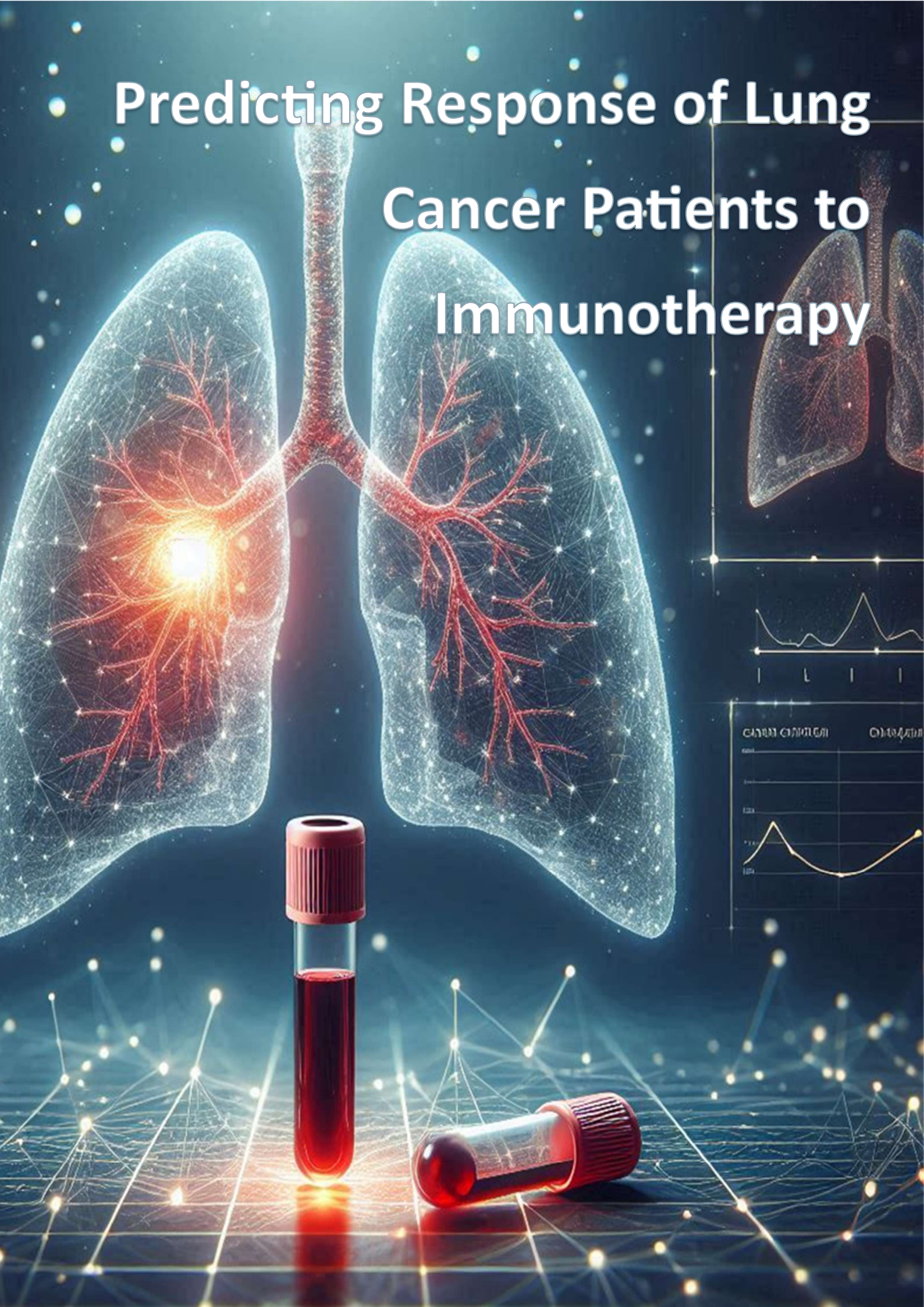
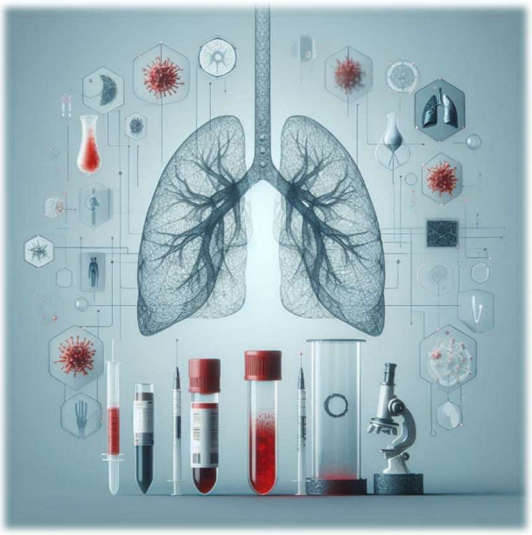
Predicting Response of Lung Cancer Patients to Immunotherapy





PREDICTING RESPONSE OF LUNG CANCER PATIENTS TO IMMUNOTHERAPY

Based on routinely determined bloodwork and CT scans

Master Thesis | Technical Medicine | Medical Imaging and Interventions

Esmee van Uum (s1899643)

Date of colloquium: 4th march 2025 13:00

Thesis committee

Chair: Prof. I. Sechopoulos, PhD
Professor at Multi-Modality Medical imaging (M3i)
Radboud University Medical Centre & University of Twente, Nijmegen & Enschede, The Netherlands

Medical supervisor: R.C. Boshuizen, MD, PhD
Pulmonologist Oncologist, Department Pulmonary Medicine
Deventer Hospital, Deventer, The Netherlands

Technical medicine supervisor: I. van der Loo, MSc
Technical Physician, Science Office
Deventer Hospital, Deventer, The Netherlands

Technical supervisor: C.O. Tan, PhD
Associate Professor, Faculty of Electrical Engineering, Mathematics and Computer Science, Robotics and Mechatronics
University of Twente, Enschede, The Netherlands

Process supervisor: R.J. Lambers, MSc
Lecturer Professional Behaviour, Faculty of Science and Technology
University of Twente, Enschede, The Netherlands

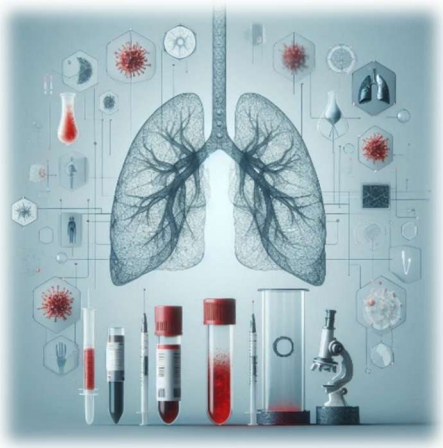
External member: ir. AG de Groot
PhD-Candidate, Faculty Electrical Engineering, Mathematics & Computer Science
University of Twente, Enschede, The Netherlands

UNIVERSITY OF TWENTE.

Deventer ziekenhuis

‘In the end, it cannot be doubted that each of us can see only a part of the picture. The doctor sees one, the patient another, the engineer a third, the economist a forth ‘…’ Human knowledge is never contained in one person. It grows from the relationships we create between each other and the world around us and still it is never complete. ‘

- When breath becomes air by Paul Kalanithi



ACKNOWLEDGEMENTS

Before you lies my master's thesis, "Predicting Response of Lung Cancer Patients to Immunotherapy". This work has been written to fulfil the graduation requirements of the Master's program in Technical Medicine, within the track Medical Imaging and Interventions, at the University of Twente in Enschede, The Netherlands. This thesis is the culmination of an incredible journey of exploration, learning and development. I would like to take this opportunity to express my gratitude to the people who guided me during this M3 year:

Ioannis, I appreciate your guidance, introducing me to helpful contacts, and facilitating support and feedback for this thesis. Rogier, thank you for constantly challenging me in clinical settings, keeping me sharp, and encouraging me to explore. Your open-mindedness and guidance have been helpful throughout this process. Iris, thank you for tirelessly reading my drafts providing me feedback, even when they stretched over 20+ pages and were still in chaos state. And thank you for your support and help in making this opportunity available for me. Can, I am grateful for the fresh insights and innovative methods you gave me, providing me with new ways or methods to solve problems and sometimes a put on place that I should not make it too complicated for myself. Rianne and my 'intervisie' group, thank you for asking the questions needed, to help me guide my professional development and allow me to grow even more.

