, never smoker vs current/former smoker, ECOG PS of 0 or ≥1, BMI under or over 18.5, steroid use yes/no, histology squamous or non-squamous, M stage M0+M1a or M1b+M1c;
- All continuous covariates have been standardized so that each value corresponds to a variable with a mean of 0 and a standard deviation of 1;
- Imputation techniques have been adopted: data have been input if the covariate had less than 40% of missing data (threshold chosen via a 5-fold cross-validation model and via C-index comparison), otherwise the covariate is deleted from the model. For categorical covariates, missing values have been imputed to the mode of the distribution. For continuous covariates: missing values have been imputed to the median of the distribution. To note, for PD-L1 expression: missing values have been imputed to the median of the corresponding treatment group (immunotherapy alone or in combination with chemotherapy).

To optimize the performance of the Cox models (maximal performance with minimal input), a Stepwise Forward Selection algorithm has been used, with a greedy approach to reduce computational load (it selects the best option at each step without searching for the global optimum). 5-fold cross-validation and median C-index are used for model comparison.

2.4.2 Software used for statistical analyses

Statistical analyses have been performed on Python with the following libraries: *pandas, numpy* (exploratory data analysis and descriptive statistics), *matplotlib* (graphical representation), *lifelines* (survival analyses), *sklearn* (k-fold cross-validation), in the Azure AI, Machine Learning Studio.

2.5 ML model development

2.5.1 Algorithms tested

We use an Azure ML pipeline to orchestrate the complete Machine Learning workflow, providing:

- End-to-end visual data lineage for reproducibility and transparency;
- Distributed computing leveraging cluster resources for parallel model training;
- Native MLflow integration for tracking configurations, hyperparameters, trained models, and performance metrics.

The first pre-analytical step is the dataset creation from REDCap (clinical data and laboratory data for external centers), integrating laboratory measurements for the internal cohort, followed by consistency checks on target variables (time-to-event and event occurrence) and administrative censoring (set at 60 months). The second step is removing from the dataset all the variables that are not populated in at least 20% of the population to ensure consistency. Then we move to model building:

- Models from the XGBSE library[137] and RandomSurvivalForest from scikit-survival[138] are trained using Nested Cross-Validation, which include (i) an inner loop for hyperparameter optimization using Bayesian sampling (Optuna library), and (ii) an outer loop for unbiased generalization estimates, followed by Leave-One-Center-Out validation for the best performing models;
- Transformer-based models[139] are trained following only the Leave-One-Center-Out validation, where each of the external centers' dataset serves as external validation dataset once, ensuring comprehensive geographic validation[140].

All models train in parallel and handle missing values natively (with some limitations for RandomSurvivalForest). The external validation on the unseen dataset yields a TRIPOD Type 3 external validation[141].

The pipeline configuration (stored in YAML) varies across multiple dimensions to explore different modelling scenarios, in particular feature robustness levels (ranked by clinicians): group 1 (highly objective features available as RWD, 26 fields), group 2 (group 1 + moderately objective features, 5 fields), group 3 (group 1 + group 2 + subjective features or those less commonly reported in retrospective studies, 3 fields).

For binary classification models, multiple algorithms in Auto-ML have been tested according to performance and research question (XGBoost, RandomForest, Ensamble).

Unsupervised algorithms (such as K-means clustering with Elbow method) have been used for exploratory analysis only.

All ML models are developed within the Azure AI, Machine Learning Studio, on the S-RACE platform.

2.5.2 Performance metrics

For time-to-event modelling (OS and PFS), model performance has been evaluated via concordance index (also called Harrell's C-index) and integrated Brier score, whenever possible. The C-index measures model discrimination (i.e. probability that, for two patients, the one with higher predicted risk experiences the event earlier), while the Brier Score measures how close predicted probabilities are to actual outcomes and the integrated Brier score averages this measure over time.

For binary classification models, model performance has been evaluated via AUC (Area Under the ROC Curve), which measures the model's ability to distinguish between classes. Additional metrics are sensitivity, specificity, and F1-score.

Chapter 3: Results

3.1 Feasibility of ML models in a single-center experience

3.1.1 Aims and methods

To assess the strengths and pitfalls of ML models applied on a real-world basis, we conducted an internal feasibility study focused on 1) limits of real-world data collection, 2) consistency of data when compared to literature, 3) ML models choice for healthcare.

This sub-study was retrospective and single center. Eligibility criteria were as follows: diagnosis of advanced or metastatic NSCLC (stage IIIC or IV according to TNM 8th ed.), treatment with at least one dose of first-line immunotherapy with pembrolizumab (between May 2017 and December 2021), PD-L1 expression on tumor cells of 50% or higher, a follow-up of at least 1 year at time of database lock on December, 31st, 2022.

Binary classification between progressive disease (PD) and no progressive disease (no-PD) at first restaging (per investigator judgment) was the selected endpoint for this preliminary work (also defined as "early progressors", see Outcome definition section 2.3).

First, we started with an Exploratory Data Analysis (EDA) and statistical univariate and multivariate analysis, to investigate associations among variables and the outcome of interest. Then, we moved into ML "black-box" experiments, leveraging on Azure Automated Machine Learning (Auto-ML). Auto-ML is an in-built tool within Azure Machine Learning that rapidly iterates over many combinations of algorithms (XGBoost, Ensemble models, Logistic regressions etc.) and hyper-parameters to help finding the best model based on a performance metric of choice (precision in our case). Last steps were "white-box" experiments (such as SHAP and Decision Trees). Missing data (NaN) were not imputed, and variables with more than 20% of missing data were excluded from ML models.

The final dataset for analysis was consistent with 123 patients with clinical and pathological data, and 106 patients with clinical, pathological and laboratory data (not all patients had laboratory data available for retrieval in the pre-specified timeframe). Clinical and pathological data were manually collected (on a local eCRF), laboratory data

were automatically collected from hospital laboratory software (see Data collection section 2.2 for details).

3.1.2 Exploratory data analysis

Baseline characteristics are summarized in *Table 3*.

Sex, female/male (N, %)	41/82 (33%/67%)
Age, years (median, IQR)	70 (62-74)
Baseline PS ECOG (N, %)	
0	44 (35%)
1	68 (55%)
≥2	11 (10%)
Histological type (N, %)	
Non-squamous	105 (85%)
Squamous	18 (15%)
Stage (N, %)	
IIIB	2 (1.5%)
IIIC	2 (1.5%)
IV *de novo*	99 (81%)
IV *relapsed*	20 (16%)
Current or former smoker (N, %)	103 (84%)
BMI, kg/m^2 (median, q1-q3)	24.6 (21.8-26.6)
PD-L1, % (median, q1-q3)	70 (60-90)

Table 3. Baseline characteristics

Our population's outcomes, according to RECIST 1.1 criteria at first restaging, were as follows: PD 28/106 (26%) or death 24/106 (23%), PR 33/106 (31%), SD 16/106 (15%), CR 5/106 (5%), resulting in 52 patients with PD or death (49%, PD group) and 54 without PD (51%, merged as no-PD group).

On laboratory data, neutrophil percentage and neutrophil/lymphocyte ratio (NLR) were significantly different between patients in the PD group and those in the no-PD group (p<0.05). Focusing on symptoms at baseline (before treatment start), asymptomatic

patients (referring to cancer-associated symptoms i.e. dyspnea, pain, weight loss) were more likely to not experience an early progression compared to symptomatic patients (p<0.05). For PD-L1 expression, we compared higher values (from 75 to 100%) to lower values (from 50 to 74%). In our population, 65 patients out of 123 had values under 75% (53%, lower group) and 58 had tumors with values above 75% (47%, higher group). Most patients in the PD-L1 higher group, did not experience early progression (60%). On the other hand, in the lower PD-L1 group, the percentage of patients experiencing early-PD was higher compared to the no-PD group (55% versus 45%, p<0.05). The impact of KRAS mutations on survival could not be estimated due to high proportion of missing values (40%), of which only 37% was in tumors with squamous histology that should not undergo genomic testing per international guidelines.

Exploratory longitudinal description of relevant laboratory data was performed, as most of the patients had laboratory tests available at different timepoints after treatment initiation. At the same time, hospital accesses were easily and automatically mapped from hospital software, so we could collect the exact number of accesses for each patient. On a time-span of 10 to 15 weeks, patients in the PD group had median values of NLR consistently higher than those in the no-PD group, as well as more frequent hospital visits, as shown in *Figure 5a* and *5b*.

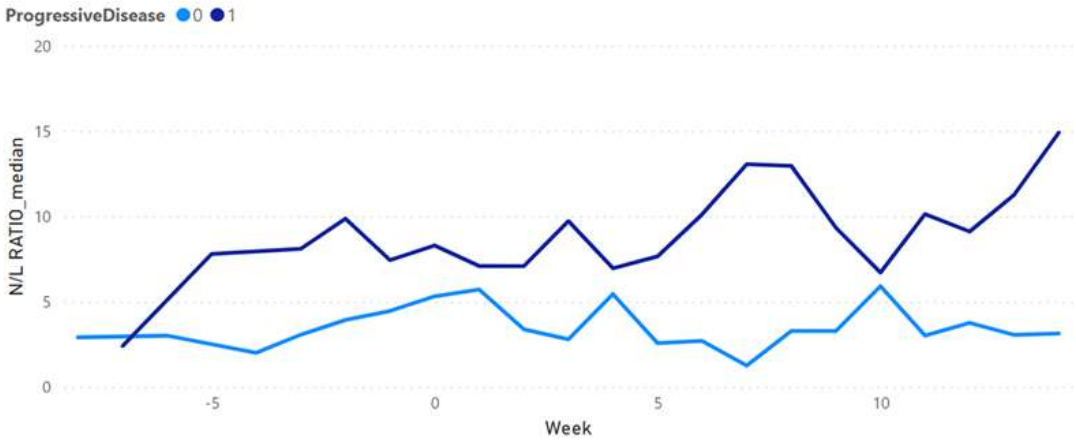

Figure 5a. NLR ratio median values by week and PD status (see Legenda)

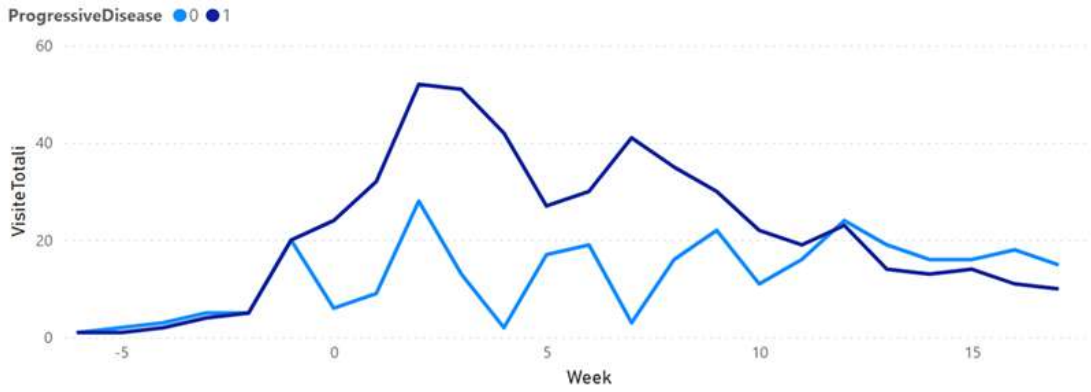

Figure 5b. Number of hospital visits by week and PD status (see Legenda)

We then investigated the optimal threshold values for the significant variables, in order to define a clear distinction between outcomes (PD and no-PD). The goal was to find significant variables with value ranges able to clearly assign each sample to either one of the two different outcomes, with a narrow confidence interval, so that the information could be used as a decision-making factor. In *Table 4*, we summarized variables whose cut-offs (in terms of percentiles) showed good separation between the two populations (PD and no-PD), with no overlapping between confidence intervals.

Variable	**Quartile [value]**	**no-PD group**	**PD-group**	**Probability of no-PD [CI]**	**Probability of PD [CI]**
Neutrophils (%)	75[81.45]	71	24	0.352 [(0.242, 0.475)]	0.917 [(0.73, 0.99)]
Lymphocytes (%)	25[10.25]	24	71	0.875 [(0.676, 0.973)]	0.366 [(0.255, 0.489)]
NLR	75[8.008]	71	24	0.366 [(0.255, 0.489)]	0.875 [(0.676, 0.973)]
PD-L1 expression (%)	50[70.0]	56	50	0.607 [(0.468, 0.735)]	0.36 [(0.229, 0.508)]

Table 4. Threshold (in quartiles) for group separation according to outcome (PD vs no-PD)

3.1.3 ML models building

We summarized our approaches in terms of ML models as follows:

1. On full dataset
2. On dataset with laboratory data only
3. On dataset with laboratory and pathological data
4. On selected features (from EDA and other models)

The first test was running Auto-ML on full dataset, without any *a priori* clinical overview (after first eCRF completion). The best model was Voting Ensemble (AUC 0.78), and all the other ML models tested had similar performances [0.76-0.69]. The features that were selected as most relevant were all laboratory and pathological data, such as neutrophil percentage, platelets and PD-L1 expression. These features recall the most important features emerged from EDA, and clinical data did not come up as impactful. Despite being statistically solid, these results did not catch the clinical complexity of this population, resulting in a lack of trust from the user's perspective (clinicians). Therefore, a second check from a clinician was applied, in particular manually re-collecting symptoms (not only yes/no, but detailed such as cough, dyspnea, pain etc) and concomitant medications (not fully captured during first collection due to inconsistencies in EHRs). After this step of data curation, we run Auto-ML again, achieving better performances with the same selected model (AUC 0.80 with Voting Ensemble), which statistically outperformed all the other classifiers (p<0.05, see *Table 5a* for details). All the other ML algorithms had performances in the range [0.75-0.71] and higher compared to non-curated full dataset. The most important features emerging were laboratory (neutrophil percentage, platelets, lymphocyte percentage, NLR), PD-L1 expression and clinical features such as performance status at baseline and use of steroids.

Algorithm name	**AUC**
Voting Ensemble	0.80320
StandardScalerWrapper, Random Forest	0.75534
SparseNormalizer, LightGBM	0.74802
StackEnsemble	0.72758
SparseNormalizer, XGBoostClassifier	0.72165

Top ranked features
NeutroPerc
Platelets
PD-L1 expression
PS at baseline
Bilirubin
Hypertension
Use of steroids
C-reactive protein
LymphoPerc
NLR

Table 5a. AutoML performance on full dataset

On laboratory data only, the best model was again Voting Ensemble (AUC 0.75), which outperformed the other classifiers (p<0.05) and confirmed neutrophil percentage, platelets, leucocyte count and hemoglobin as most informative features, see *Table 5b* for details.

Algorithm name	**AUC**
Voting Ensemble	0.74756
MaxAbsScaler, LightGBM	0.71890
StandardScalerWrapper, XGBoostClassifier	0.71591
RobustScaler, KNN	0.70158
StandardScalerWrapper, LightGBM	0.70152
RobustScaler, RandomForest	0.69801
Top ranked features	
NeutroPerc	
Platelets	
Leucocytes	
Haemoglobin	
Bilirubin	
Alanine-aminotransferase	
C-reactive protein	
Aspartate-aminotransferase	
NLR	
LymphoPerc	

Table 5b. AutoML performance on laboratory data only

The Auto-ML test on pathological and laboratory data (clinical data excluded) outperformed the results on laboratory data only: AUC 0.79 (p<0.05). Again, the best performing model was a Voting Ensemble and the other classifiers had precision scores in the range [0.76-0.71]. The only non-laboratory feature ranked, in third place, was PD-L1 expression. Results are summarized in *Table 5c*.

Algorithm name	AUC
Voting Ensemble	0.79603
RobustScaler, KNN	0.76540
SparseNormalizer, ExtremeRandomTrees	0.75483
MaxAbsScaler, LightGBM	0.74726
TruncatedSVDWrapper, RandomForest	0.74151
SparseNormalizer, LightGBM	0.73836
Top ranked features NeutroPerc Platelets PD-L1 expression Bilirubin Leucocytes NLR Haemoglobin C-reactive protein Alanine-aminotransferase LymphoPerc	

Table 5c. AutoML performance on laboratory and pathological data

Finally, the last Auto-ML test included all the features that were found significant during the EDA and those emerging from other models. The top performing model (Voting Ensemble) had the highest AUC of 0.82 among all the experiments (p<0.05). The most important features still were laboratory data except for PD-L1 expression and hypertension (probably due to overfitting of this last feature), see *Table 5d* for details.

Algorithm name	AUC
Voting Ensemble	0.82504
SparseNormalizer, XGBoostClassifier	0.75635
SparseNormalizer, LightGBM	0.75198
SparseNormalizer, RandomForest	0.74556
MaxAbsScaler, ExtremeRandomTrees	0.73247
StandardScalerWrapper, LightGBM	0.72992
Top ranked features NeutroPerc PD-L1 expression Hypertension Platelets LymphoPerc Bilirubin C-reactive protein	

Table 5d. AutoML performance on selected features

Auto-ML models showed that our input data were reliable and their high performance-scores further supported the good-quality of extracted data. Therefore, we proceeded with

the addition of explainable tools on top of Auto-ML models. The first methodology that we applied was SHAP on all the previously tested datasets. To note: running SHAP models, the ranking of features slightly differed from the one originated from AutoML. SHAP results are summarized in *Figures 6* (*a-d*).

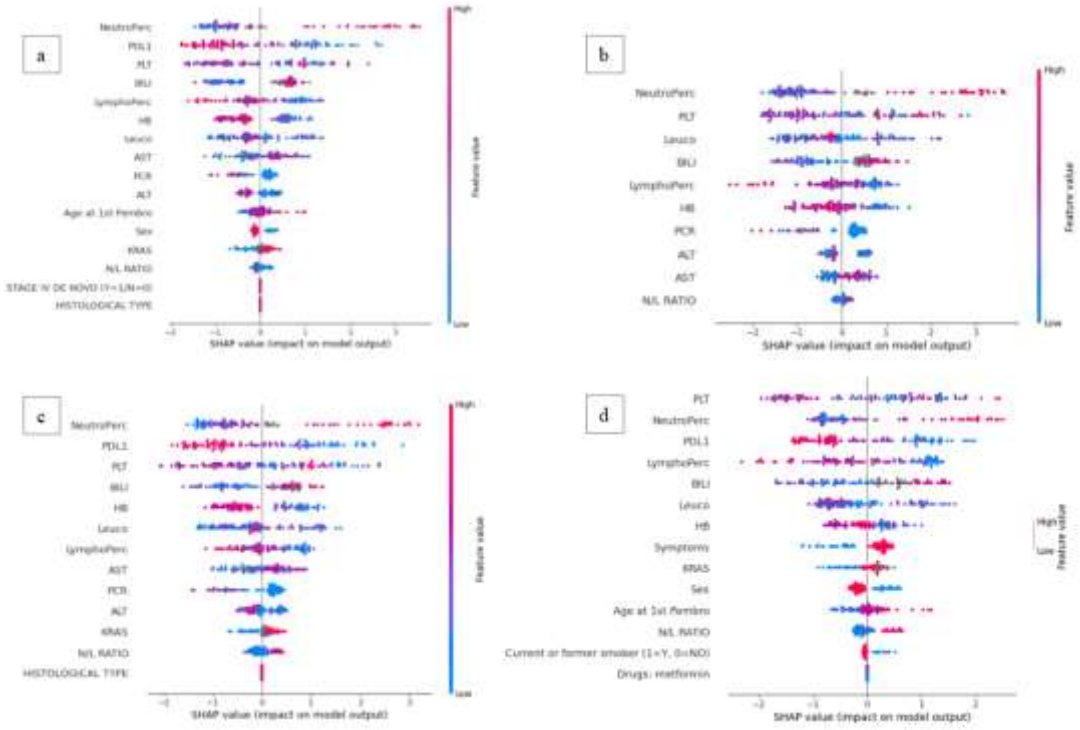

Figure 6. SHAP performance on different datasets, a) on full dataset, b) on dataset with laboratory data only, c) on dataset with laboratory and pathological data, d) on selected features (from EDA and other models)

Another white-box model used in healthcare is the Decision Tree. The model works by asking questions and splitting the dataset according to these questions, where each question can be seen as a branch of the tree and each answer (yes or no) separates the sample according to its behavior (in our case, PD versus no-PD). We performed a Decision Tree on our dataset (all data), as shown in *Figure 7*. The set of rules the algorithm found can be immediately observed and are consistent with AutoML results (top features: neutrophil percentage, leucocyte count, platelets, PD-L1 expression).

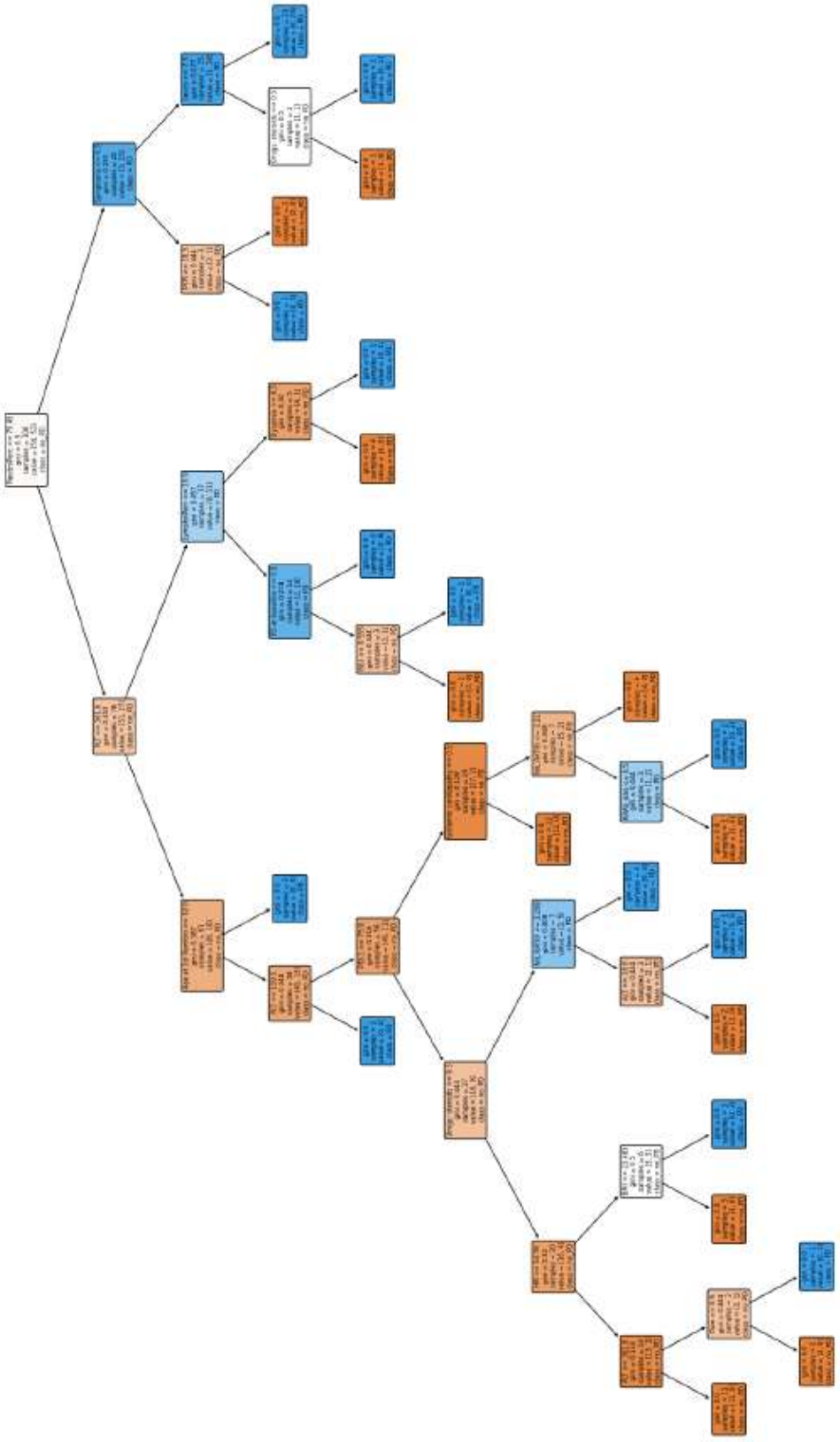

Figure 7. Decision Tree built on full dataset

3.2 ML models encompassing radiomic features: an overview

3.2.1 Aims and methods

As discussed before, radiomics has not entered the clinical workflows yet because of reproducibility issues and lack of consistency, above all in small datasets. Therefore, we explored different approaches for radiomic features inclusion in ML models for outcome prediction in a population with advanced NSCLC treated with first line immunotherapy.

This sub-study was retrospective and bicentric (Ospedale San Raffaele and Spedali Civili di Brescia). Eligibility criteria were as follows: diagnosis of advanced or metastatic NSCLC (stage IIIC or IV according to TNM 8th ed.), treatment with at least one dose of first-line immunotherapy or chemo-immunotherapy (between May 2017 and December 2022), a follow-up of at least 1 year at time of database lock on December, 31st, 2023.

Binary classification between progressive disease (PD) and no progressive disease (no-PD) at first restaging (per investigator judgment) was the selected endpoint for this preliminary work (also defined as "early progressors", see Outcome definition section 2.3).

First, we collected clinical and laboratory data (as per the previously presented sub-study), and baseline thorax CT-scans. For this sub-study, the primary tumor was manually contoured on baseline imaging by one expert radiologist (for Ospedale San Raffaele) and by one or two expert radiologists (for ASST Spedali Civili). Radiomic features were extracted via PyRadiomics (see Radiological data and radiomics pipeline section 2.2.3).

In total, 100 features were initially extracted from each volume of interest (VOI), encompassing a variety of classes (*first-order, shape-based, texture-based*). After extraction, a feature reduction process was applied to mitigate the risk of overfitting due to the high dimensionality of the data relative to the limited sample size. This involved removing features with strong intercorrelation, defined by a Spearman correlation coefficient of greater than 0.8. Features that demonstrated volume dependence were also eliminated to ensure the final feature set was independent of tumor size, as size is often an inherent confounder in radiomics analyses. Following this selection process, a total of 20 radiomic features were retained for further analysis. The final set of features is summarized in *Table 6*.

First-order statistics	Shape-based features	Texture features
10Percentile	Elongation	GLCM_Cluster Prominence
90Percentile	Flatness	GLCM_Correlation
Kurtosis	Least Axis	GLCM_Imc1
Maximum	Length	GLCM_Idmn
Mean	Sphericity	GLSZM_SmallAreaEmphasis
Minimum		GLSZM_SizeZoneNonUniformityNormalized
Skewness		GLSZM_Zone Entropy
Uniformity		GLDM_Dependence Entropy

Table 6. Final set of radiomic features for analysis

3.2.2 Exploratory data analysis

A total of 76 patients meeting inclusion and exclusion criteria were included in the exploratory data analysis cohort (Ospedale San Raffaele, OSR, only). Baseline characteristics are summarized in *Table 7*.

Age, median (interquartile range)	70 years (IQR 63.7-74.0)
Sex (n, %)	
Female	18 (23%)
Male	58 (77%)
Smoking (n, %)	
Current/former	70 (92%)
Never	6 (8%)
ECOG Performance Status (n, %)	
ECOG 0	38 (50%)
ECOG 1	34 (45%)
ECOG ≥ 2	4 (5%)
Histology (n, %)	
Non-squamous	65 (86%)
Squamous	11 (14%)
PD-L1 expression (TPS) (n, %)	
< 1 %	17 (22%)

1-49%	16 (21%)
≥50%	39 (51%)
Unknown	4 (6%)
Treatment (n, %)	
Pembrolizumab	38 (50%)
Pembrolizumab + chemotherapy	38 (50%)

Table 7. Baseline characteristics of the 76 included patients

Among the PD versus no-PD groups, two texture-related features demonstrated statistically significant differences, specifically GLSZM SizeZoneNonUniformityNormalized (p=0.0255) and GLSZM SmallAreaEmphasis (p=0.0263). Patients in the PD-group had lower median values for both features compared to those in the no-PD group.

3.2.3 ML models building: unsupervised approach

Due to limited sample size, on the OSR cohort, we tried to apply unsupervised models, which are less explainable, but potentially powerful in case of limited datasets. Our aim was to test if in radiomic features only (without clinical data and without labelled outcomes) there were any signals of disease evolution (i.e. predicting PD).

We tested a clustering approach, with K-means clustering algorithm, using the Elbow method for identifying the optimal number of clusters. In the OSR cohort, the "elbow" point is observed at 3 clusters, as there is a diminishing return in inertia reduction with additional clusters (see Figure 8).

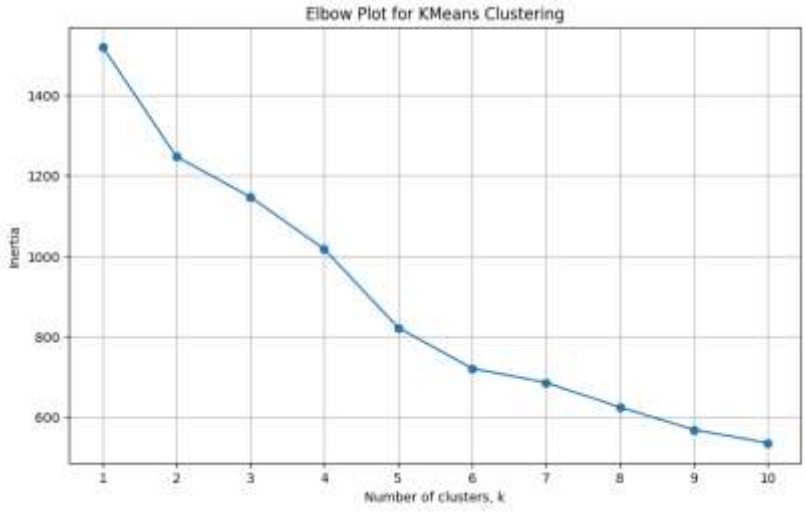

Figure 8. Elbow plot showing the decrease of inertia with increasing number of clusters

The median age across clusters was similar, with Cluster A showing a median of 72.0 years (IQR 66.5–78.0), Cluster B at 69 years (IQR 63.0–71.0), and Cluster C at 71 years (IQR 67.5–74.0). In terms of sex distribution, there were no significant differences, as Cluster A comprised 82% men, Cluster B 74%, and Cluster C 74%. Similarly, smoking status and histology were similarly distributed across clusters. However, a higher proportion of patients in Cluster C (52%) had a performance status ECOG 0, compared to 32% in Cluster A and 26% in Cluster B.