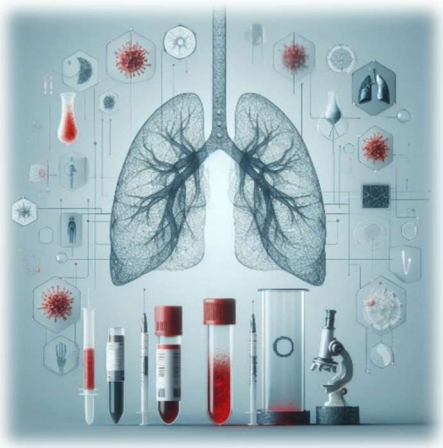
My parents, thank you for your unwavering support throughout my studies. Mom, your emotional support, grocery deliveries (despite my protests), and car rides home were always appreciated and much enjoyed. Above all, your and dads love and belief in me. Dad, thank you for always being willing to read my drafts and discuss my work, even when it was not always clear exactly what I was doing. Your willingness to engage and offer new perspectives has been deeply meaningful. And of course thanks for the amount of kilometres I was allowed to put on your car. ;) Stan, thank you for your patience, understanding, and love, even when my studies took precedence. My friends, thank you for providing me with plenty of joyful memories of my studying years. Nadine, as a fellow M3 student, it has been a pleasure to share this journey with you. Your companionship made this year all the more memorable.

To the reader, I hope you will enjoy reading a project I am proud to have created and completed.

Sincerely,

Esmee van Uum

Enschede, March 4, 2025



ABSTRACT

Research aim

This research investigates the predictive power of routinely collected blood values and CT scans in forecasting the response to immunotherapy in patients with stage IV non-small cell lung cancer.

Method

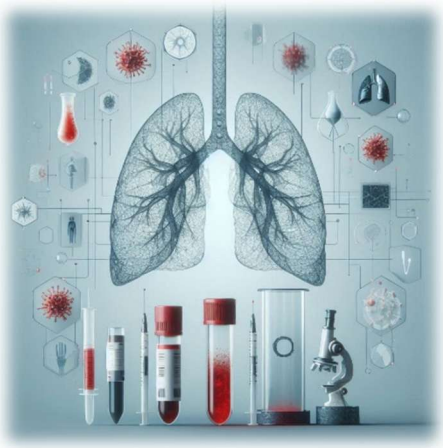
The study involved a retrospective analysis of patient data, including demographic information, clinical characteristics, blood values, and CT images. Various statistical and machine learning methods were applied, including Kaplan-Meier analysis, log-rank tests, multinomial regression, mixed-effects models, and random forest.

Results

Several blood biomarkers, particularly CRP, emerged as significant predictors of overall survival (OS) at various time points during treatment. In addition, immune cell ratios such as NLR, PLR, and LMR demonstrated notable prognostic value. Blood values obtained after the initiation of therapy showed a stronger association with OS compared to baseline values. Although the nnU-Net achieved the highest Dice scores (0.56) for automated tumor segmentation from CT scans, these scores were insufficiently high to reliably extract radiological features. Mixed-effects models (MEM) and random forest (RF) models that integrated blood values and clinical data demonstrated potential for more accurate prediction of immunotherapy response. However, the low feature importance scores in the RF models indicated that the response to immunotherapy is shaped by a complex interplay of factors rather than by a single dominant feature.

Conclusion

This study provides initial insights into the relationship between clinical parameters and treatment outcomes in immunotherapy for NSCLC. Although it did not achieve the goal of identifying patients who would benefit the most from treatment, it demonstrated potential for developing models that can potentially identify non-responders.



CONTENTS

ACKNOWLEDGEMENTS 4

ABSTRACT 5

CONTENTS 6

INTRODUCTION 11

CHAPTER 1: CLINICAL BACKGROUND 12

1.1 Non-Small Cell Lung Cancer 12

1.2 Treatment 12

1.3 Immunotherapy 12

1.4 Imaging 14

1.5 Biomarkers in NSCLC 14

CHAPTER 2: TECHNICAL BACKGROUND 16

2.1 Statistics 16

2.1.1 Kaplan Meier Method & Log-Rank Test 16

2.1.2 Cox Hazard Model 16

2.1.3 Multinomial regression 17

2.1.3 Mixed Effect Model 17

2.2 Machine learning 18

2.3 Image segmentation 18

CHAPTER 3: SEGMENTATION NETWORK 20

3.1 Introduction 20

3.3 Results 21

3.4 Discussion 24

3.4.1 Clinical implications 25

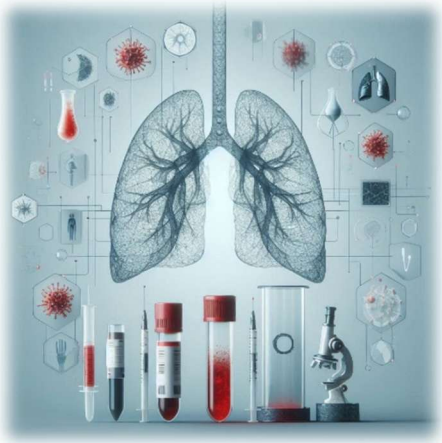
CHAPTER 4: BASIC STATISTICS 26

4.1 Introduction 26

4.2 Method 26

4.2.1 Patient Population & Data Collection 26
4.2.2 Statistical Analysis 27
4.3 Results 27
4.3.1 Kaplan-Meier and log-rank results 28
4.3.2 Multinomial Regression Analysis 30
4.3.3 Cox Regression Analysis 31
4.4 Discussion 32
4.4.1 Comparison with Existing Literature 33
4.4.2 Limitations 34
4.4.3 Future Directions 34
4.5 Conclusion 35
CHAPTER 5: MIXED EFFECT MODELS 36
5.1 Introduction 36
5.2 Method 36
5.3 Results 37
5.4 Discussion 38
5.4.1 Comparison with literature 39
5.4.2 Future directions 39
5.5 Conclusion 40
CHAPTER 6: RANDOM FOREST 41
6. 1 Introduction in the random forest 41
6.2 Method 41
6.3 Results 42
6.4 Discussion 44
6.4.1 Clinical implications 44
6.4.2 Limitations 45
6.4.3 Recommendations for Future Research 46
6.5 Conclusion 46
CHAPTER 7: SUMMARY & RECOMMENDATIONS 48
7.1 Summary 48
7.2 Clinical Relevance 49
7.3 Recommendations for Future Research 49
7.4 Conclusion 50
REFERENCES 51
Appendix A: Mixed Effect Models explained 56
Appendix B: Preprocessing types 58

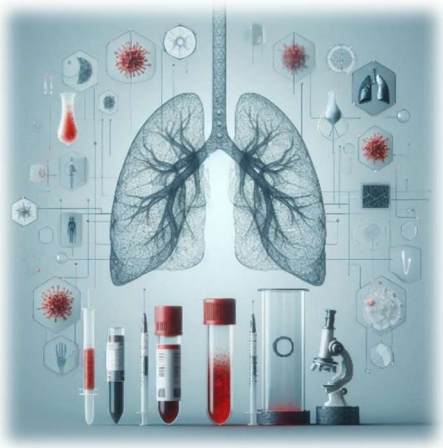
Appendix C: Neural Networks explained 60
Convolutional Neural Networks 60
The U-Net 61
Appendix D: Descriptive statistics 63
Appendix E: Survival function PD-L1 65
Appendix F: Blood variables per model 67
Appendix G: Results linear models 69



ABBREVIATIONS

AI	Artificial Intelligence
AIC	Akaike Information Criterion
ALAT	Alanine aminotransferase
ANC	Absolute Neutrophil Count
ASAT	Aspartate transaminase
AUC	Area Under the Curve
BIC	Bayesian Information Criterion
CNN	Convolutional Neural Network
CR	Complete Response
CRP	C-reactive Protein
CT	Computed Tomography
CTLA-4	Cytotoxic T-lymphocyte associated protein 4
DSC	Dice Similarity Coefficients
FT4	Free T4
GA(M)M	Generalized Additive (Mixed) Models
GLMM	Generalized Linear Mixed effect Model
GTV	Gross Tumour Volume
ICI	Immune Checkpoint Inhibitors
IL	Interleukin
irAEs	Immune-Related Adverse Events
JI	Jaccard index
KNN	K-Nearest Neighbours
LDH	Lactate dehydrogenase
LIPI	Lung Immune Prognostic Index
LMR	Lymphocyte-to-Monocyte Ratio
LMM	Linear Mixed effect Model
MAR	Missing at Random
MCV	Mean Corpuscular Volume
MDSC	Myeloid-Derived Suppressor Cell
MEM	Mixed Effect Model
ML	Machine Learning
MR	Mixed Response
NK	Natural Killer
NLR	Neutrophil-to-Lymphocyte Ratio
NPV	Negative Predictive Value
NSCLC	Non-Small Cell Lung Cancer
OS	Overall Survival
PD	Progressive Disease
PD-1	Programmed cell Death protein 1
PD-L1	Programmed cell Death Ligand 1
PFS	Progression-Free Survival
PLR	Platelet-to-Lymphocyte Ratio
PR	Partial Response
rCR	Radiological Complete Response
(i)RECIST	Immune Response Evaluation Criteria In Solid Tumours
ReLU	Rectified Linear Unit