Patients treated with mono-immunotherapy were 41% in Cluster A, 71% in Cluster B and 37% in Cluster C, at least partially mirroring the PD-L1 distribution. PD-L1 expression levels across the three clusters are shown in *Figure 9*. In Cluster A, the median PD-L1 value is around 0.4 (40%), with no relevant outliers. In Cluster B, the median PD-L1 value is slightly higher (0.6, 60%), with only one very low outlier (0.05, 5%). In Cluster C, the median PD-L1 value is around 0.2 (20%), showing the lowest median PD-L1 value compared to the other two clusters.

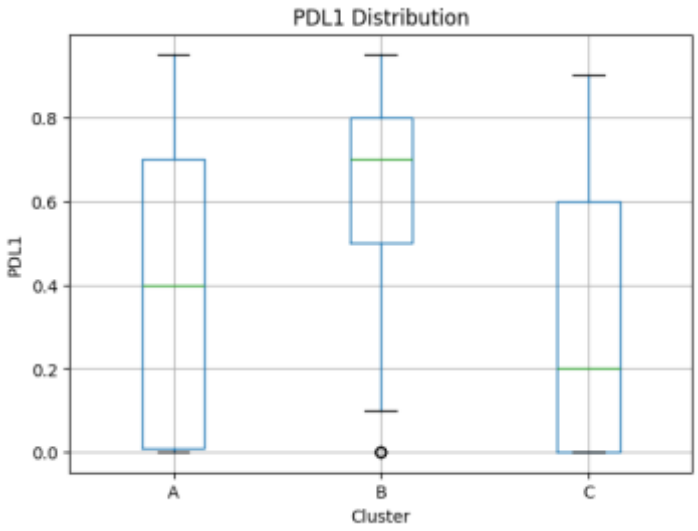

Figure 9. Distribution of PD-L1 expression among Clusters A, B, and C

Considering PD and no-PD group, the distribution within clusters was as follows: Cluster A PD 40% vs no-PD 60%, Cluster B PD 30% vs no-PD 70%, Cluster C PD 30% vs no-PD 70%. When comparing OS, Cluster A demonstrated the poorest survival outcomes with lowest median OS (median of 13.0 months) and a narrower interquartile

range indicating less variability within the cluster. Conversely, Cluster C showed the most favorable survival, with a wider distribution, indicating greater variability but overall better outcomes (median of 19.0 months for all PD-L1 levels). None of these differences tested statistically significant at the chi-square or Log-rank tests.

3.2.4 ML models building: supervised approach

Beyond preliminary unsupervised analysis, we aimed at including radiomic features in ML models for early PD prediction, despite the lack of standardized approaches for radiomic features selection. We therefore compared two different feature selection methods in three datasets (one independent, and two paired) of patients treated with first line mono-immunotherapy for metastatic NSCLC: OSR cohort (36 patients and related baseline CT-scan that were contoured by one radiologist), and Spedali Civili di Brescia (or BS, 112 patients and related baseline CT-scan that were contoured by one radiologist [BS1] and 70 patients and related baseline CT-scan that were contoured by two radiologists, resulting in 2 contours per patient [BS2]). Patients treated with first-line chemo-immunotherapy were excluded from this analysis to ensure homogeneity among the samples.

Two feature selection approaches were compared:

(A) ML-based selection using SelectKBest with 100 Monte Carlo cross-validation (MCCV) simulations to identify the optimal number of features;

(B) Selection based on intraclass correlation coefficient (ICC) thresholds between paired datasets (BS1 and BS2).

Three ML models (Random Forest, Logistic Regression, DeepNet) were used to classify patients in PD or no-PD group.

In Approach A, there is no human supervision, and the ML models automatically choose the most relevant features for the predefined prediction when given tabular data as input data (in this experiment, 8 essential clinical variables – such as age, ECOG PS, TNM staging, location of metastatic sites – and 100 radiomic features). These ML models achieved AUCs between 0.56±0.16 and 0.70±0.18, with DeepNet performing best using fewer features (results summarized in *Table 8*).

	BS1	**BS2**	**OSR**

Random Forest	AUC 0.64 (±0.11) Nr of feat. 23 (±28)	AUC 0.56 (±0.16) Nr of feat. 36 (±34)	AUC 0.58 (±0.20) Nr of feat. 63 (±34)
Logistic Regression	AUC 0.65 (±0.10) Nr of feat. 28 (±28)	AUC 0.60 (±0.13) Nr of feat. 29 (±25)	AUC 0.70 (±0.18) Nr of feat. 71 (±30)
DeepNet	AUC 0.65 (±0.12) Nr of feat. 10 (±6)	AUC 0.65 (±0.15) Nr of feat. 14 (±4)	AUC 0.60 (±0.17) Nr of feat. 14 (±5)

Table 8. Models' performances across different datasets (AUC±SD) and number of features used (±SD) (SD=standard deviation)

The best ranking features were then summarized into heatmaps (see *Figure 10*). It is relevant to underline that feature selection appears to be consistent across different cohorts despite different sample sizes, as brain metastases and Skewness appear in the top ranked features for DeepNet models in all the three datasets.

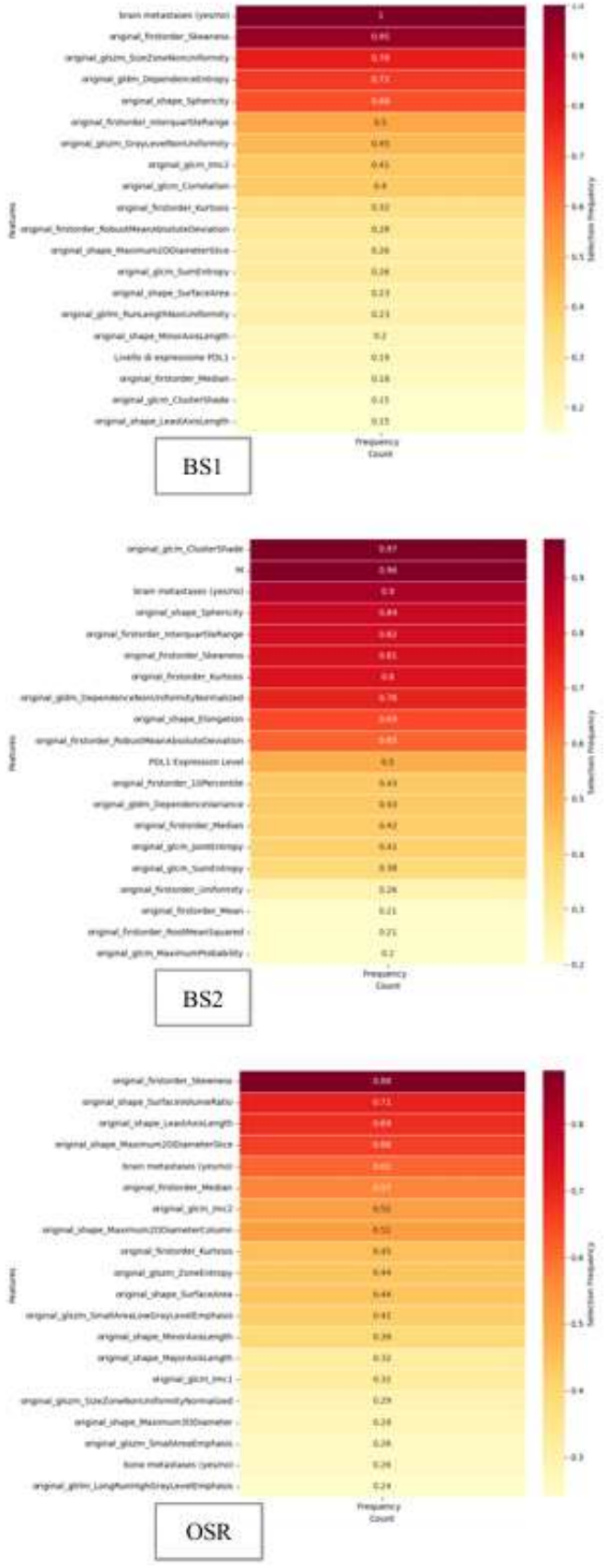

Figure 10. Heatmaps for feature ranking in DeepNet models among the three datasets

In Approach B, selection is based on ICC between the paired datasets (and then applied on the independent dataset). Given that we did not have multiple scans per patient, but only multiple contourings per patient, the ICC measured "reproducibility" (and not "repeatability") across repeated segmentations of the same subject. In radiomic literature, features with ICC≥0.75 (or ≥0.9 in stricter workflows) are typically considered robust enough for further development[142]. We explored different ICC thresholds within our datasets, and selected features were then included in ML models together with clinical variables.

As expected, the number of selected features dropped with increasing ICC thresholds (see Figure *11a* and *11b*).

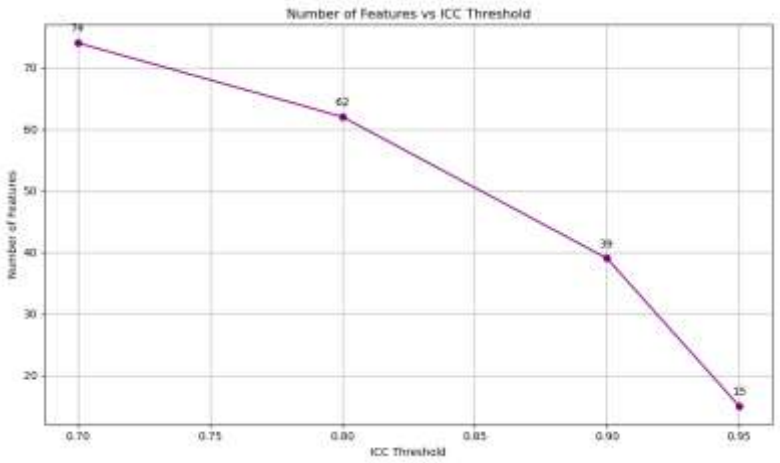

Figure 11a. Number of features according to different ICC thresholds

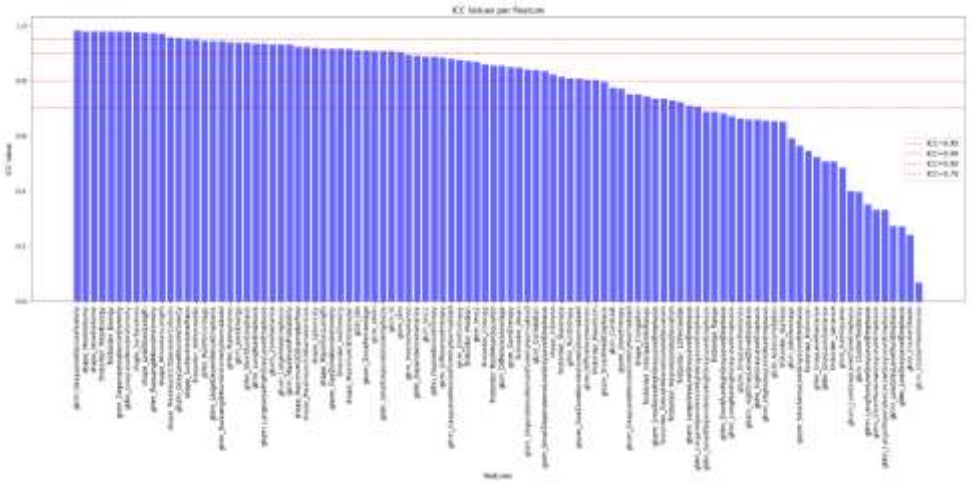

Figure 11b. List of features according to different ICC thresholds

Higher ICC thresholds improved Random Forest performance in larger datasets (BS1 and BS2), but reduced accuracy in smaller datasets (OSR). Logistic Regression and DeepNet performed best at intermediate-low ICC thresholds (0.70 and 0.80) across all datasets. Results are summarized in *Table 9*.

	BS1	**BS2**	**OSR**
RF, ICC 0.70	AUC 0.69 (±0.10)	AUC 0.59 (±0.15)	AUC 0.62 (±0.18)
RF, ICC 0.80	AUC 0.69 (±0.10)	AUC 0.61 (±0.15)	AUC 0.61 (±0.19)
RF, ICC 0.90	AUC 0.72 (±0.09)	AUC 0.61 (±0.15)	AUC 0.60 (±0.17)
RF, ICC 0.95	AUC 0.71 (±0.10)	AUC 0.61 (±0.14)	AUC 0.54 (±0.20)
LR, ICC 0.70	AUC 0.69 (±0.11)	AUC 0.58 (±0.15)	AUC 0.75 (±0.18)
LR, ICC 0.80	AUC 0.69 (±0.11)	AUC 0.57 (±0.15)	AUC 0.74 (±0.18)
LR, ICC 0.90	AUC 0.65 (±0.10)	AUC 0.55 (±0.16)	AUC 0.68 (±0.19)
LR, ICC 0.95	AUC 0.65 (±0.10)	AUC 0.58 (±0.15)	AUC 0.61 (±0.21)
DN, ICC 0.70	AUC 0.69 (±0.11)	AUC 0.60 (±0.15)	AUC 0.76 (±0.17)
DN, ICC 0.80	AUC 0.69 (±0.10)	AUC 0.60 (±0.15)	AUC 0.73 (±0.19)
DN, ICC 0.90	AUC 0.66 (±0.11)	AUC 0.57 (±0.15)	AUC 0.66 (±0.18)
DN, ICC 0.95	AUC 0.63 (±0.11)	AUC 0.58 (±0.14)	AUC 0.63 (±0.20)

Table 9. Models' performances across different datasets (AUC±SD) and ICC-thresholds (SD=standard deviation, RF=Random Forest, LR=Logistic Regression, DN=DeepNet)

In conclusion, models' performances were very similar between ML-based approach and ICC-based approach, with reduced workload in the pre-analytical phase for ML-based models.

3.3 Development and comparison of time-to-event ML models

3.3.1 Aims and methods

Time-to-event ML models are more consistent in clinical practice, thanks to their ability to go beyond binary prediction and generate more individualized predictions. Moreover, their informative content is higher, because they can provide information on a larger timespan rather than on a single timepoint. Therefore, we trained different time-to-event models to ensure robustness and compare performances in our real-world cohort of patients with metastatic NSCLC treated with first line immunotherapy.

This sub-study was retrospective and multicentric (Ospedale San Raffaele [OSR], Maastricht University Medical Center [UMC], Maxima MC and VieCuri Medisch Centrum, from now on summarized as the Dutch centers). Eligibility criteria were as follows: diagnosis of advanced or metastatic NSCLC (stage IIIC or IV according to TNM 8th ed.), treatment with at least one dose of first-line immunotherapy (between May 2017 and December 2023), a follow-up of at least 1 year at time of database lock on December, 31st, 2024.

The co-primary endpoints were OS and PFS in the overall population (see Outcome definition section 2.3).

First, we started with an Exploratory Data Analysis (EDA) and statistical univariate and multivariate analysis, to investigate associations among variables and the outcome of interest. Then, we moved into time-to-event modelling (see ML model development section 2.5) and descriptively compare models' performances. Missing data (NaN) were handled as described in Statistical analyses section 2.4 and ML model development section 2.5.

The final dataset for analysis was consistent with 498 patients with tabular data only (clinical, pathological, laboratory). Clinical and pathological data were manually collected (on a local eCRF), laboratory data were automatically collected from hospital laboratory software for OSR cohort and manually in the eCRF for external centers (see Data collection section 2.2 for details).

3.3.2 Exploratory data analysis

Baseline characteristics are summarized in *Table 10*.

Age, classes (n, %)	
36-60 years	113 (23%)
61-70 years	184 (37%)
71-80 years	169 (34%)
≥81 years	32 (6%)
Sex (n, %)	
Female	181 (36%)
Male	317 (64%)
Smoking (n, %)	
Current/former smoker	428 (86%)
Never smoker	27 (5%)
Unknown	43 (9%)
ECOG Performance Status (n, %)	
ECOG 0	192 (38%)
ECOG 1	243 (49%)
ECOG ≥ 2	63 (13%)
Histology (n, %)	
Non-squamous	421 (85%)
Squamous	77 (15%)
Stage IV *de novo*	444 (89%)
Stage IV *relapsed*	54 (11%)
M1 status (n, %)	
M0	54 (11%)
M1a	94 (19%)
M1b	85 (17%)
M1c	265 (53%)
Sites of metastases (n, %)	
Bone	200 (40%)
Lung	170 (34%)
Lymphnodes	146 (29%)
Pleura	143 (29%)

Brain	114 (23%)
Adrenal glands	104 (21%)
Liver	77 (15%)
PD-L1 expression (TPS) (n, %)	
< 1 %	74 (15%)
1-49%	160 (32%)
≥ 50%	259 (52%)
Unknown	5 (1%)
Treatment (n, %)	
Pembrolizumab	242 (49%)
Pembrolizumab + chemotherapy	256 (51%)

Table 10. Baseline characteristics of the 498 included patients

We conducted a comparison analysis between OSR and the Dutch centres, to assess differences in baseline characteristics between the two cohorts. Age, sex and ECOG PS were differently distributed, in particular OSR patients tend to be older (age > 70 years: 45% vs 32% in the Dutch centers, p=0.01), more frequently male (male: 69% vs 56% in the Dutch centers, p<0.01) and with poorer ECOG PS (PS ≥ 1: 69% vs 51% in the Dutch centers, p<0.01).

In the overall cohort, median OS was 13 months (95% CI, 10.5-15.5) and median PFS was 5.9 months (95%CI, 5.1-6.8), as shown in *Figure 12a-b*.

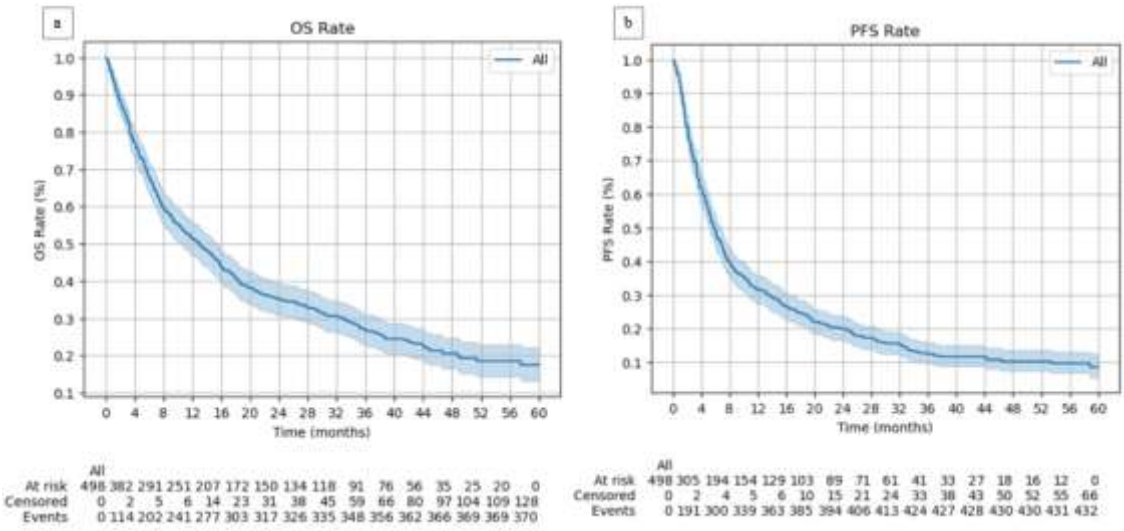

Figure 12. Kaplan-Meier curves for OS (a) and PFS (b) in the overall population

Survivals have also been estimated separately for OSR and the Dutch centers, showing a statistically significant difference in both OS and PFS in favour of the Dutch cohort (OSR median OS 10.7 months, 95% CI 7.8-13.5, vs Dutch centers median OS 15.9 months, 95% CI 12.3-20.5, *p*=0.03; OSR median PFS 4.9 months, 95% CI 4.0-6.0, vs Dutch centers median OS 7.1 months, 95% CI 5.8-9.4, *p*=0.02). Survival curves are reported in *Figure 13a-b*.

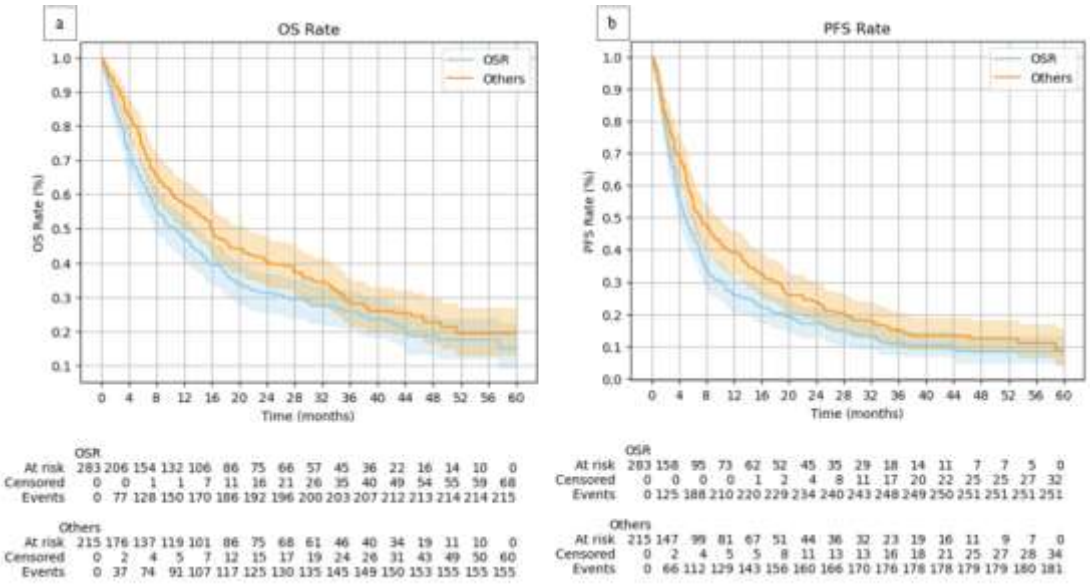

Figure 13. Kaplan-Meier curves for OS (a) and PFS (b) in OSR and Dutch centers cohorts

We then evaluated the impact of different variables on survival with Cox regression analyses. Multivariate analysis results for OS indicate ECOG PS, BMI, statin use, PD-L1 expression, metastatic burden, neutrophil count, hemoglobin and LDH as relevant prognostic factors, results are summarized in *Table 11.* The median C-index for Cox regression model for OS was 0.65, and a reduced model with only 4 variables (ECOG PS, metastatic burden, neutrophil count and LDH) reached the same performance (median C-index 0.65).

Variable	**HR (95% CI)**	***p*-value**
Age (ref: ≤ 70 years)	1.14 (0.91-1.42)	0.25
Sex (ref: male)	0.81 (0.64-1.03)	0.08
Smoking status (ref: Never smoker)	0.65 (0.42-1.02)	0.06

ECOG PS (ref: 0)	1.78 (1.42-2.23)	**<0.005**
BMI (ref: < 18.5)	0.62 (0.42-0.92)	**0.02**
Use of statins (ref: no)	0.71 (0.55-0.91)	**0.01**
Use of steroids (ref: no)	1.37 (0.91-2.05)	0.13
Histology (ref: Non-squamous)	0.95 (0.70-1.29)	0.77
PD-L1 expression	0.77 (0.69-0.85)	**<0.005**
KRAS mutation (ref: no)	1.06 (0.84-1.33)	0.64
M status (ref: M0+M1a)	1.44 (1.11-1.87)	**0.01**
White blood cell count	1.04 (0.95-1.14)	0.41
Neutrophil count	1.17 (1.04-1.3)	**0.01**
Platelet count	0.92 (0.81-1.04)	0.18
Hemoglobin	0.86 (0.76-0.98)	**0.03**
LDH	1.11 (1.03-1.2)	**0.01**
Creatinine	1.03 (0.92-1.16)	0.61

Table 11. Multivariate analysis for OS

For PFS, male sex, smoking status, ECOG PS, use of statins and steroids, PD-L1 expression, neutrophil count, hemoglobin and LDH levels emerged as significantly correlated with the outcome (see *Table 12*). The Cox regression model for PFS achieved a C-index of 0.63, while the reduced 3-variable model (sex, ECOG PS, neutrophil count) reached a median C-index of 0.61.

Variable	**HR (95% CI)**	***p*-value**
Age (ref: ≤ 70 years)	1.07 (0.87-1.31)	0.51
Sex (ref: male)	0.76 (0.61-0.95)	**0.01**
Smoking status (ref: Never smoker)	0.63 (0.42-0.95)	**0.03**

ECOG PS (ref: 0)	1.42 (1.16-1.75)	**<0.005**
BMI (ref: < 18.5)	0.69 (0.48-1.01)	0.05
Use of statins (ref: no)	0.71 (0.56-0.90)	**0.01**
Use of steroids (ref: no)	1.85 (1.28-2.68)	**<0.005**
Histology (ref: Non-squamous)	0.84 (0.63-1.12)	0.23
PD-L1 expression	0.81 (0.74-0.90)	**<0.005**
KRAS mutation (ref: no)	1.17 (0.94-1.45)	0.15
M status (ref: M0+M1a)	1.21 (0.96-1.54)	0.11
White blood cell count	1.02 (0.93-1.11)	0.71
Neutrophil count	1.16 (1.04-1.29)	**0.01**
Platelet count	0.94 (0.83-1.07)	0.35
Hemoglobin	0.89 (0.79-1.00)	**0.04**
LDH	1.08 (1.01-1.16)	**0.03**
Creatinine	1.00 (0.89-1.12)	1.0

Table 12. Multivariate analysis for PFS

3.3.3 Time-to-event models building

We then moved to ML models building, according to Methods described in ML Model development section 2.5. In particular, we compared performances of transformer-based models, models coming from the XGBSE library and a RandomSurvivalForest, as benchmark, for OS prediction. We used, as input data, tabular data only (clinical and laboratory), ranked for feature robustness as described in ML Model development section 2.5. Due to imbalanced variables among different centers (as previously reported), we proceeded with training and validation according to the Leave-One-Center-Out (LOCO) method.

In *Table 13*, we report the sample sizes for training and validation cohorts according to different LOCO experiments.

Center-out	Ospedale San Raffaele	Maastricht UMC	Maxima MC	Viecuri Medisch Centrum
Samples				
Training	215	428	416	435
Validation	283	70	82	63
Events				
Training	157	322	319	321
Validation	216	51	54	52

Table 13. Sample size for training and validation cohorts according to different LOCO experiments

The first models to be tested were 4 models from the XGBSE library (XGBSE KaplanTree, XGBSE KaplanNeighbors, XGBSE DebiasedBCE, XGBSE StackedWeibull) and 1 RandomForest for survival analysis, results are summarized in *Table 14*.

Center-out / Models	Ospedale San Raffaele	Maastricht UMC	Maxima MC	Viecuri Medisch Centrum
group1 features				
XGBSE-KT	0.54	0.60	0.57	0.63
XGBSE-KN	0.57	0.63	0.62	0.63
XGBSE-DE	0.57	0.66	0.63	0.65
XGBSE-SW	0.60	0.66	0.61	0.65
RF-SA	0.60	0.67	0.54	0.65
group1+2 features				
XGBSE-KT	0.53	0.61	0.56	0.64
XGBSE-KN	0.58	0.63	0.60	0.66
XGBSE-DE	0.56	0.66	0.63	0.64
XGBSE-SW	0.60	0.65	0.64	0.64
RF-SA	0.59	0.68	0.64	0.65
group1+2+3 features				
XGBSE-KT	0.53	0.61	0.56	0.64
XGBSE-KN	0.58	0.63	0.63	0.65
XGBSE-DE	0.53	0.66	0.64	0.66
XGBSE-SW	0.57	0.64	0.63	0.64
RF-SA	0.58	0.66	0.65	0.66

Table 14. Models' C-index values according to different feature groups and different LOCO experiments in Nested Cross Validation (XGBSE-KT=XGBSE KaplanTree, XGBSE-

KN=XGBSE KaplanNeighbors, XGBSE-DE=XGBSE DebiasedBCE, XGBSE-SW=XGBSE StackedWeibull, RF-SA=RandomForest Survival Analysis)

Best performing model on group1 features was RandomForest SurvivalAnalysis, based on mean C-index of outer loop of Nested Cross Validation. It was then re-validated on external datasets and reached C-index values ranging from 0.71 (if VieCurie was Center-out) to 0.79 (if OSR was Center-out). Considering group1+2 features, RandomForest SurvivalAnalysis, XGBSE KaplanNeighbors and XGBSE StackedWeibull emerged as the best performing models, and their external validation achieved the highest C-index among all the experiments (0.83 if OSR was Center-out). Lastly, including group1+2+3 features, best performing models were again RandomForest SurvivalAnalysis and XGBSE DebiasedBCE, reaching a C-index of 0.79 with OSR as external validation dataset.

We then moved to transformer-based models. These models achieved mediocre performances in terms of C-index across all LOCO experiments and showed no improvement according to different feature input (group1, group1+2, or group 1+2+3). Interestingly, global C-index was higher with only group1 features (mean global C-index 0.58) than with group1+2 or group 1+2+3 (mean global C-index 0.54), and the integrated Brier score (iBS) resulted lower (mean iBS 0.28 versus 0.30 with group1+2 and 0.29 with group 1+2+3). Focusing only on models using group1 features (the most robust), the best result was achieved using VieCuri Medisch Centrum as validation center (global C-index 0.62, global iBS 0.26), which meant training the model on the larger sample size possible (435 samples for training cohort). LOCO results according to different feature groups are summarized in *Table 15*.

Center-out	**Ospedale San Raffaele**	**Maastricht UMC**	**Maxima MC**	**Viecuri Medisch Centrum**
group1 features				
C-index	0.61	0.52	0.56	0.62
iBS	0.24	0.31	0.32	0.26
group1+2 features				
C-index	0.55	0.52	0.62	0.48
iBS	0.25	0.33	0.32	0.29
group1+2+3 features				
C-index	0.45	0.54	0.59	0.61
iBS	0.24	0.34	0.33	0.24

Table 15. Transformer-based models' performances according to different feature groups and different LOCO experiments

Moreover, transformer-based models allow better visualization of survival over time, resulting in higher C-index in the first quartile of follow-up time, except when Maastricht UMC served as the validation cohort. On the other hand, the iBS tended to be lower in the first quartile across all LOCO experiments. Trends are summarized in *Figure 14a-b*.