RF	Random Forest
RMSE	Root Mean Square Error
RNN	Recurrent Neural Networks
SCLC	Small Cell Lung Cancer
SD	Stable Disease
SHAP	Shapley Additive Explanations
SII	Systemic Immune Inflammatory Index
SVM	Support Vector Machines
TIL	Tumour-Infiltrating Lymphocytes
TLS	Tumour Lysis Syndrome
TMB	Tumour Mutation Burden
TNM	Tumour, Lymph Node and Metastasis
TSH	Thyroid-Stimulating Hormone
VIF	Variance Inflation Factor
XGB	Extreme Gradient Boosting



INTRODUCTION

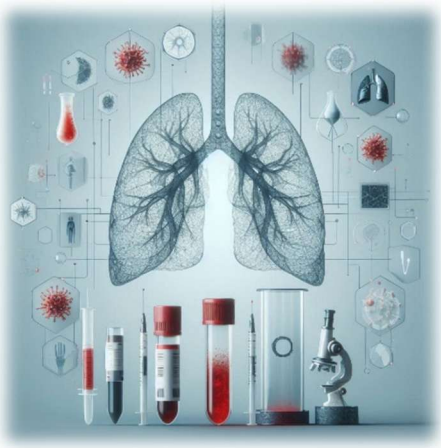
Immunotherapy has increasingly been adopted as one of the most effective approaches to treating Stage IV non-small cell lung cancer (NSCLC) [1-5]. However, not everyone benefits from immunotherapy and it can potentially lead to side effects. Therefore, it is important to identify those who will benefit from therapy and those who will not. This will not only help minimize side effects and minimize losing time to useless therapy but also reduces the overall cost of therapy. To address this need, Deventer Hospital collected data from all Stage IV lung cancer patients who received at least one dose of (chemo)immunotherapy between January 3, 2017, and June 1, 2024. This database provides an opportunity to analyse the routinely determined bloodwork to predict response to immunotherapy and to analyse available CT scans to provide radiological features to identify long-term survival in the early stages.

The primary aim of this research is to develop a predictive model that can identify individuals who will benefit from immunotherapy and those who will not.

Research questions

Based on the research objective the following research questions are formulated:

1. Can routinely determined bloodwork help predict response to immunotherapy?
    - Which routinely determined blood markers can be a marker to predict response to immunotherapy?
    - Could combinations of blood markers be used to predict response to immunotherapy?
2. Can CT scans help determine the long-term survival of patients to immunotherapy?
    - How can tumours automatically be segmented out of CT images?
    - How can the information needed to determine the prognosis be automatically generated from CT images?
    - What information in CT scans is helpful in determining the prognosis?
3. Can an automatic algorithm or model be created that predicts the response of patients to immunotherapy?
    - Which algorithms can and will be used by healthcare providers?
    - How is an algorithm or model created to predict the response to immunotherapy?



CHAPTER 1:
CLINICAL BACKGROUND

1.1 Non-Small Cell Lung Cancer

Lung cancer is a global public health concern, representing the most commonly diagnosed cancer [6, 7]. With more than 1.4 million deaths annually, it accounts for up to 18% of all cancer-related deaths with a poor 5-year survival rate of approximately 15% [7-9]. Lung cancer is broadly categorized into two main types based on histology: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). NSCLC accounts for approximately 80% to 85% of all lung cancer cases, while SCLC constitutes about 10% to 15% of cases [2, 9]. Although NSCLC tends to grow and metastasize less rapidly than SCLC, it often does not show a good response to first-line chemotherapy and/or radiation therapy [10]. Advancements in treatment, like immunotherapy, offer hope for improved outcomes. However, much remains to be done to improve the overall 5-year survival rate of 92% in stage IA1, decreasing to 13% in stage IIIC and below 5% for stage IV [2, 11-13].

1.2 Treatment

The treatment options for NSCLC have been evolving over time and are currently based on classified stage (Figure 1) [2, 3, 12, 14]. For resectable NSCLC, guidelines recommend surgery for stages I–II and a select set of stage III patients, with post-surgery chemotherapy for stages II–III. [3, 15, 16]. In stage IV, and some advanced stage III cases, treatment with chemotherapy and radiotherapy is often not successful, but recent advancements in immunotherapy have significantly prolonged overall survival (OS) and progression-free survival (PFS) [1-5].

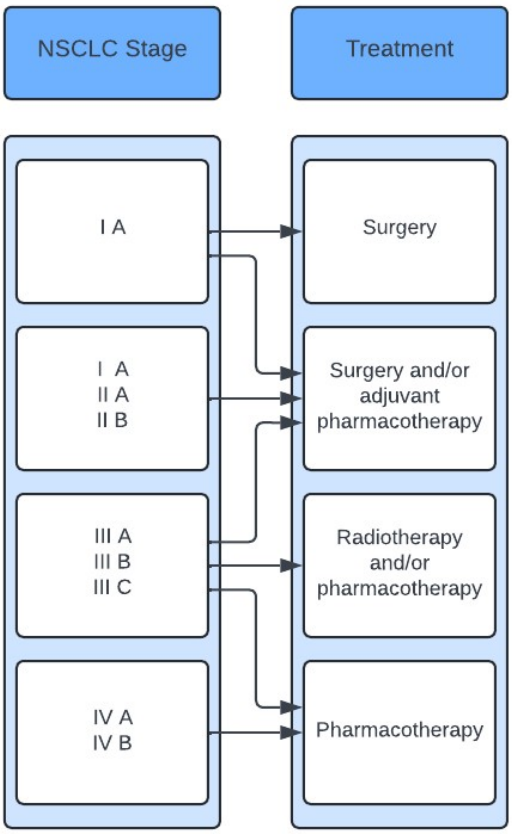


Figure 1: Treatment options in NSCLC.

1.3 Immunotherapy

Immunotherapy is a form of cancer treatment that harnesses the body's immune system to identify and destroy cancer cells. The immune system is designed to recognize and eliminate abnormal cells through a process known as immunosurveillance. In the context of cancer, some genetic mutations in tumour cells can produce abnormal antigens that the immune system can detect, leading to the activation of immune cells such as T cells to target and destroy cancer cells. However, tumours can evolve through a process called cancer immunoediting, which allows them to escape immune control and continue growing. Immunotherapy aims to overcome this

immune evasion by boosting the immune system's ability to recognize and attack tumour cells [17]. An important immunotherapeutic strategy in NSCLC is the immune checkpoint inhibitors (ICIs), which block the pathways tumours use to suppress immune responses. Examples of ICIs are anti-programmed cell death protein 1 (PD-1) and anti-cytotoxic T-lymphocyte-associated protein 4 (CTLA-4) antibodies (Figure 2) which have become central to treatment, particularly in advanced cases [18, 19].

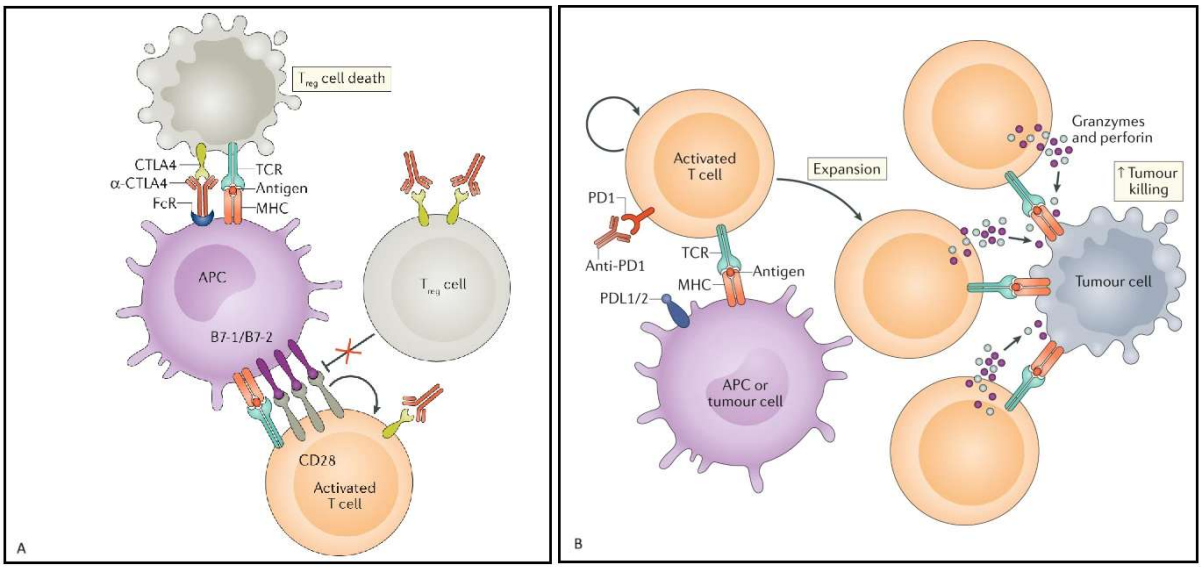


Figure 2: ICIs mechanisms. A) CTLA-4 blocking mechanism. B) PD1/PD-L1 mechanism. Retrieved from: https://www.nature.com/articles/s41577-020-0306-5