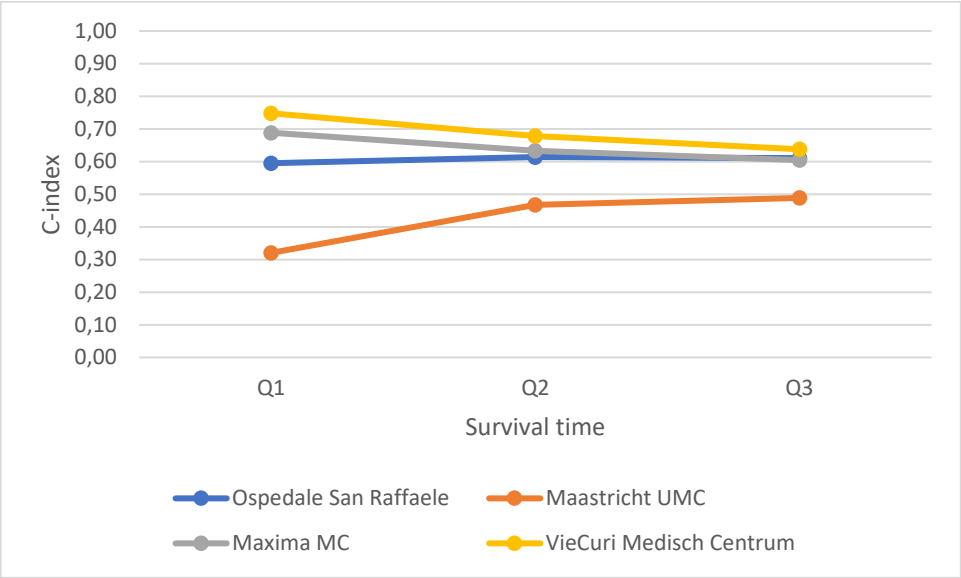

Figure 14a. C-index variations over time (see Legenda for Center-out)

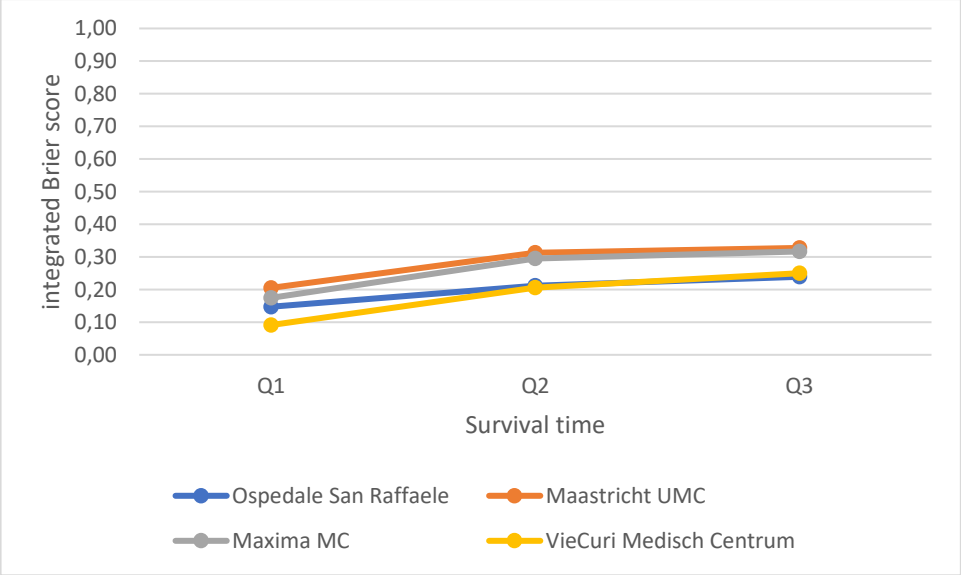

Figure 14b. Integrated Brier score variations over time (see Legenda for Center-out)

Chapter 4: Discussion

Machine learning models for RWD: are data sources ready yet?

In our study, ML models emerged as a reliable tool for outcome prediction in patients with metastatic NSCLC. We demonstrated that "black-box" ML models can achieve consistent performances (AUC between 0.74 and 0.82) with minimal risk of overfitting in a small real-world dataset, and that they can be seamlessly paired with explainability tools such as SHAP. Beyond accuracy, we assessed the feasibility of an integrated interface between the models and hospital sources, improving data quality and reducing manual workload. Other real-world studies published by Italian groups (such as the AI-ON Lab of Fondazione IRCCS Istituto Nazionale Tumori) have never provided details about source integration, and the output models have probably been developed relying on manual data extraction, with minimal automation for imaging processing[143,144]. The S-RACE platform, on the other hand, has the higher aim of directly integrating into clinical workflows, reducing the need for manual check, therefore speeding up data collection and ensuring data quality. The AI-HOPE study has served as one of the pilot studies inside Ospedale San Raffaele for this platform integration and development.

Laboratory data are the easiest source to map, thanks to the default settings of laboratory software and to the pre-defined structure of the data (provided normal values range, international measurement units, etc)[96]. In our feasibility study, it had a clear repercussion on feature importance, with laboratory values emerging in the top ranked features across all ML models. Despite the well-known relevance of neutrophil count and NLR for immunotherapy response[72], the strong presence of laboratory values in the models is widely dependent on their robustness thanks to automatic retrieval and low proportion of missing data.

On the other hand, clinical features did not emerge as relevant for outcome prediction, despite they have been considered the most consistent prognostic factors for decades[80]. In this case, model performance improved when a second layer of manual data curation was applied, in particular recollecting detailed data on concomitant medications. This procedure is a clear proof-of-concept of the "garbage in, garbage out" principle, established in computer science since 1957 and applicable to medical practice, too[145]. Only multiple human manual check could ensure the necessary data quality to reach

results consistent with literature, at least for variables that can be found in free text format (as EHR in Ospedale San Raffaele have very few structured fields). Some research institutions are now partnering with high-tech companies to introduce large language models (LLMs) or NLP for clinical data collection from reports[146], but this use is still hindered by economical and privacy concerns, above all in most of the non-academic hospitals. Moreover, in the near future, LLMs could be enriched with specialized toolsets (i.e., image-analysis modules, guideline retrieval systems, web/document search) to process multimodal data and work autonomously to plan and execute multiple tasks[147]. Despite brilliant performances in simulated oncologic patient cases and clear impact on manual workload reduction, real-world deployment is still limited and these studies, conducted in a research environment, may not reflect the variability and complexity of routine care.

Another approach to overcome this limitation is data standardization, which aims at linking single data fields to standardized categories in larger ontologies. Many hospitals, including Ospedale San Raffaele and Maastricht University Medical Center, are collaborating in European networks for data standardization, such as the European Economic Interest Grouping Digicore[148]. The aim of this consortium is sharing a common data model (CDM) that defines how clinical information should be structured and stored according to the OMOP standards (Observational Medical Outcomes Partnership), and links the CDM to an international ontology for the medical field (SNOMED, LOINC or RxNorm[149]). This approach is optimal for larger institutions with structured EHRs and multiple international collaborations that highlight the need for ready-to-exchange data. On the other hand, in case of non-structured EHR, it does not overcome the need of manual data curation before the CDM. Nevertheless, our group was part of the core-group for OMOP standardization in the Digicore consortium, collaborating with other European centers in order to create the first standardized pan-cancer hospital network. Within this study, and alongside the S-RACE platform development, we could highlight the areas of improvement for in-hospital data mining as well as working on algorithms that could work on few structured but solid variables. Accordingly, we presented some preliminary results on OMOP standardization of real-world data of patients with metastatic NSCLC, describing location of metastases and survivals via a federated learning approach[150]. On this same topic, a line-of-treatment algorithm was designed, to extract subsequent

oncological therapies with a set of rules and without any disease-related information (i.e. progression date, imaging scans results, etc)[151]. All these efforts have led to a fruitful international collaboration and substantial progress in internal data structuring, but much remains to be done in the context of automatic data extraction.

The upside of this feasibility study is the seamless integration of "black-box" and "white-box" models. In the introductory section, we already discussed the strengths and pitfalls of powerful but non-interpretable models (also called "black-box") and the need of explainability tools to improve users' trust ("white-box"). In our study, SHAP was applied on the Voting Ensemble models, preserving both good explainability and consistent performances, as previously reported from other groups[143]. We also added another explainable tool that is easy to interpret for medical reasoning, which is the Decision Tree. The structure of the tree is intuitive (branches and ramifications) and closely mirrors multi-step medical reasoning, where attention is first given to the most relevant variables and then to progressively less important ones. Therefore, it could rapidly evolve into a quick risk-scoring system for everyday clinical practice use as already proven in literature[152].

Altogether, our study is the first study proving this paired approach in a real-world population of patients with metastatic NSCLC treated with first-line immunotherapy and published in an international peer-reviewed journal at the beginning of 2025[153]. Our study succeeded in matching a solid workflow together with a very tailored prediction, because a previous study on this same topic considered patients treated with immunotherapy in first- and second-line line setting (N=480)[143], thus reducing the generalizability of results in the current first-line scenario. Recently, another similar study has been published on ML models for prediction in 1050 patients treated with first-line pembrolizumab for metastatic NSCLC[154]. Despite using an innovative architecture (NAIM, a transformer-based artificial intelligence model designed to handle missing data without imputation), this time-to-event model achieved mediocre performances (C-index for OS 0.623±0.21), and the authors acknowledged a certain degree of overfitting and limited generalizability due to selection bias, which are inherent to a retrospective real-world data collection. Nevertheless, ECOG PS, age, and metastatic burden emerged as the most relevant risk factors, consistently with our results and with literature. However, the Pembro-real 5Y registry is a comprehensive real-world, global, dataset derived from 61 institutions[43], in

which data quality was ensured by research personnel and structured eCRF, making it fully dependent on manual human check. Our study, on the other hand, aimed to automate data processing and develop algorithms capable of interfacing directly with multiple sources to build a comprehensive eCRF for model development, representing a pioneering solution in the healthcare domain.

In conclusion, our approach showed promising clinical application in a real-world scenario. Results were consistent with scientific evidence, despite the small sample size and the intrinsic risk of overfitting of ML models. Data extraction was solid and validated by a manual check, which might be replaced by artificial intelligence tools in the near future, reducing human workload. The pairing of "white-box" and "black-box" models merges the power and the transparency of the two methodologies, paving the way for an integrated technology, ready-to-use aside EHRs. Together with a non-deferrable update of Italian hospital information systems, these integrated platforms will truly be able, within a few years, to make a difference in routine clinical practice for prognosis estimation and treatment allocation.

Imaging derived biomarkers and related pre-analytical issues

Despite combining different approaches for radiomics integration into predictive modelling, our experiments did not achieve a satisfactory balance between workload and significance. Regardless of the feature selection method, ML models incorporating radiomic features did not outperform models excluding radiomics (see Section 3.1), yet required substantially more manual work and supervision due to the absence of automated workflows. Nevertheless, the limited dataset posed challenges for the generalizability of both approaches and results, as well as for the interpretation of outputs from unsupervised models.

The inclusion of radiomic features in outcome prediction has been a challenge since their discovery in early 2000s[105]. There is often a lack of understanding of radiomic biological meaning[155], and the overall workflow for feature extraction and interpretation is not reproducible due to multistep manual processing[134]. We tried to dig deeper into these issues in order to optimize our approach and make it feasible for further inclusion in multisource algorithms for outcome prediction in patients with metastatic NSCLC.

Firstly, feature extraction was conducted as per literature standards, namely proceeding with manual contouring by expert radiologist(s) and subsequently extracting radiomic features from each ROI via PyRadiomics. Nevertheless, even in a single-center cohort, not all the CT-scans were performed with the same scanner (changing over time, or patients tested outside the hospital), increasing the variability in image acquisition procedures[156]. Moreover, manual contouring is considered the benchmark also for automatic or semi-automatic segmentation, but it is resource-intensive and biased by inter-operator inconsistencies[156]. In our pipeline, post-image acquisition pre-processing tools have not been implemented, resulting in problems with feature scaling when merging two different datasets (OSR and BS, *data not shown*). On the other hand, the datasets of Spedali Civili (BS1 and BS2) were used to assess and improve reproducibility, thanks to double manual contouring. Nevertheless, we decided to focus our experiments on feature selection, which is a crucial step for further implementation of radiomics into more complex models, acknowledging the reduced sample size and the lack of harmonization algorithms as the main areas for future improvement.

In different experiences, feature selection has been approached differently, in order to achieve a significant reduction in dimensionality (many radiomic features for few patients may increase the risk of overfitting). The primary objective of feature selection is to create a reduced subset of features that accurately mirror the whole set of features. In literature, most of the studies used embedded approaches (such as LASSO, Least Absolute Shrinkage and Selection Operator), which assess feature relevance based on the classifier's intrinsic ability to identify the most informative features, or filters (i.e. Spearman correlation or intra-class correlation), which rely on data properties that are independent from the classifier[157]. In our study, we compared a traditional filtering technique (ICC-based) with agnostic ML-approach, achieving similar results in terms of performance and feature ranking. However, hybrid approaches are now emerging as the best-performing approaches[157], although not widely used in literature. In a recent paper by the AI-ON Lab of Fondazione IRCCS Istituto Nazionale Tumori on radiomics in metastatic NSCLC[158], only filtering techniques were applied (Spearman correlation, maximum relevance minimum redundancy) and final models encompassing radiomics and clinical data reached interesting results in terms of performances (AUC 0.74 for clinical benefit rate, 0.79 for OS status at 2 years). Still, limited information about pre-

processing harmonization is available, and a consistent sample size of 375 patients could be reached only by including different settings (any-line immunotherapy), thus limiting the generalizability of the results in the current first-line scenario. In our study, given that we aimed to maintain a homogeneous population by including only first-line patients, and considering that image transfer is currently slow and has not allowed in-house processing of CT-scans from the external center (thus preventing the implementation of harmonization techniques), we limited our work to a few proof-of-concept experiments, which were nonetheless time- and labor-intensive.

Another paper published in 2022 has included radiomic features of 337 chest lesions (lung, pleura, lymphnodes) from 187 patients treated with any-line immunotherapy for advanced NSCLC[159]. ML models including some (or all) the contoured lesions reached disappointing results (AUC 0.61-0.65 for response discrimination), and highlight some constraints that are shared in the radiomic field: lack of automatic segmentation, focus on thoracic region only, multi-step non standardized feature selection[159]. From a clinical perspective, satisfactory performances were achieved only layering complementary feature sets (PD-L1 score, genomic and radiomic features)[159]. Nonetheless, this study was single-center and lacks external validation.

In our opinion, the balance between the robustness of radiomic features and their informative content is clearly unfavorable. As we discussed so far, there are many issues in all the phases of the radiomic workflow that insurmountably limit the transition from research to a clinical standard. On the other hand, deep learning methods skip many pre-processing steps (above all, segmentation) and learn without explicit feature definitions[160]. They have been shown to outperform conventional radiomics models, at the cost of requiring larger sample sizes and offering reduced interpretability[160], although radiomic features themselves are not highly interpretable (beyond metrics such as volume and intensity, for instance). Moreover, identifying a single ROI may introduce bias in metastatic cancers due to tumor heterogeneity and the presence of multiple metastatic sites, whereas delineating multiple ROIs would be highly time-consuming and not sustainable[161]. New methods are emerging that could build an individual "fingerprint" starting from multiple lesions and exploiting dimensionality reduction methods such as PCA (Principal Component Analysis), providing new information about intra-patient lesion similarity and its impact on survival outcomes[162]. Therefore, we believe that

standard radiomic approaches that have proven solid in small lung nodules cannot be applied blindly to the metastatic scenario, and that new tools will speed up imaging analysis and its inclusion in ML models for outcome prediction in the upcoming future.

In addition, we tried to explore the role of unsupervised models in radiomics, believing that unlabeled data could help identifying new patterns also in small datasets. Clustering methods tend to group patients according to their similarity, and to maximize the distance of different clusters[163]. In our study, given the retrospective nature, we could not deepen our understanding of tumor biology, and we could compare only basic clinical variables among the three resulting clusters. Overall, Cluster A tended to show a slightly poorer prognosis (higher rates of PD, shorter OS) compared to Clusters B and C, suggesting an underlying worse biology, potentially reflected by radiomic features. Cluster B, for instance, had higher median PD-L1 expression levels (60%), which is usually a good prognostic factor[58], but the best overall survival was reached by Cluster C patients (median of 19 months), which had a median PD-L1 expression of 20%. Cluster C, on the other hand, had more patients with good performance status (52% of ECOG 0) compared to Cluster A and B (32% and 26%, respectively), which could at least partially explain the good risk profile. However, the clustering algorithm grouped patients based solely on radiomic features, leaving room for discussion about the individual contribution of biological and clinical factors. In particular, it is interesting to explore the extent to which underlying biology may influence the clinical prognostic factors we already know, opening a path toward anticipating risk profiles in earlier disease-stages by examining imaging-derived biomarkers before they evolve into a specific clinical presentation.

As previously reported in a different setting (EGFR-mutant NSCLC), clustering-based radiomic phenotyping appears promising in its ability to discriminate survival outcomes in patients treated with targeted therapy[164]. In the study by Yousefi et al., radiomic features alone identified two patient groups with a difference in progression-free survival exceeding 6 months[164]. Beyond this specific result, the authors support the concept that radiomics may reflect tumor heterogeneity, with patients with a favorable prognosis exhibiting more homogeneous and well-defined tumors, compared with patients whose neoplasms are more irregular and poorly defined[164]. Similarly, another study in patients with stage III–IV NSCLC treated with any-line immunotherapy confirmed the feasibility of exploiting unsupervised approaches such as clustering to predict survival outcomes[165].

In this study, the radiomic cluster was an independent predictor of OS and appeared to correlate with T/NK lymphocyte infiltration[165]. The authors therefore suggest that radiomics alone may play a role not only in capturing tumor heterogeneity, but also in revealing the biological underpinnings of the tumor microenvironment[165]. Our pilot experiment perfectly aligns with these results, even if a prospective validation is needed to combine them with wide-genome sequencing or other deeper "omics" analyses.

Another major limitation of our radiomic sub-study is that it is limited to CT-scans, although analyses on 18F-FDG-PET are planned within the study. From a practical perspective, we chose to primarily focus on CT-scans because they were more widely available than 18F-FDG-PET scans and because of internal expertise on the topic. Nevertheless, we aim at including also PET-derived biomarkers in our models, extracted both from primary tumors, from metastatic sites and from peritumoral areas, which are known to be useful in terms of predictive performance in NSCLC[166]. However, the processing of PET-scans is hindered by many standardization problems, which include inter-scanners comparisons, but also voxel sizes, intensity resolution, SUV normalization etc[167]. In my experience at Maastricht University Medical Center, I had the opportunity to delve into PET-analysis, even if limited to basic metabolic features such as metabolic tumor volume (MTV) and total lesion glycolysis (TLG). Both these variables, despite being well described in literature, are hindered by pre-analytical issues that vary from acquisition protocols (studies can be included in the analysis only if they followed international guidelines) to patient-related information (blood glucose, fasting, timing, etc). As we reported in our study, both MTV and TLG appear to have a prognostic value for patients with oligo-metastatic NSCLC[168], but their real impact might be diluted by the difficulties in centralizing and processing images.

In conclusion, imaging-derived biomarkers are a promising source of information in metastatic NSCLC, above all if considered with a more holistic approach, beyond the primary tumor. In the future of the AI-HOPE study there is further improvement in this area, moving towards deep learning models and exploiting long-term prediction, aiming to identify the (positive) flip side of tumor biology. International collaborations on this topic are ongoing and AI-HOPE was selected as one of only four projects (out of more than twenty) to be further developed within the first European "Endeavour Lung Cancer

Programme", based on its innovative design and its recognized value for the oncological community.

Time-to-event machine learning models: the neglected sibling in oncology

In the recent years, ML models have boosted the generation of predictive models in oncology. Classification models, for instance, predict discrete outcomes (i.e., yes/no, responder/non-responder), while time-to-event survival models predict *when* an event occurs, accounting for censoring (patients who have not experienced the event yet)[169]. Classification models are usually easier to build and validate, because they do not need to handle censoring, outcome data are simple (binary), and the modelling frameworks are well-standardized[86]. On the other hand, despite performing very well in pure classification tasks (i.e. for biomarker evaluation in histological slides[170]), they are not so consistent in survival prediction. Many published studies in metastatic NSCLC exploit classification models for this task[121,143,144], predicting OS at different timepoints, but they risk to prove inconsistent for real-world applications because they miss the continuous nature of survival outcomes, overlooking patients whose events occur close to or far from the pre-specified timepoint.

Given that our final aim is to bridge the gap between a purely academic result and a possible clinical application, we chose to develop time-to-event survival models in a real-world population of patients with metastatic NSCLC and treated with first-line ICI-based regimens. In order to deploy the best performing model, we started from a comparison of two different architectures to identify our benchmark for future evaluations: the XGBoost (ensemble of gradient-boosted decision trees adapted for censored data[137]) and the transformer-based (deep neural network architecture with self-attention layers[171]). In recently published studies, transformer-based models outperformed standard methods (i.e. Cox proportional hazards model, random survival forest) in performance metrics for survival predictions, despite higher computational costs and lower interpretability[172]. Moreover, new architectures have been developed for this task, that build – on top of classical transformer-based models – innovative complementary approaches, such as NAIM (Not Another Imputation Method), specifically designed to handle missing data without imputation[154]. However, in our study, the highest results in terms of C-index were achieved by XGBSE StackedWeibull (C-index 0.83 with feature group1+2 and OSR as

Center-out) and the RandomForest SurvivalAnalysis (C-index 0.79 with feature group1+2 and OSR as Center-out). As well described in literature[173], for tabular data only and with small to medium datasets, RandomForest SurvivalAnalysis and the XGBSE library perform better than neural networks, and they are robust despite the background noise. In our study, best performing models were the simplest ones and they did not improve significantly when larger number of features were considered as input data. On the other hand, the overall satisfactory results achieved from the transformer-based models allow us to further work in this direction, exploiting their strength in handling sequential, multi-modal or high-dimensionality data sources[174].

Delving into the details of the models, transformer-based experiments clearly show a fluctuation of performances over time. The C-index was overall higher at a shorter follow-up (first quartile of survival time), a part when Maastricht UMC was the Center-out, and it slightly decreased at median and third quartile timepoints. Nevertheless, we included also the integrated Brier score value at those same timepoints, observing a reverse phenomenon (lower at a shorter follow-up, and then increasing over time). The interpretation of these combined parameters supports that our transformer-based model performed best at the first quartile of survival time (first months of treatment), while losing some power in the later phases. However, despite C-index values were mediocre compared to the existing literature[154,172], results proved consistent thanks to the integration of the integrated Brier score, which complements the C-index by jointly assessing both discrimination and calibration over time. While the C-index measures the model's ability to correctly rank patients by risk (one after one another, but without information on the estimated time of the event), the Integrated Brier Score quantifies how punctual the prediction is over the entire follow-up period[175]. We strongly believe that models integrating both these metric parameters have higher interpretability and greater potential for transparent refinement according to the user's specific application needs[176,177].

Nevertheless, models extracted from the XGBSE library and the RandomForest SurvivalAnalysis proved solid among all the experiments. The double step validation (Nested Cross Validation and proper external validation with LOCO) guided the choice of the best performing model according to Center-out and group of input features, improving its performance on the external validation dataset after re-training. C-index

values over 0.83 clearly outperform available literature evidence for XGBSE survival models[178] and the Cox regression reference model for OS in our dataset (C-index 0.65). Accordingly, the RandomForest SurvivalAnalysis performed very well, achieving a C-index of 0.79 in different experiments, which is in line with literature[179]. Both these architectures, with sequential or parallel boosting, have good interpretability and low risk of overfitting in smaller datasets, and are therefore considered a good benchmark for OS prediction with tabular data in oncology[173].

In the upcoming future, we are planning to explore multi-view models as modelling framework to improve performances of the aforementioned architectures. So far, only tabular data were available, and imaging processing was slower than expected, resulting in a single-view data strategy. In literature, multi-view models have already been published in the lung cancer field[180], achieving consistent results for models including imaging and RNA-sequencing (C-index 0.76). Given the observational design of the AI-HOPE study, only data generated from clinical practice are used as input data, but solid retrospective models can lay the basis for future prospective studies expanding the data sources.

Nevertheless, the ultimate goal is the seamless integration of these prognostic models into routine clinical practice: clinicians collecting clinical and laboratory data during medical visits could obtain a real-time prediction of the prognosis of the patient in front of them. The impact on the patient's care pathway would therefore span multiple domains, including more accurate physician–patient communication, better support for caregivers, implementation of supportive care strategies, and enrollment in clinical trials with novel agents for patients who are unlikely to benefit from standard therapies. Models of this kind represent the final objective of a project such as AIHOPE.

Conclusions

Machine learning models are reshaping outcome prediction in oncology, but real-world data struggle to keep up with rapid methodological advancements. The European and, above all, the Italian scenario is challenged by many initiatives aiming to address the key issues of data standardization, data quality, and automatic data collection, yet much remains to be done. We have demonstrated that the future lies in integrating AI-models within hospital software, which can concretely optimize processes beyond research institutions.

Imaging-scans are gold mines for artificial intelligence, because of their intrinsic link with tumor biology and their wide availability in clinical practice. However, reproducibility of analytical approaches remains a major limitation in studies exploring image-related biomarkers, including (but not limiting to) radiomics. Future developments in this field – overcoming manual work – will enhance the potential of imaging as one of the integrated sources in multi-modal models, despite losing interpretability and posing users' trust at risk.

To achieve faster integration into clinical workflows, models evaluating survivals should shift from classification to time-to-event modelling. Many architectures can be exploited in this setting, favoring those that handle missing data without imputation and that balance good performances with explainability. Robust time-to-event models are expected to be the backbone of future developments in this field, enabling exploration of new data sources and integration strategies.

Bibliography

1. Bray, F. *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 74, 229–263 (2024).
2. https://gco.iarc.who.int/media/globocan/factsheets/cancers/15-trachea-bronchus-and-lung-fact-sheet.pdf.
3. Barta, J. A., Powell, C. A. & Wisnivesky, J. P. Global Epidemiology of Lung Cancer. *Ann. Glob. Health* 85, (2019).
4. Santucci, C. *et al.* European cancer mortality predictions for the year 2024 with focus on colorectal cancer. *Annals of Oncology* 35, 308–316 (2024).
5. Hendriks, L. E. L. *et al.* Non-small-cell lung cancer. *Nat. Rev. Dis. Primers* 10, 71 (2024).
6. SEER. Cancer Stat Facts: Lung and Bronchus Cancer. Available at: https://seer.cancer.gov/statfacts/html/lungb.html.
7. Oxnard, G. R., Nguyen, K.-S. H. & Costa, D. B. Germline Mutations in Driver Oncogenes and Inherited Lung Cancer Risk Independent of Smoking History. *JNCI Journal of the National Cancer Institute* 106, djt361–djt361 (2014).
8. Adams, S. J. *et al.* Lung cancer screening. *The Lancet* 401, 390–408 (2023).
9. de Koning, H. J. *et al.* Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *New England Journal of Medicine* 382, 503–513 (2020).
10. Hill, W. *et al.* Lung adenocarcinoma promotion by air pollutants. *Nature* 616, 159–167 (2023).
11. Siegel, R. L., Miller, K. D., Wagle, N. S. & Jemal, A. Cancer statistics, 2023. *CA Cancer J. Clin.* 73, 17–48 (2023).
12. Nicholson, A. G. *et al.* The 2021 WHO Classification of Lung Tumors: Impact of Advances Since 2015. *Journal of Thoracic Oncology* 17, 362–387 (2022).
13. Devarakonda, S. *et al.* Genomic Profiling of Lung Adenocarcinoma in Never-Smokers. *Journal of Clinical Oncology* 39, 3747–3758 (2021).
14. Remon, J., Soria, J.-C. & Peters, S. Early and locally advanced non-small-cell lung cancer: an update of the ESMO Clinical Practice Guidelines focusing on diagnosis, staging, systemic and local therapy. *Annals of Oncology* 32, 1637–1642 (2021).
15. Asamura H, Farjah F, Gill R, et al. AJCC Cancer Staging System: Lung (Version 9). In: AJCC Cancer Staging Manual, 9th ed. Chicago, IL: Springer; 2024. in.

16. Ganti, A. K., Klein, A. B., Cotarla, I., Seal, B. & Chou, E. Update of Incidence, Prevalence, Survival, and Initial Treatment in Patients With Non–Small Cell Lung Cancer in the US. *JAMA Oncol.* 7, 1824 (2021).

17. Barlesi, F. *et al.* Routine molecular profiling of patients with advanced non-small-cell lung cancer: results of a 1-year nationwide programme of the French Cooperative Thoracic Intergroup (IFCT). *The Lancet* 387, 1415–1426 (2016).

18. Gálffy, G. *et al.* Targeted therapeutic options in early and metastatic NSCLC-overview. *Pathology and Oncology Research* 30, (2024).

19. Marei, H. E., Hasan, A., Pozzoli, G. & Cenciarelli, C. Cancer immunotherapy with immune checkpoint inhibitors (ICIs): potential, mechanisms of resistance, and strategies for reinvigorating T cell responsiveness when resistance is acquired. *Cancer Cell Int.* 23, 64 (2023).

20. Borghaei, H. *et al.* Nivolumab versus Docetaxel in Advanced Nonsquamous Non–Small-Cell Lung Cancer. *New England Journal of Medicine* 373, 1627–1639 (2015).

21. Brahmer, J. *et al.* Nivolumab versus Docetaxel in Advanced Squamous-Cell Non–Small-Cell Lung Cancer. *New England Journal of Medicine* 373, 123–135 (2015).

22. Herbst, R. S. *et al.* Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. *The Lancet* 387, 1540–1550 (2016).

23. Rittmeyer, A. *et al.* Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *The Lancet* 389, 255–265 (2017).

24. Garon, E. B. *et al.* Pembrolizumab for the Treatment of Non–Small-Cell Lung Cancer. *New England Journal of Medicine* 372, 2018–2028 (2015).

25. Mino-Kenudson, M. Immunohistochemistry for predictive biomarkers in non-small cell lung cancer. *Transl. Lung Cancer Res.* 6, 570–587 (2017).

26. Carbone, D. P. *et al.* First-Line Nivolumab in Stage IV or Recurrent Non–Small-Cell Lung Cancer. *New England Journal of Medicine* 376, 2415–2426 (2017).

27. Reck, M. *et al.* Pembrolizumab versus Chemotherapy for PD-L1–Positive Non–Small-Cell Lung Cancer. *New England Journal of Medicine* 375, 1823–1833 (2016).

28. Reck, M. *et al.* Five-Year Outcomes With Pembrolizumab Versus Chemotherapy for Metastatic Non–Small-Cell Lung Cancer With PD-L1 Tumor Proportion Score ≥ 50%. *Journal of Clinical Oncology* 39, 2339–2349 (2021).