In the neoadjuvant setting, immunotherapy can accelerate the immune system's activation, targeting both the primary tumour and any micro metastatic disease. This can lead to a higher rate of mediastinal nodal clearance and potentially a greater rate of pathological complete response [18, 19]. For patients with advanced NSCLC, anti-PD-1 (an ICI) is recommended as a first-line treatment when PD-L1 expression is ≥50%. Combination therapies with ICIs and chemotherapy are preferred for lower PD-L1 expressions [4, 9, 12]. Unlike traditional chemotherapy, which targets rapidly dividing cells indiscriminately, ICIs enhance the immune system's precision in targeting cancer cells, potentially leading to long-lasting remissions [20]. Despite its benefits, immunotherapy is associated with a distinct set of side effects that can occur at any point during and after treatment. Many side effects happen when the activated immune system also acts against healthy cells and tissues in the body, causing immune-related adverse events (irAEs) [4, 21]. An example is the PD-1/PD-L1 pathway, which plays a crucial role in immune homeostasis and the suppression of T cell activity against autoantigens. With immunotherapy blocking this pathway, an immune attack can be triggered causing irAEs. Currently, it is unknown when or if side effects will occur or how serious they will be. Although immunotherapy for advanced NSCLC generally has fewer high-grade toxicities than traditional chemotherapy [22], patients treated with ICIs can experience unpredictable and potentially fatal toxicities affecting almost all tissues and organs, particularly the skin, colon, endocrine glands, liver, and lungs [22]. While most irAEs are mild to moderate, severe irAEs occur in up to 20% of patients on single-agent therapy and about 60% of those on combination of anti-PD-1 and anti-CTLA-4 drugs [5]. These severe side effects can lead to treatment interruptions [5, 8] and the need for steroid therapy, which may be associated with poorer survival outcomes [23]. The response to immunotherapy in NSCLC patients can vary between patients, with a subset of patients achieving significant clinical benefits. While immunotherapy can produce durable responses and long-term remissions in some patients, it remains unclear why others fail to respond [4, 21, 24, 25]. The high cost of treatment, which can exceed $100,000 per patient per year, or even $200,000 with combination therapies in the USA, further underscores the need to identify patients who are most likely to benefit from these therapies [5]. Therefore, there is a need for predictive biomarkers that can distinguish between responders and non-responders, helping to optimise treatment decisions and avoid unnecessary side effects and financial burdens [24, 25]. Developing reliable predictive tools would be a significant advancement in the field of immunotherapy, improving its effectiveness while minimising costs and adverse effects.

1.4 Imaging

Computed tomography (CT) is very important in the management of lung cancer patients because it allows for non-invasive visualization of tumours before, during, and after treatment. CT imaging allows for the assessment of tumour size, mediastinal and vascular invasions, and the presence of distant metastases, helping to stage the cancer [3, 26]. Additionally, it is also used to evaluate the effect of treatment on the cancer. Response evaluation is needed in determining the effectiveness of the treatment, which can be done with radiological response or pathological response. CT imaging is used to evaluate the radiological response to treatment. A standardized way for response assessment are the Response Evaluation Criteria In Solid Tumours (RECIST) [27]. This method assesses the largest diameter of tumours and categorizes the treatment effect into complete response (CR), partial response (PR), progressive disease (PD), or stable disease. For patients receiving immunotherapy the iRECIST (immune RECIST) was created, it addresses how tumours respond differently to immunotherapies compared to chemotherapies [28]. This was created because of the phenomenon called pseudoprogression, pseudoprogression mimics disease progression although it is not. What makes it hard to distinguish pseudoprogression form progression is that only time tells if the progression is in fact progression or a reduction in tumour size will become apparent [21]. iRECIST builds on RECIST 1.1 but includes specific criteria for immunotherapy responses. This approach ensures a consistent way of tracking tumour response in trials, facilitating better data collection and analysis of treatment response. Several radiographic features can be evaluated in CT imaging, including:

- Size: Determination of tumour size measured as diameter in X, Y, Z direction.
- Volume: Determination of full tumour volume.
- Density: Evaluation of tumour density.
- Internal Features: Characteristics of the tumour's internal structure, such as necrosis, calcifications, or cystic areas.
- External Features and associated findings: Characteristics of the tumours external and surrounding, including spiculation, lymph node involvement, vascular invasion, surrounding tissue reaction or amount of volume/size reduction over time.

Combinations of the mentioned radiographic features can potentially also be used in algorithms to predict cancer status and to predict response over time [29, 30].

1.5 Biomarkers in NSCLC

In oncology, biomarkers are particularly valuable for identifying patients who are most likely to benefit from specific treatments, such as immunotherapy. By enabling personalized treatment approaches, biomarkers help optimise therapeutic efficacy and minimise unnecessary side effects, thus improving patient outcomes and resource utilisation [4, 9, 21]. For NSCLC several biomarkers have been researched for their role in predicting response to immunotherapy:

- PD-L1 Expression: High PD-L1 expression is associated with better response rates to ICIs like pembrolizumab and nivolumab. It is therefore nowadays the most used predictive biomarker for immunotherapy, in which in mainly determines if a patient will receive solely immunotherapy or a combination with chemotherapy [15, 22, 33].
- Tumour Mutation Burden (TMB): TMB measures the total number of mutations per mega base of DNA. High TMB has been linked to improved responses to ICIs due to the increased likelihood of tumour cells presenting antigens. The Checkmate-227 and Checkmate-568 studies showed that TMB predicts clinical benefits of nivolumab plus ipilimumab in patients with NSCLC, regardless of PD-L1 expression [31, 32]. However, The Keynote-021 and Keynote-189 studies show inconsistent outcomes, creating an uncertainty around the predictive value of TMB [33].

In gastric cancers the presence and density of tumour-Infiltrating Lymphocytes (TILs), indicated a favourable prognosis and potential responsiveness to immunotherapy, as it reflects the immune system's engagement with the tumour [30]. Despite the significance of the biomarkers, the clinical use is limited by invasive biopsies, which are often infeasible, unsuitable for monitoring disease response, and may not represent tumor heterogeneity accurately [34].

Peripheral blood-based biomarkers offer a less invasive alternative for patient selection and treatment monitoring that is continuously easy to access. Some blood markers (Figure 3) have already shown their potential in previous studies within different tumour types [4, 25, 35], like lymphocyte and neutrophil count [36], or neutrophil to lymphocyte ratio [4]. Some of the routinely determined blood-based markers include:

- **Eosinophils**: The accumulation of eosinophils has been associated with diverse prognostic outcomes in various cancers. Responders to ICI treatment showed a significant increase in eosinophil counts compared to non-responders, suggesting that eosinophil counts could serve as an early predictive marker for immunotherapy response [37]. When looking at the relation between eosinophil count and survival the literature shows incongruent results, sometimes being a positive predictor and otherwise predicting a negative outcome [25, 37-39].
- **Peripheral Blood Cell Counts**: Markers such as lymphocyte count, neutrophil count, and the neutrophil-to-lymphocyte ratio (NLR) are routinely measured and have demonstrated potential in predicting responses to immunotherapy. For example, a high NLR might signal a more favourable response to treatment [15, 36].

Nevertheless, most of the available results are preliminary, so the potential biomarkers still have to be investigated within NSCLC or cannot be implemented into routine clinical practice until they are validated in large-scale trials. These trials should also take into account the differences in the application of the potential biomarkers alone (or combinations of markers), and standardise thresholds for the guidance of clinical decision making. Therefore, further research opportunities remain abundant, with the potential to refine biomarker applications and enhance their clinical utility.

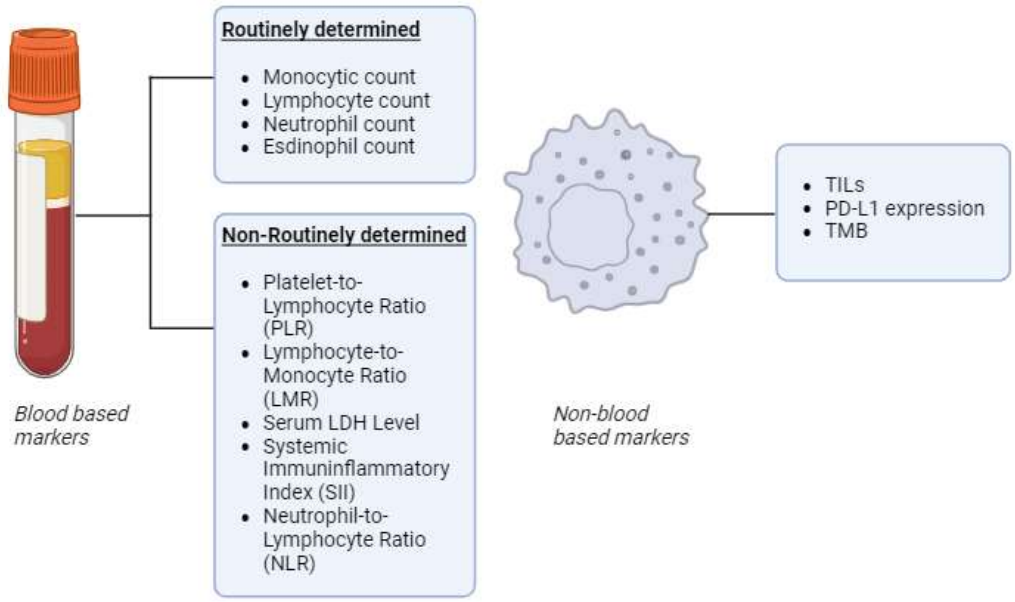
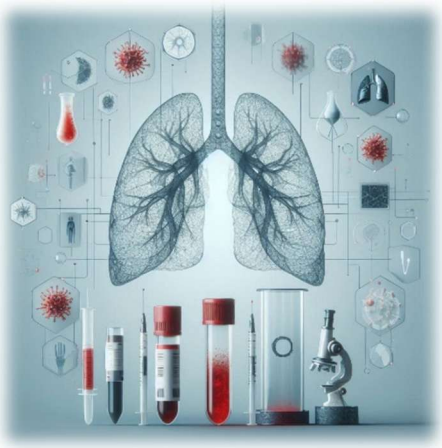


Figure 3: (Possible) markers for immunotherapy [4, 9, 35, 40, 41]. Created in BioRender.com



CHAPTER 2:
TECHNICAL BACKGROUND

To address these challenges and advance the clinical applicability of potential biomarkers, a robust analytical approach is essential. The following chapter delves into the technical background, exploring the statistical methods, AI models, and segmentation networks that can be used to determine whether a biomarker can predict a patient's response to immunotherapy.

2.1 Statistics

2.1.1 Kaplan Meier Method & Log-Rank Test

Kaplan-Meier curves and the log-rank test are statistical tests that are often used in survival analysis [42]. The Kaplan-Meier method estimates the survival function, which is the probability of "surviving" beyond a certain time point [42]. The Kaplan-Meier curve plots survival probability (y-axis) against time (x-axis). The curve is a step function where the survival probability drops vertically whenever one or more outcome events occur, and remains horizontal between events. The general formula for Kaplan-Meier survival probability at a failure time is expressed in equation 1 [43].

$$(1) \quad S\big(t(f)\big) = S\big(t(f-1)\big) \times \left(\frac{number\ of\ individuals\ surviving\ just\ before\ t(f)}{number\ of\ individuals\ at\ risk\ at\ t(f)}\right)$$

The log-rank test provides an overall comparison of the Kaplan-Meier curves [48]. It tests the hypothesis that there is no difference in survival between two or more groups over time. The log-rank test statistic is a chi-square statistic calculated using the observed and expected event counts at each failure time as depicted in equation 2 and 3 [44].

$$(2) \quad X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$(3) \quad E_i = \frac{d_j \times n_{ij}}{n_j}$$

In which O_i is the observed number of events in group i and Ei is the expected number of events in group i. d_j is the total number of events at time j, n_{ij} is the number of individuals at risk in group i at time j and n_j is the total number of individuals at risk at time j across all groups.

2.1.2 Cox Hazard Model

The Cox Proportional Hazards Regression Model predicts the OS and PFS outcomes in relation to one or more predictive variables. It can be used to assess the relationship between different blood values and survival outcomes [45, 46]. The Cox model is a statistical technique used for survival-time outcomes, assessing the probability that the event of interest (death or progression) occurs before a given time. It models the hazard function $\lambda(t)$ (equation 4) as an exponential function of an arbitrary baseline hazard $\lambda 0(t)$ where all covariates are zero. The regression coefficient *b* quantifies the effect of the covariate *x* on the hazard.

(4) $\lambda(t) = \lambda_0(t)e^{bx}$

The model assumes that survival curves for different groups have hazard functions proportional over time and that the relationship between the log hazard and each covariate is linear [46]. Given the availability of multiple measurements of blood values at different time points, time-dependent covariates in the Cox model can be used to account for changes in the values over time. This approach allows for an assessment of how values influence survival outcomes throughout the treatment time. To explore combinations of values, multiple values can be used as predictor values in the Cox model. Creating interaction between terms allows for assessment of association between the combined effect of blood values and its effect on overall survival and progression-free survival. This approach offers insights into potential relationships between blood values and their impact on patient outcomes even with time-dependent changes and competing risks [45-47].

2.1.3 Multinomial regression

Multinomial regression is a statistical method used when the outcome variable is categorical with more than two possible groups. When predicting patient response to immunotherapy, it estimates the likelihood that a patient will fall into one of several response categories, such as complete response, partial response, stable disease, or progression, based on a set of predictor variables. In multinomial logistic regression, each response category is compared to a reference response category through dummy coding (1/0 variables) [48]. This results in one less than the amount of response categories, binary logistic regression models, each with its own intercept and coefficients. The models reveal how predictors influence the probability of each outcome category relative to the reference group. Each category has its own intercept and coefficients, reflecting that the predictors may impact each category differently [48, 49]. Equation 5 is the general equation of the probability that the outcome Y is in category j given the predictors X, equation 6 the reference category and equation 7, the likelihood of the parameters belonging to a specific category compared to the reference category.

(5) $$P\,(Y = j \mid X\,) = \frac{e^{\beta_{j0} + \beta_{j1}x_1 + \beta_{j2}x_2 + \cdots + \beta_{jp}x_p}}{1 + \sum_{m=2}^{k} e^{\beta_{m0} + \beta_{m1}x_1 + \beta_{m2}x_2 + \cdots + \beta_{mp}x_p}}$$

(6) $$P\,(Y = 1 \mid X\,) = \frac{1}{1 + \sum_{m=2}^{k} e^{\beta_{m0} + \beta_{m1}x_1 + \beta_{m2}x_2 + \cdots + \beta_{mp}x_p}}$$

(7) $$\log\left(\frac{P\,(Y = j \mid X\,)}{P(Y = 1 \mid X\,)}\right) = \beta_{j0} + \beta_{j1}x_1 + \beta_{j2}x_2 + \cdots + \beta_{jp}x_p$$

Where Y is a categorical outcome variable with k categories, X represents the predictor variables (x_1, x_2, … , x_p). $P\,(Y = j \mid X\,)$ *is the probability that the outcome Y is in category j given the predictors X.* β_{j0} *is the intercept for category j and* $\beta_{j1}, \cdots, \beta_{jp}$ *are the regression coefficients for the predictors corresponding to category j. Each category has its own set of coefficients* β_j,

2.1.3 Mixed Effect Model

A Mixed Effects Model (MEM) is a statistical tool used to predict a single continuous variable like OS or PFS using two or more other variables [50-52]. MEMs can be suitable to predict a patient's response to immunotherapy due to their ability to handle repeated measures. MEMs can incorporate multiple independent variables, such as blood values and CT image metrics like size and volume, measured at various points during treatment. This allows for an analysis of how these variables interact and influence OS or PFS over time. It provides insights into the relationships between the input variables with the output variable [50, 51, 53, 54]:

(8) $$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_j + e_{ij}$$

Where Y_{ij} *is the response variable for observation i within group j,* β_0 *is the intercept (fixed effect),* $\beta_1 X_{ij}$ *is the fixed effect for predictor X,* u_j *is the random effect for group j, which varies between groups and* ϵ_{ij} *is the residual error term.*

To create a MEM, the hierarchical structure of the data must be identified. For example, in medical research, repeated measures from the same patient can be considered hierarchical data. Fixed and random effects must be determined, variables of primary interest and variables representing random samples from a larger population [52]. By modelling the repeated measurements and accounting for both fixed and random effects, MEMs help to control for potential confounding factors and improve the accuracy of predictions. This approach

reduces the likelihood of false positives and negatives [54]. However, proper model specification and validation are essential to ensure accurate and interpretable results. For more information, appendix A.

2.2 Machine learning

Artificial Intelligence (AI) performs tasks that previously required human-like intelligence, including learning, reasoning and problem-solving. Machine Learning (ML), a subset of AI, enables computers to perform specific tasks by learning from data instead of explicit programming [55]. ML has two main categories [56]:

1. Supervised learning uses a labelled datasets where algorithms learn by comparing predictions to known outcomes, adjusting parameters to minimize errors. This works well with limited, labelled data.
2. Unsupervised learning identifies patterns or structures in the data without predefined labels, used in larger datasets.

In healthcare, ML algorithms can enhance treatment processes like treatment planning [57]. AI has also enabled the development of predictive models for assessing treatment possibilities, responses and potential side effects in cancer therapy [58, 59].