29. Herbst, R. S. *et al.* Atezolizumab for First-Line Treatment of PD-L1–Selected Patients with NSCLC. *New England Journal of Medicine* 383, 1328–1339 (2020).

30. Jassem, J. *et al.* Updated Overall Survival Analysis From IMpower110: Atezolizumab Versus Platinum-Based Chemotherapy in Treatment-Naive Programmed Death-Ligand 1–Selected NSCLC. *Journal of Thoracic Oncology* 16, 1872–1882 (2021).

31. Sezer, A. *et al.* Cemiplimab monotherapy for first-line treatment of advanced non-small-cell lung cancer with PD-L1 of at least 50%: a multicentre, open-label, global, phase 3, randomised, controlled trial. *The Lancet* 397, 592–604 (2021).

32. Kilickap, S. *et al.* Cemiplimab Monotherapy for First-Line Treatment of Patients with Advanced NSCLC With PD-L1 Expression of 50% or Higher: Five-Year Outcomes of EMPOWER-Lung 1. *Journal of Thoracic Oncology* 20, 941–954 (2025).

33. Hendriks, L. E. *et al.* Non-oncogene-addicted metastatic non-small-cell lung cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Annals of Oncology* 34, 358–376 (2023).

34. Yang, S. *et al.* Autoimmune Effects of Lung Cancer Immunotherapy Revealed by Data-Driven Analysis on a Nationwide Cohort. *Clin. Pharmacol. Ther.* 107, 388–396 (2020).

35. Gandhi, L. *et al.* Pembrolizumab plus Chemotherapy in Metastatic Non–Small-Cell Lung Cancer. *New England Journal of Medicine* 378, 2078–2092 (2018).

36. Paz-Ares, L. *et al.* Pembrolizumab plus Chemotherapy for Squamous Non–Small-Cell Lung Cancer. *New England Journal of Medicine* 379, 2040–2051 (2018).

37. Garassino, M. C. *et al.* Pembrolizumab Plus Pemetrexed and Platinum in Nonsquamous Non–Small-Cell Lung Cancer: 5-Year Outcomes From the Phase 3 KEYNOTE-189 Study. *Journal of Clinical Oncology* 41, 1992–1998 (2023).

38. Novello, S. *et al.* Pembrolizumab Plus Chemotherapy in Squamous Non–Small-Cell Lung Cancer: 5-Year Update of the Phase III KEYNOTE-407 Study. *Journal of Clinical Oncology* 41, 1999–2006 (2023).

39. Paz-Ares, L. *et al.* First-line nivolumab plus ipilimumab combined with two cycles of chemotherapy in patients with non-small-cell lung cancer (CheckMate 9LA): an international, randomised, open-label, phase 3 trial. *Lancet Oncol.* 22, 198–211 (2021).

40. Gogishvili, M. *et al.* Cemiplimab plus chemotherapy versus chemotherapy alone in non-small cell lung cancer: a randomized, controlled, double-blind phase 3 trial. *Nat. Med.* 28, 2374–2380 (2022).

41. Zwanenburg, L. C. *et al.* Living in the twilight zone: a qualitative study on the experiences of patients with advanced cancer obtaining long-term response to immunotherapy or targeted therapy. *Journal of Cancer Survivorship* 18, 750–760 (2024).

42. Pons-Tostivint, E. *et al.* Comparative Analysis of Durable Responses on Immune Checkpoint Inhibitors Versus Other Systemic Therapies: A Pooled Analysis of Phase III Trials. *JCO Precis. Oncol.* 1–10 (2019) doi:10.1200/PO.18.00114.

43. Cortellini, A. *et al.* Determinants of 5-year survival in patients with advanced NSCLC with PD-L1≥50% treated with first-line pembrolizumab outside of clinical trials: results from the Pembro-real 5Y global registry. *J. Immunother. Cancer* 13, e010674 (2025).

44. Eisenhauer, E. A. *et al.* New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur. J. Cancer* 45, 228–247 (2009).

45. Seymour, L. *et al.* iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. *Lancet Oncol.* 18, e143–e152 (2017).

46. Garralda, E., Laurie, S. A., Seymour, L. & de Vries, E. G. E. Towards evidence-based response criteria for cancer immunotherapy. *Nat. Commun.* 14, 3001 (2023).

47. Borcoman, E. *et al.* Novel patterns of response under immunotherapy. *Annals of Oncology* 30, 385–396 (2019).

48. Di Giacomo, A. M. *et al.* Therapeutic efficacy of ipilimumab, an anti-CTLA-4 monoclonal antibody, in patients with metastatic melanoma unresponsive to prior systemic treatments: clinical and immunological evidence from three patient cases. *Cancer Immunology, Immunotherapy* 58, 1297–1306 (2009).

49. Sharma, P., Hu-Lieskovan, S., Wargo, J. A. & Ribas, A. Primary, Adaptive, and Acquired Resistance to Cancer Immunotherapy. *Cell* 168, 707–723 (2017).

50. Champiat, S. *et al.* Hyperprogressive Disease Is a New Pattern of Progression in Cancer Patients Treated by Anti-PD-1/PD-L1. *Clinical Cancer Research* 23, 1920–1928 (2017).

51. Viscardi, G. *et al.* Comparative assessment of early mortality risk upon immune checkpoint inhibitors alone or in combination with other agents across solid malignancies: a systematic review and meta-analysis. *Eur. J. Cancer* 177, 175–185 (2022).

52. Tawbi, H. A. *et al.* Society for Immunotherapy of Cancer (SITC) checkpoint inhibitor resistance definitions: efforts to harmonize terminology and accelerate immuno-oncology drug development. *J. Immunother. Cancer* 11, e007309 (2023).

53. Schoenfeld, A. J. *et al.* Clinical definition of acquired resistance to immunotherapy in patients with metastatic non-small-cell lung cancer. *Annals of Oncology* 32, 1597–1607 (2021).

54. Oresti, S. *et al.* Impact of platinum-based chemotherapy and CTLA-4 inhibition on acquired resistance to first-line anti-PD-1/PD-L1 agents in non-small cell lung cancer: a systematic review and reconstructed individual patient data analysis. *EClinicalMedicine* 88, 103482 (2025).

55. Clark, G. M. Prognostic factors versus predictive factors: Examples from a clinical trial of erlotinib. *Mol. Oncol.* 1, 406–412 (2008).

56. Harvey, R. D. *et al.* Impact of broadening clinical trial eligibility criteria for advanced non-small cell lung cancer patients: Real-world analysis. *Journal of Clinical Oncology* 37, LBA108–LBA108 (2019).

57. Brueckl, W. M., Ficker, J. H. & Zeitler, G. Clinically relevant prognostic and predictive markers for immune-checkpoint-inhibitor (ICI) therapy in non-small cell lung cancer (NSCLC). *BMC Cancer* 20, 1185 (2020).

58. Aguilar, E. J. *et al.* Outcomes to first-line pembrolizumab in patients with non-small-cell lung cancer and very high PD-L1 expression. *Annals of Oncology* 30, 1653–1659 (2019).

59. Dall'Olio, F. G. *et al.* ECOG performance status ≥2 as a prognostic factor in patients with advanced non small cell lung cancer treated with immune checkpoint inhibitors—A systematic review and meta-analysis of real world data. *Lung Cancer* 145, 95–104 (2020).

60. Facchinetti, F. *et al.* First-line pembrolizumab in advanced non–small cell lung cancer patients with poor performance status. *Eur. J. Cancer* 130, 155–167 (2020).

61. Magri, V. *et al.* Correlation of body composition by computerized tomography and metabolic parameters with survival of nivolumab-treated lung cancer patients. *Cancer Manag. Res.* Volume 11, 8201–8207 (2019).

62. Conforti, F. *et al.* Cancer immunotherapy efficacy and patients' sex: a systematic review and meta-analysis. *Lancet Oncol.* 19, 737–746 (2018).

63. Cortellini, A. *et al.* Another side of the association between body mass index (BMI) and clinical outcomes of cancer patients receiving programmed cell death protein-1 (PD-1)/ Programmed cell death-ligand 1 (PD-L1) checkpoint inhibitors: A multicentre analysis of immune-related adverse events. *Eur. J. Cancer* 128, 17–26 (2020).

64. Rounis, K. *et al.* Cancer cachexia syndrome and clinical outcome in patients with metastatic non-small cell lung cancer treated with PD-1/PD-L1 inhibitors: results from a prospective, observational study. *Transl. Lung Cancer Res.* 10, 3538–3549 (2021).

65. Kawachi, H. *et al.* Association between metastatic sites and first-line pembrolizumab treatment outcome for advanced non–small cell lung cancer with high PD-L1 expression: a retrospective multicenter cohort study. *Invest. New Drugs* 38, 211–218 (2020).

66. Sridhar, S. *et al.* Prognostic Significance of Liver Metastasis in Durvalumab-Treated Lung Cancer Patients. *Clin. Lung Cancer* 20, e601–e608 (2019).

67. Vilariño, N., Bruna, J., Bosch-Barrera, J., Valiente, M. & Nadal, E. Immunotherapy in NSCLC patients with brain metastases. Understanding brain tumor microenvironment and dissecting outcomes from immune checkpoint blockade in the clinic. *Cancer Treat. Rev.* 89, 102067 (2020).

68. Dong, H. *et al.* Prognostic significance of bone metastasis and clinical value of bone radiotherapy in metastatic non-small cell lung cancer receiving PD-1/PD-L1 inhibitors: results from a multicenter, prospective, observational study. *Transl. Lung Cancer Res.* 13, 2603–2616 (2024).

69. Scott, S. C. & Pennell, N. A. Early Use of Systemic Corticosteroids in Patients with Advanced NSCLC Treated with Nivolumab. *Journal of Thoracic Oncology* 13, 1771–1775 (2018).

70. Rousseau, A. *et al.* Concomitant Comedications and Survival With First-Line Pembrolizumab in Advanced Non–Small-Cell Lung Cancer. *JAMA Netw. Open* 8, e2529225 (2025).

71. Routy, B. *et al.* Gut microbiome influences efficacy of PD-1–based immunotherapy against epithelial tumors. *Science (1979).* 359, 91–97 (2018).

72. Cao, D., Xu, H., Xu, X., Guo, T. & Ge, W. A reliable and feasible way to predict the benefits of Nivolumab in patients with non-small cell lung cancer: a pooled analysis of 14 retrospective studies. *Oncoimmunology* 7, (2018).

73. Zhang, Z. *et al.* Pretreatment lactate dehydrogenase may predict outcome of advanced non small-cell lung cancer patients treated with immune checkpoint inhibitors: A meta-analysis. *Cancer Med.* 8, 1467–1473 (2019).

74. Riedl, J. M. *et al.* C-Reactive Protein (CRP) Levels in Immune Checkpoint Inhibitor Response and Progression in Advanced Non-Small Cell Lung Cancer: A Bi-Center Study. *Cancers (Basel).* 12, 2319 (2020).

75. Sun, L. *et al.* Association Between KRAS Variant Status and Outcomes With First-line Immune Checkpoint Inhibitor–Based Therapy in Patients With Advanced Non–Small-Cell Lung Cancer. *JAMA Oncol.* 7, 937 (2021).

76. Ricciuti, B. *et al.* Diminished Efficacy of Programmed Death-(Ligand)1 Inhibition in STK11- and KEAP1-Mutant Lung Adenocarcinoma Is Affected by KRAS Mutation Status. *Journal of Thoracic Oncology* 17, 399–410 (2022).

77. Budczies, J. *et al.* KRAS and TP53 co-mutation predicts benefit of immune checkpoint blockade in lung adenocarcinoma. *Br. J. Cancer* 131, 524–533 (2024).

78. Zhou, H., Shen, J., Liu, J., Fang, W. & Zhang, L. Efficacy of Immune Checkpoint Inhibitors in SMARCA4-Mutant NSCLC. *Journal of Thoracic Oncology* 15, e133–e136 (2020).

79. Mezquita, L. *et al.* Association of the Lung Immune Prognostic Index With Immune Checkpoint Inhibitor Outcomes in Patients With Advanced Non–Small Cell Lung Cancer. *JAMA Oncol.* 4, 351 (2018).

80. Banna, G. L. *et al.* The lung immuno-oncology prognostic score (LIPS-3): a prognostic classification of patients receiving first-line pembrolizumab for PD-L1 ≥ 50% advanced non-small-cell lung cancer. *ESMO Open* 6, (2021).

81. Obermeyer, Z. & Emanuel, E. J. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine* 375, 1216–1219 (2016).

82. Bertsimas, D. & Wiberg, H. Machine Learning in Oncology: Methods, Applications, and Challenges. *JCO Clin. Cancer Inform.* 885–894 (2020) doi:10.1200/CCI.20.00072.

83. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* 2, 56–67 (2020).

84. Pathan, R. K. *et al.* The efficacy of machine learning models in lung cancer risk prediction with explainability. *PLoS One* 19, (2024).

85. Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S. & Stodtmann, S. Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clin. Transl. Sci.* 17, (2024).

86. *Fundamentals of Clinical Data Science*. (Springer International Publishing, Cham, 2019). doi:10.1007/978-3-319-99713-1.

87. The Lancet Regional Health – Europe. The Italian health data system is broken. *The Lancet Regional Health - Europe* 48, 101206 (2025).

88. Garneau, W. *et al.* 76 Lessons learned during implementation of OMOP common data model across multiple health systems. *J. Clin. Transl. Sci.* 8, 20–20 (2024).

89. Saesen, R. *et al.* Defining the role of real-world data in cancer clinical research: The position of the European Organisation for Research and Treatment of Cancer. *Eur. J. Cancer* 186, 52–61 (2023).

90. Lever, J. *et al.* Facing & mitigating common challenges when working with real-world data: The Data Learning Paradigm. *J. Comput. Sci.* 85, 102523 (2025).

91. *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences*. (Springer International Publishing, Cham, 2024). doi:10.1007/978-3-031-39355-6.

92. Alafari, F., Driss, M. & Cherif, A. Advances in natural language processing for healthcare: A comprehensive review of techniques, applications, and future directions. *Comput. Sci. Rev.* 56, 100725 (2025).

93. https://learn.microsoft.com/en-us/azure/cognitive-services/language-service/text-analytics-for-health/overview?tabs=ner.

94. Torri, V., Ercolanoni, M., Bortolan, F., Leoni, O. & Ieva, F. A NLP-based semi-automatic identification system for delays in follow-up examinations: an Italian case study on clinical referrals. *BMC Med. Inform. Decis. Mak.* 24, 107 (2024).

95. Ganguly, R. & Chakraborty, S. Semi-structured Patient Data in Electronic Health Record. in 219–233 (2023). doi:10.1007/978-981-19-5184-8_12.

96. Muñoz Monjas, A., Rubio Ruiz, D., Pérez del Rey, D. & Palchuk, M. B. Enhancing real world data interoperability in healthcare: A methodological approach to laboratory unit harmonization. *Int. J. Med. Inform.* 193, 105665 (2025).

97. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018 (2016).

98. van Genderen, M. E., Kant, I. M. J., Tacchetti, C. & Jovinge, S. Moving Toward Implementation of Responsible Artificial Intelligence in Health Care. *JAMA* 333, 1483 (2025).

99. Golpayegani, D., Pandit, H. J. & Lewis, D. Comparison and Analysis of 3 Key AI Documents: EU's Proposed AI Act, Assessment List for Trustworthy AI (ALTAI), and ISO/IEC 42001 AI Management System. in 189–200 (2023). doi:10.1007/978-3-031-26438-2_15.