2.3 Image segmentation

A common application of AI in medical imaging is tumour segmentation - isolating and outlining the tumour on the image. Segmentation can be performed manually, semi-automatically, or through deep learning algorithms [60]. The process requires preprocessing to improve the efficiency and performance of the model, as images are often obtained from different scanners with various image acquisition and reconstruction protocols [61, 62]. Preprocessing can include steps like normalization [63], noise reduction, standardization of size and data augmentation [64], see appendix B for more information about preprocessing steps. Once preprocessing has been done, deep learning is often used for segmentation in cancer research [65]. Deep learning [55, 66] is a subset of ML that is based on artificial neural networks, like the brains neural network, see figure 4 and appendix C for more explanation. These neural networks, inspired by the human brain's structure, consist of interconnected layers of nodes or neurons. Deep learning networks like Convolutional Neural Networks (CNNs) excel in processing high-dimensional data such as images [60, 67]. A popular architecture for segmentation tasks is the U-Net [64], which is a type of CNN designed to be suitable for image segmentation. It consists of an encoder-decoder structure: the encoder takes an image as the input of the model and extracts necessary features and relevant information, whereas the decoder learns to generate the corresponding predictions (probability maps) [68]. U-Net is often used in medical image segmentation due to its ability to achieve adequate performance even with limited datasets. Together, the CNN-based methods provide powerful tools for lung cancer segmentation, combining efficiency, accuracy, and the ability to work with diverse data.

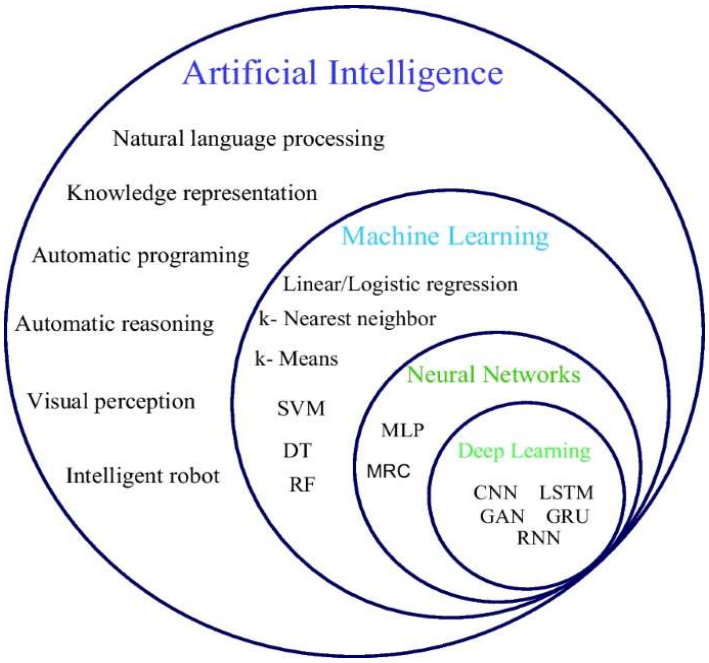
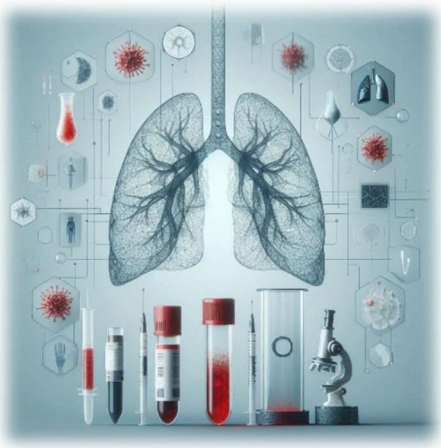


Figure 4: Architecture of AI to deep learning networks. Retrieved from: https://link.springer.com/article/10.1007/s10661-024-12443-2/figures/2



CHAPTER 3:
SEGMENTATION NETWORK

3.1 Introduction

Accurately identifying lung tumours in NSCLC is important for extracting radiomic features that can predict patient responses to treatment. Segmentation networks can play a role in this process by isolating and outlining tumours in CT images, enabling analysis of tumour characteristics. This chapter evaluates the performance of four segmentation networks designed to optimize tumour segmentation in CT images. These networks were selected based on their reported Dice similarity coefficients (DSC) in previous studies [61, 68-70], which make them suitable candidates for this task, see table 1. This chapter systematically compares the networks based on their segmentation performance on NSCLC CT images, to find their ability to detect primary and secondary tumour lesions.

Table 1: information per segmentation network.

	Network 1 (Dune) [61]	**Network 2 (DS) [68]**	**Network 3 (Vermond) [69]**	**Network 4 (nnU-net) [70]**
Dice score	0.82	0.88	0.71	0.84
training data	1328 scans of 8 different institutions	64 scans from the medical segmentation decathlon challenge	1781 scans from three datasets: Medical segmentation decathlon, NSCLC radiomics Lung-PET-CT-Dx	96 scans from Medical Decathlon Segmentation Challenge
Type of network	2D U-Net architecture	MobileNetV2 encoder with a U-Net decoder	teacher-student framework	nnU-net

3.2 Method

A total of 100 CT images from a publicly available NSCLC dataset [71] were used. Corresponding ground truth segmentation masks annotated by radiologists served as the reference standard. These masks were used to evaluate the performance of each network using a range of metrics. The DSC was used to measure the overlap between predicted tumour masks and ground truth annotations. The Jaccard Index (JI) was calculated to assess the similarity between predicted and reference masks. Hausdorff Distance was employed to measure the maximum spatial discrepancy between tumour boundaries. Additionally, precision, sensitivity, and specificity were computed to evaluate the classification accuracy of the segmentations. Finally, the Spearman correlation was used to investigate the relationship between the Dice score (segmentation quality) and tumour volume as most networks are trained or on small nodes or on larger tumours. A second Spearman correlation was done between volume of radiologist and of the AI prediction, to address consistency in the AI network. Spearman's rank correlation is a non-parametric test that assesses the strength and direction of the association between

two ranked variables. All the above mentioned metrics were computed twice for each network: (1) considering only the primary lesion and (2) including all detected lesions in the scan. This allowed for a comprehensive comparison of the networks in both single and multiple tumour detection scenarios.

3.2.1 Segmentation Networks and Steps

Network 1 [61] begins with preprocessing steps aimed at standardizing the CT images. These images were normalized to ensure consistent intensity values across scans and harmonized to account for different image acquisition protocols. The images were then cropped and padded to ensure uniform dimensions. The first step in the segmentation process was lung isolation, where a region-growing algorithm was applied to the CT scans, followed by morphological operations like dilation and erosion to remove small irrelevant structures. After lung isolation, a modified 2D U-Net architecture was used to segment the tumours. The output consisted out of 2D binary masks that were stacked to generate a 3D volume. Post-processing steps included extracting connected components from the generated mask, followed by resampling it to match the original dimensions of the CT scans.

Network 2 [68] uses a hybrid architecture that combines MobileNetV2 as the encoder and U-Net as the decoder. The first step in preprocessing involved normalizing the CT images using min-max scaling, followed by standardization to resize the images to a fixed size of 256×256 pixels. The preprocessed images were used in the mobileNetV2 encoder and then the U-Net decoder. The tumour segmentation masks are reconstructed in this model by skip connections with the ReLU activation function to link the encoder and decoder layers.

Network 3 [69] introduces a teacher-student framework, which uses two types of annotations: semantic 3D annotations (strong annotations) and 2D bounding boxes in axial planes (weak annotations). The preprocessing step involves converting the CT scans into 3D volumes and normalizing the voxel intensities. The teacher model generates pseudo-strong labels for the weakly annotated data. The student model performs end-to-end segmentation.

Network 4 [70], nnU-Net, is a self-configuring framework designed to automatically adapt to new datasets and segmentation tasks. No extra manual preprocessing is needed. The model uses 3D full-resolution U-Net with setting the lung segmentation weights, it predicts segmentations. Post-processing steps in nnU-Net include removing small regions and selecting the largest connected components to ensure that the tumour segmentation is accurate and free of artifacts.

The steps of implementing these networks were followed and then the metrics were computed on the resulting segmentations. 3D slicer was used to provide the CT scans in the right format (Nifti, NRRD). For analyses python was used, version 3.7-3.10 depending on the usage needed per network.

3.3 Results

All metrics results can be found in table 2. Figure 5 shows segmentations of all the networks of one patient out of the dataset. The analysis of the correlation between volumes and dice scores provided the following results: the Pearson correlation coefficient for the dune network was 0.1007 with a P-value of 0.3188. The network provided 3 empty predictions. For the DS network, the correlation coefficient was 0.3167 with a P-value of 0.0013. The network provided 10 empty predictions. For the Vermund Mask network, the correlation coefficient was -0.0035 with a P-value of 0.9723. The network provided 1 empty prediction. For the nnU-Net network, the Spearman correlation coefficient was -0.1137 with a P-value of 0.2601. The network provided 14 empty predictions. Figure 6 shows all scatterplots of the networks on dice score versus volume. Figure 7 shows the scatterplots of radiologist volume vs AI volume. Additionally, the analysis of the correlation between AI volume and Radiologist volume yielded the following Spearman correlation coefficients:

- Network 1: 0.7292, P-value: 0.0000
- Network 2: 0.3524, P-value: 0.0003
- Network 3: 0.6979, P-value: 0.0000
- Network 4: 0.4237, P-value: 0.0000

Table 2: Results metrics for all four networks, for both primary tumour and all locations of tumour.

	DSC	Jaccard Index	Hausdorff Distance	Precision	Sensitivity	Specificity
Network 1: Dune (all)	0.479	0.372	53.6	0.767	0.405	1
Network 1: Dune (primary)	0.522	0.419	33.8	0.734	0.46	1
Network 2: DS (all)	0.212	0.144	33.8	0.476	0.172	1
Network 2: DS (primary)	0.215	0.147	46.49	0.415	0.184	1
Network 3: Vermond (all)	0.365	0.258	103.34	0.55	0.342	1
Network 3: Vermond (primary)	0.415	0.302	123.49	0.533	0.434	1
Network 4: nnU-net (all)	0.523	0.409	96.09	0.688	0.492	1
Network 4: nnU-Net (primary)	0.561	0.45	91.76	0.645	0.574	1

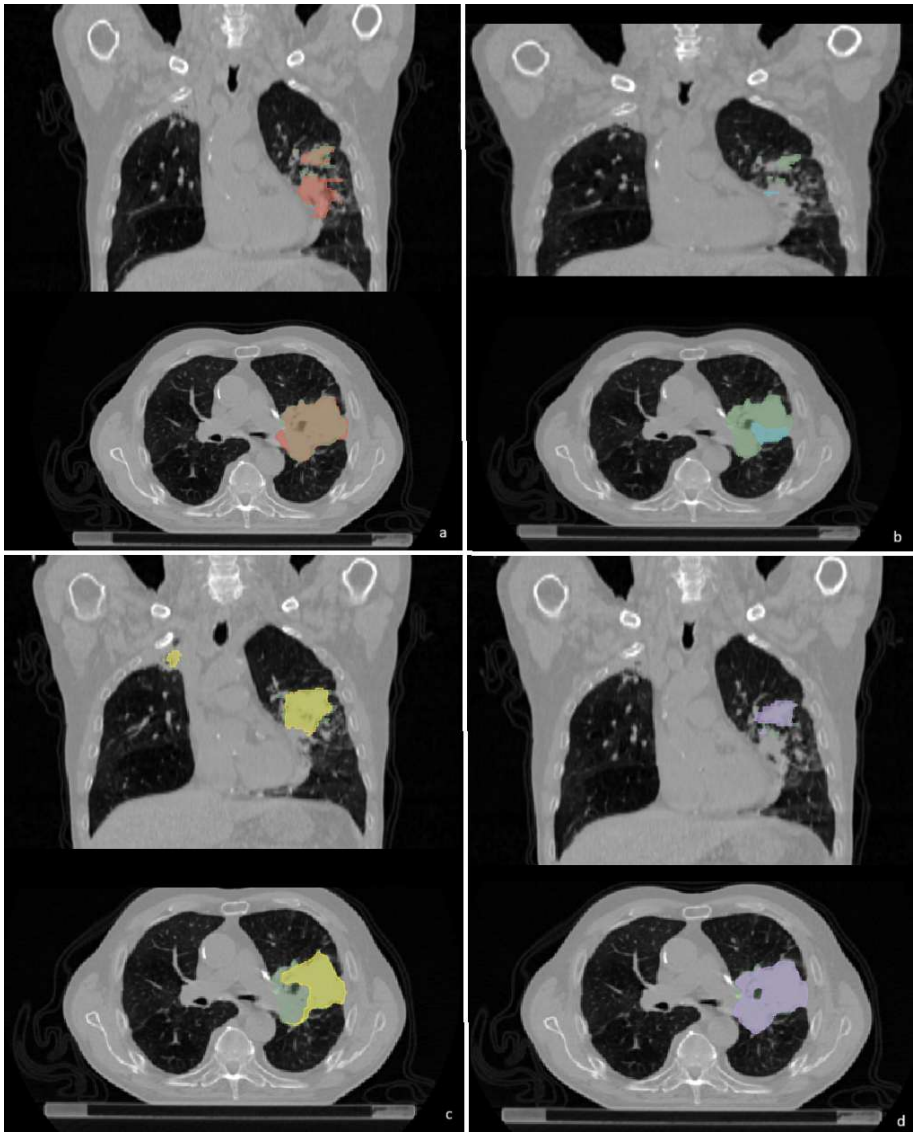


Figure 5: Segmentation results of the networks combined with the ground truth masks (green). a) Network 1: the dune network, b) Network 2: the DS network, c) Network 3: the Vermond network and, d) Network 4: the nnU-Net. All images are within 10 slices of each other, showcasing the slices that best illustrate the networks in action.

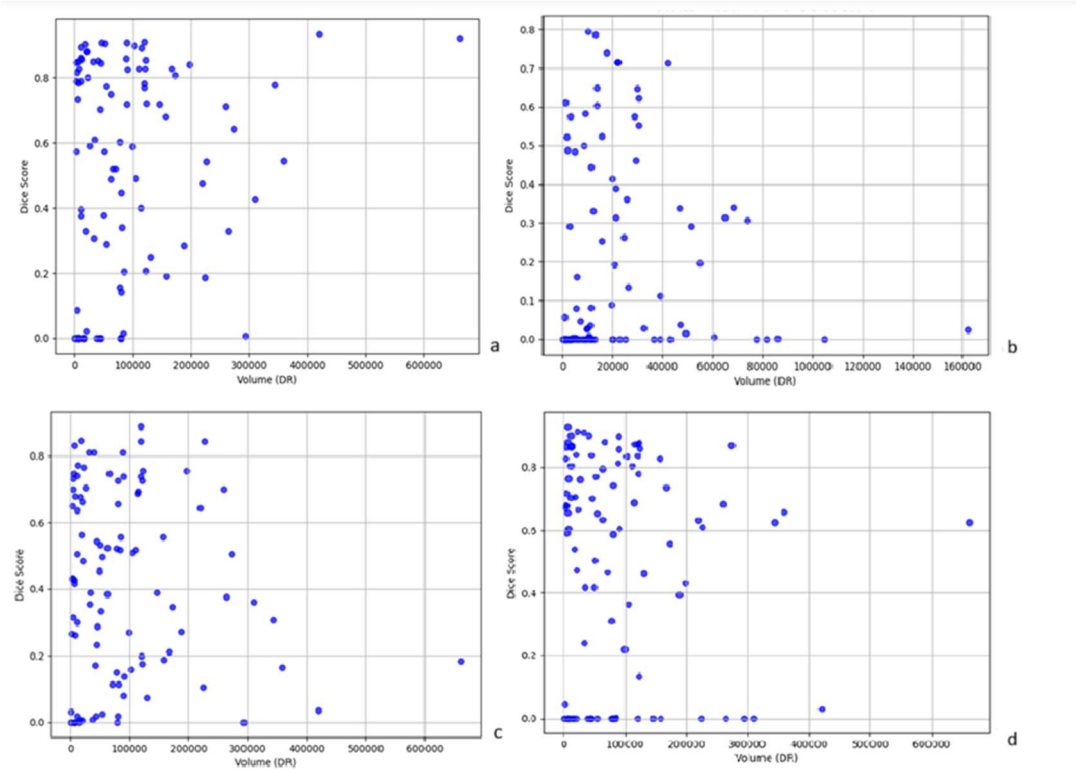


Figure 6: Scatterplots of volume vs dice scores. a) network 1: the dune network, b) network 2: the DS network, c) network 3: the Vermond network and, d) network 4: the nnU-Net.