100. https://www.unisr.it/news/2022/10/unisr-con-microsoft-per-costruire-la-sanita-del-futuro.

101. Mariotti, F. *et al.* Insights into radiomics: a comprehensive review for beginners. *Clinical and Translational Oncology* 27, 4091–4102 (2025).

102. Libling, W. A., Korn, R. & Weiss, G. J. Review of the use of radiomics to assess the risk of recurrence in early-stage non-small cell lung cancer. *Transl. Lung Cancer Res.* 12, 1575–1589 (2023).

103. Chen, M., Copley, S. J., Viola, P., Lu, H. & Aboagye, E. O. Radiomics and artificial intelligence for precision medicine in lung cancer treatment. *Semin. Cancer Biol.* 93, 97–113 (2023).

104. https://pyradiomics.readthedocs.io/en/latest/features.html?

105. Nwogu, I. & Corso, J. J. Exploratory Identification of Image-Based Biomarkers for Solid Mass Pulmonary Tumors. in 612–619 (2008). doi:10.1007/978-3-540-85988-8_73.

106. Chetan, M. R. & Gleeson, F. V. Radiomics in predicting treatment response in non-small-cell lung cancer: current status, challenges and future perspectives. *Eur. Radiol.* 31, 1049–1058 (2021).

107. Ye, G. *et al.* AI-Derived Longitudinal and Multi-Dimensional CT Classifier for Non-Small Cell Lung Cancer to Optimize Neoadjuvant Chemoimmunotherapy Decision: A Multicentre Retrospective Study. Preprint at https://doi.org/10.2139/ssrn.5347781 (2025).

108. Sun, R. *et al.* A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. *Lancet Oncol.* 19, 1180–1191 (2018).

109. Wang, J. H. *et al.* Radiomic biomarkers of tumor immune biology and immunotherapy response. *Clinical and Translational Radiation Oncology* vol. 28 97–115 Preprint at https://doi.org/10.1016/j.ctro.2021.03.006 (2021).

110. Chang, R. *et al.* Predicting chemotherapy response in non-small-cell lung cancer via computed tomography radiomic features: Peritumoral, intratumoral, or combined? *Front. Oncol.* 12, (2022).

111. Coroller, T. P. *et al.* CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiotherapy and Oncology* 114, 345–350 (2015).

112. Keek, S. A. *et al.* Investigation of the added value of CT-based radiomics in predicting the development of brain metastases in patients with radically treated stage III NSCLC. *Ther. Adv. Med. Oncol.* 14, (2022).

113. Trebeschi, S. *et al.* Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. *Annals of Oncology* 30, 998–1004 (2019).

114. Schroeder, K. E., Acharya, L., Mani, H., Furqan, M. & Sieren, J. C. Radiomic biomarkers from chest computed tomography are assistive in immunotherapy response prediction for non-small cell lung cancer. *Transl. Lung Cancer Res.* 12, 1023–1033 (2023).

115. Tonneau, M. *et al.* Generalization optimizing machine learning to improve CT scan radiomics and assess immune checkpoint inhibitors' response in non-small cell lung cancer: a multicenter cohort study. *Front. Oncol.* 13, (2023).

116. Yolchuyeva, S. *et al.* Multi-institutional prognostic modeling of survival outcomes in NSCLC patients treated with first-line immunotherapy using radiomics. *J. Transl. Med.* 22, 42 (2024).

117. Mei, T., Wang, T. & Zhou, Q. Multi-omics and artificial intelligence predict clinical outcomes of immunotherapy in non-small cell lung cancer patients.

Clinical and Experimental Medicine vol. 24 Preprint at https://doi.org/10.1007/s10238-024-01324-0 (2024).

118. Mohammed, S. *et al.* The Effects of Data Quality on Machine Learning Performance on Tabular Data. https://doi.org/10.1016/j.is.2025.102549 (2025) doi:10.1016/j.is.2025.102549.

119. Chang, T.-G. *et al.* LORIS robustly predicts patient outcomes with immune checkpoint blockade therapy using common clinical, pathologic and genomic features. *Nat. Cancer* 5, 1158–1175 (2024).

120. Li, Y. *et al.* Machine learning models for identifying predictors of clinical outcomes with first-line immune checkpoint inhibitor therapy in advanced non-small cell lung cancer. *Sci. Rep.* 12, 17670 (2022).

121. Yoo, S.-K. *et al.* Prediction of checkpoint inhibitor immunotherapy efficacy for cancer using routine blood tests and clinical data. *Nat. Med.* https://doi.org/10.1038/s41591-024-03398-5 (2025) doi:10.1038/s41591-024-03398-5.

122. Saad, M. B. *et al.* Machine-learning driven strategies for adapting immunotherapy in metastatic NSCLC. *Nat. Commun.* 16, 6828 (2025).

123. van de Sande, D., van Genderen, M. E., Huiskens, J., Gommers, D. & van Bommel, J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. *Intensive Care Med.* 47, 750–760 (2021).

124. van de Sande, D. *et al.* To warrant clinical adoption AI models require a multi-faceted implementation evaluation. *NPJ Digit. Med.* 7, 58 (2024).

125. Palmisano, A. *et al.* AI-SCoRE (artificial intelligence-SARS CoV2 risk evaluation): a fast, objective and fully automated platform to predict the outcome in COVID-19 patients. *Radiol. Med.* 127, 960–972 (2022).

126. Boellaard, R. *et al.* FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur. J. Nucl. Med. Mol. Imaging* 42, 328–354 (2015).

127. Wasserthal, J. *et al.* TotalSegmentator: Robust Segmentation of 104 Anatomical Structures in CT images. https://doi.org/10.5281/zenodo.6802613 (2023) doi:10.5281/zenodo.6802613.

128. https://orthanc.uclouvain.be/book/plugins/ohif.html.

129. Im, H.-J., Bradshaw, T., Solaiyappan, M. & Cho, S. Y. Current Methods to Define Metabolic Tumor Volume in Positron Emission Tomography: Which One is Better? *Nucl. Med. Mol. Imaging* 52, 5–15 (2018).

130. Nioche, C. *et al.* LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity. *Cancer Res.* 78, 4786–4789 (2018).

131. https://www.siemens-healthineers.com/en-us/molecular-imaging/pet-ct/syngo-via.

132. van Griethuysen, J. J. M. *et al.* Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* 77, e104–e107 (2017).

133. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J. Digit. Imaging* 26, 1045–1057 (2013).

134. Jha, A. K. *et al.* Repeatability and reproducibility study of radiomic features on a phantom and human cohort. *Sci. Rep.* 11, 2055 (2021).

135. https://dctd.cancer.gov/research/ctep-trials/for-sites/adverse-events/ctcae-v5-5x7.pdf.

136. Schemper, M. & Smith, T. L. A note on quantifying follow-up in studies of failure time. *Control. Clin. Trials* 17, 343–346 (1996).

137. https://loft-br.github.io/xgboost-survival-embeddings/how_xgbse_works.

138. https://github.com/sebp/scikit-survival.

139. Caruso, C. M., Soda, P. & Guarrasi, V. Not Another Imputation Method: A Transformer-based Model for Missing Values in Tabular Datasets. http://arxiv.org/abs/2407.11540 (2025).

140. Yordanov, T. R. *et al.* An integrated approach to geographic validation helped scrutinize prediction model performance and its variability. *J. Clin. Epidemiol.* 157, 13–21 (2023).

141. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med.* 13, 1 (2015).

142. Zwanenburg, A. *et al.* The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* 295, 328–338 (2020).

143. Prelaj, A. *et al.* Real-world data to build explainable trustworthy artificial intelligence models for prediction of immunotherapy efficacy in NSCLC patients. *Front. Oncol.* 12, (2023).

144. Prelaj, A. *et al.* Machine Learning Using Real-World and Translational Data to Improve Treatment Selection for NSCLC Patients Treated with Immunotherapy. *Cancers (Basel).* 14, (2022).

145. Teno, J. M. Garbage in, Garbage out—Words of Caution on Big Data and Machine Learning in Medical Practice. *JAMA Health Forum* 4, e230397 (2023).

146. Munzone, E. *et al.* Development and Validation of a Natural Language Processing Algorithm for Extracting Clinical and Pathological Features of Breast

Cancer From Pathology Reports. *JCO Clin. Cancer Inform.* https://doi.org/10.1200/CCI.24.00034 (2024) doi:10.1200/CCI.24.00034.

147. Ferber, D. *et al.* Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nat. Cancer* 6, 1337–1349 (2025).

148. https://digicore-cancer.eu/.

149. https://ohdsi.github.io/CommonDataModel/index.html.

150. Öjlert, Å. K. *et al.* 1364P Federated analysis of overall survival (OS) by location of metastases (mets) in patients (pts) with metastatic NSCLC (mNSCLC) from the Digital Oncology Network for Europe (DigiONE). *Annals of Oncology* 35, S859 (2024).

151. https://www.ohdsi.org/wp-content/uploads/2024/10/85-Moreira-Clinically-validated-line-of-therapy-algorithm-for-patients-with-metastatic-Non-Small-Cell-Lung-Cancer-Jie-Yeap.pdf.

152. Sarrio-Sanz, P. *et al.* A Novel Decision Tree Model for Predicting the Cancer-Specific Survival of Patients with Bladder Cancer Treated with Radical Cystectomy. *J. Clin. Med.* 13, 2177 (2024).

153. Ogliari, F. R. *et al.* Exploring machine learning tools in a retrospective case-study of patients with metastatic non-small cell lung cancer treated with first-line immunotherapy: A feasibility single-centre experience. *Lung Cancer* 199, 108075 (2025).

154. Cortellini, A. *et al.* Transformer-based AI approach to unravel long-term, time-dependent prognostic complexity in patients with advanced NSCLC and PD-L1 ≥50%: insights from the pembrolizumab 5-year global registry. *J. Immunother. Cancer* 13, e012423 (2025).

155. Tomaszewski, M. R. & Gillies, R. J. The Biological Meaning of Radiomic Features. *Radiology* 298, 505–516 (2021).

156. Zhang, W., Guo, Y. & Jin, Q. Radiomics and Its Feature Selection: A Review. *Symmetry (Basel).* 15, 1834 (2023).

157. Perniciano, A., Loddo, A., Di Ruberto, C. & Pes, B. Insights into radiomics: impact of feature selection and classification. *Multimed. Tools Appl.* 84, 31695–31721 (2024).

158. Provenzano, L. *et al.* Integrating radiomics and real-world data to predict immune checkpoint inhibitor efficacy in advanced non-small-cell lung cancer. *ESMO Real World Data and Digital Oncology* 10, 100182 (2025).

159. Vanguri, R. S. *et al.* Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. *Nat. Cancer* 3, 1151–1164 (2022).

160. Demircioğlu, A. Are deep models in radiomics performing better than generic models? A systematic review. *Eur. Radiol. Exp.* 7, 11 (2023).

161. Cavinato, L. *et al.* Imaging-based representation and stratification of intra-tumor heterogeneity via tree-edit distance. *Sci. Rep.* 12, 19607 (2022).

162. Sollini, M. *et al.* Methodological framework for radiomics applications in Hodgkin's lymphoma. *Eur. J. Hybrid Imaging* 4, 9 (2020).

163. Rizzo, S. *et al.* Radiomics: the facts and the challenges of image analysis. *Eur. Radiol. Exp.* 2, 36 (2018).

164. Yousefi, B. *et al.* Combining radiomic phenotypes of non-small cell lung cancer with liquid biopsy data may improve prediction of response to EGFR inhibitors. *Sci. Rep.* 11, 9984 (2021).

165. Guo, Y. *et al.* Non-invasive prediction of NSCLC immunotherapy efficacy and tumor microenvironment through unsupervised machine learning-driven CT radiomic subtypes: a multi-cohort study. *International Journal of Surgery* 111, 6592–6603 (2025).

166. Lin, X. *et al.* Intratumoral and peritumoral PET/CT-based radiomics for non-invasively and dynamically predicting immunotherapy response in NSCLC. *Br. J. Cancer* https://doi.org/10.1038/s41416-025-02948-z (2025) doi:10.1038/s41416-025-02948-z.

167. Ramlee, S., Manavaki, R., Aloj, L. & Escudero Sanchez, L. Mitigating the impact of image processing variations on tumour [18F]-FDG-PET radiomic feature robustness. *Sci. Rep.* 14, 16294 (2024).

168. Ogliari, F. R. *et al.* Association of metabolic tumour volume (MTV) and total lesion glycolysis (TLG) with survival in patients with oligometastatic non-small-cell lung cancer treated with immunotherapy: a multicentre retrospective study. *Eur. J. Nucl. Med. Mol. Imaging* https://doi.org/10.1007/s00259-025-07557-9 (2025) doi:10.1007/s00259-025-07557-9.

169. Suresh, K., Severn, C. & Ghosh, D. Survival prediction models: an introduction to discrete-time modeling. *BMC Med. Res. Methodol.* 22, 207 (2022).

170. Kourou, K. *et al.* Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis. *Comput. Struct. Biotechnol. J.* 19, 5546–5555 (2021).

171. https://proceedings.mlr.press/v146/hu21a/hu21a.pdf.

172. Arango-Argoty, G. *et al.* Pretrained transformers applied to clinical studies improve predictions of treatment efficacy and associated biomarkers. *Nat. Commun.* 16, 2101 (2025).

173. Shwartz-Ziv, R. & Armon, A. Tabular data: Deep learning is not all you need. *Information Fusion* 81, 84–90 (2022).

174. Zhou, X. *et al.* Scalable Transformer for High Dimensional Multivariate Time Series Forecasting. in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* 3515–3526 (ACM, New York, NY, USA, 2024). doi:10.1145/3627673.3679757.

175. Park, S. Y., Park, J. E., Kim, H. & Park, S. H. Review of Statistical Methods for Evaluating the Performance of Survival or Other Time-to-Event Prediction Models (from Conventional to Deep Learning Approaches). *Korean J. Radiol.* 22, 1697 (2021).

176. Astley, J. R. *et al.* Explainable deep learning-based survival prediction for non-small cell lung cancer patients undergoing radical radiotherapy. *Radiotherapy and Oncology* 193, 110084 (2024).

177. Luo, Q. *et al.* Time-dependent interpretable survival prediction model for second primary NSCLC patients. *Int. J. Med. Inform.* 195, 105771 (2025).

178. Xie, Y. *et al.* Development of a machine learning-based prognostic model for survival prediction in patients with lung cancer brain metastases using multicenter clinical data. *Int. J. Med. Inform.* 203, 106025 (2025).

179. Koyama, J. *et al.* Artificial intelligence-based personalized survival prediction using clinical and radiomics features in patients with advanced non-small cell lung cancer. *BMC Cancer* 24, 1417 (2024).

180. Farooq, A., Mishra, D. & Chaudhury, S. Survival Prediction in Lung Cancer through Multi-Modal Representation Learning. in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* 3907–3915 (IEEE, 2025). doi:10.1109/WACV61041.2025.00384.

This thesis was produced using resources from **Ministerial Decree No. 351 of April 9, 2022**, under the National Recovery and Resilience Plan (PNRR) – funded by the European Union – NextGenerationEU – Mission 4 – *"Education and Research"*, Component 1 – *"Enhancement of the education services offer: from early childhood to university"*, Investment 4.1 – *"Extension of the number of PhD courses and innovative PhDs for public administration and cultural heritage"*