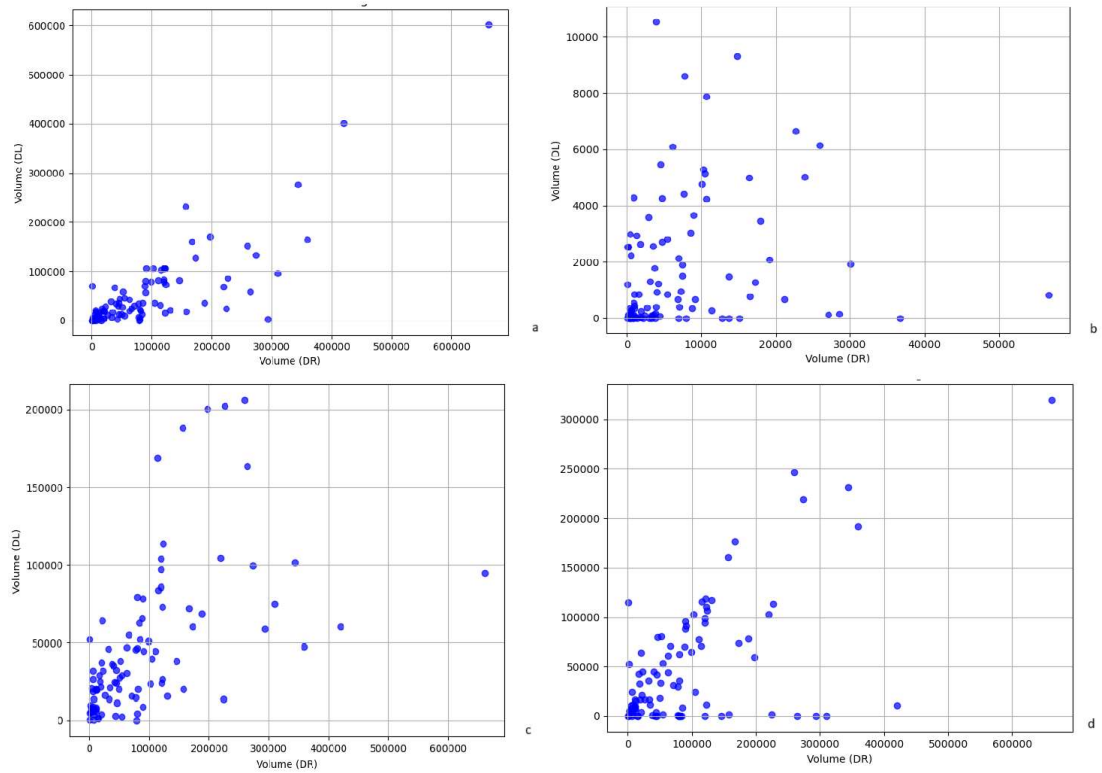


Figure 7: Scatterplots of volume of radiologist (DR) vs volume of AI network (DL). a) network 1: the dune network, b) network 2: the DS network, c) network 3: the Vermond network and, d) network 4: the nnU-Net.

3.4 Discussion

Accurate segmentation of tumours in NSCLC is an important step in understanding tumour characteristics and extracting features for therapy strategy. The four segmentation networks evaluated in this study demonstrated varying degrees of performance in segmenting lung tumours from CT images, as reflected in DSC, precision, and other metrics. Below, the performance of each network is discussed, contextualizing their strengths and limitations in light of their original design and application to the current research.

Networks 2 showed a small correlation between tumour size and DSC scores. Unlike network 1, 3 and 4 which show no significant correlation. This suggests that segmentation accuracy of these networks is less influenced by tumour size, making these networks more robust for segmenting tumours of varying sizes.

The correlation analysis between AI-predicted volumes and radiologist-determined volumes provides insight into the consistency of different networks. Network 1 and Network 3 demonstrated the highest correlation with radiologist volumes (0.7292 and 0.6979, respectively), suggesting that these models generate segmentations that are more proportional to those of the radiologist. Network 2 and Network 4 showed lower correlations (0.3524 and 0.4237), indicating higher variability in AI-predicted volumes relative to the radiologist's assessment. A high correlation does not necessarily imply perfect agreement but rather suggests that the AI model maintains a consistent relationship with the radiologist's measurements. This consistency is particularly important when considering the use of AI-generated volumes in predictive models, as stability in predictions may be more valuable than exact volume replication. Further analysis could explore the potential impact of AI volume variations on downstream clinical applications and whether models with higher volume correlations also yield more reliable predictions in patient outcome modelling.

The first network originally reported strong performance, with metrics including a DSC of 0.82, a Jaccard Index of 0.72, and an H95 of 9.43 mm [61]. However, on our dataset, these scores were lower (higher on the Hausdorff distance), possibly due to the network's focus on segmenting primary NSCLC tumours. This focus may lead to the exclusion of secondary tumour components, which lowers accuracy, although the effect is expected to show in dice score of all tumour lesions and not or less in the primary dice scores. The networks reliance on identifying the largest connected component (GTV-1) as the tumour can help cause this issue, as it risks losing portions of the tumour when components are not fully connected. Addressing this limitation could involve improving the network's handling of disconnected tumour regions.

The second network achieved a DSC of 0.88, a recall of 0.86, and a precision of 0.93 in its original study [68]. Its potential lays in the transfer learning that offers computational efficiency. For future use, the computational efficiency is important in clinical applications where time and resources are constrained, although the primary focus now is the current network struggle with segmentation consistency. The main limitation of this network is the small dataset, which contains only 64 annotated CT images. This limited training dataset could hinder the network's ability to generalize to more diverse cases. A significant limitation noted in the original study [68] is that the model's performance was only validated on the same dataset, and broader testing on independent and more diverse medical imaging datasets is necessary to confirm its generalizability to more diverse cases. In our experience, this limitation became evident when evaluating its accuracy on our dataset, where it revealed limitations with dice scores of just above 0.2.

Network 3 originally reported a DSC of 0.71 [69]. In this study, it achieved Dice scores just below 0.5, reflecting average performance compared to the other networks tested. According to its original research, the network's performance is sensitive to preprocessing parameters, in particular voxel spacing, which is not accounted for in their described preprocessing steps [69]. Improving these parameters through methods such as linear or nearest-neighbour interpolation could help align the input data with the conditions of the training dataset, potentially enhancing performance. Another concern is the overlap between the training and testing datasets. The network was trained using three public datasets: the Medical Segmentation Decathlon (MSD)-Lung, the NSCLC-Radiomics, and Lung-PET-CT-Dx. This includes part of the test dataset used in this research, potentially introducing bias and inflating performance metrics.

The last network achieved DSC scores of 0.82 for lung nodules and 0.84 for lung tumours in its original study [70]. In our research, it outperformed the other networks, with a DSC of 0.561 for primary tumours and 0.523 for all tumour regions. Its automated configuration capabilities and robust preprocessing pipeline might have allowed it to generalize better to our dataset, although there is definitely room for improvement. The network also showed limitations in a high rate of empty predictions (14 out of 100 cases). While its automated

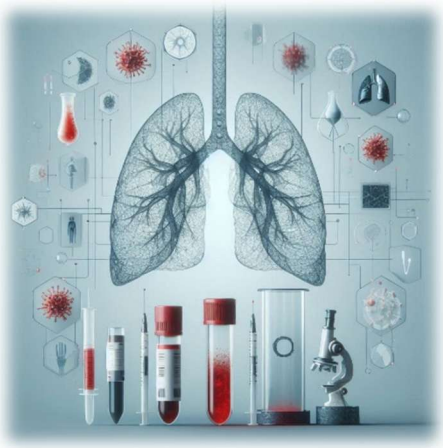
design simplifies application, it may restrict manual optimization for specific challenges within a dataset. Despite these issues, nnU-Net's adaptability and user-friendliness make it a strong candidate for NSCLC segmentation tasks.

3.4.1 Clinical implications

A limitation across all networks was their reliance on pretreatment CT scans for training. This training approach may lead to decreased segmentation accuracy on post-treatment scans, where treatment-induced changes complicate tumour segmentation. For instance, networks may over-segment areas of pseudoprogression or miss smaller regions altered by treatment, providing a higher change in volume than actually has been accomplished with treatment. Therefore, the networks should be trained in segmenting post treatment CT-scans as well as the pretreatment scans. Another consideration and the main aim of this research is the use of automated segmentations for extracting radiological features. Current Dice scores are insufficient for reliably identifying tumour areas and therefore the extraction of radiological features. Developing better-performing networks tailored to the specific dataset is necessary. The nnU-Net is a promising option for training on the intended dataset as it is a small dataset. However, using the same dataset for training and validation poses challenges. Using some of the data for training and validation would reduce the number of patients available for testing later in the research. This trade-off makes nnU-Net training on this dataset less feasible. In the future, this approach could still be viable if manually segmented CT radiological features were incorporated into the statistical or machine learning models along with the automatically segmented CT radiological features of the other patients. However, this would induce variability in predictions due to differences in how the segmentations were gathered. While this method was not pursued in the current study due to time constraints, it could serve as a focus for future research aimed at finding or training a better-functioning network. As automated segmentation offers significant advantages over manual methods, including faster processing times and reproducibility. Radiologist and radiation oncologists preferred automated segmentation in 56% of cases [61], underscoring its potential value, but also current precaution. While current networks require further refinement, automated segmentation tools could significantly improve the accuracy, speed, and reproducibility of NSCLC tumour segmentation, ultimately enhancing diagnosis and in words of this research: treatment planning/prediction.

3.5 conclusion

Network 4 was the best performing network researched, although the average dice score only reached 0.56. The results highlight that there is no "one-size-fits-all" solution for lung tumour segmentation. Further research is needed to compare the performance of segmentation networks across diverse datasets and to evaluate the factors influencing segmentation accuracy. Therefore, these networks will not be used to add radiological features to predicting response to therapy.



CHAPTER 4:
BASIC STATISTICS

4.1 Introduction

The recent advances in immunotherapy have shown promise in improving outcomes for patients with advanced NSCLC. While past research has mainly concentrated on the mentioned markers in chapter 1, it could be advantageous to broaden the scope of inquiry. Additional blood markers, or a combination of those markers in ratios, that are regularly tested during therapy could be examined. By incorporating the testing of additional markers, an easily accessible marker will potentially be provided. In general, follow-up response assessment every 6–12 weeks is recommended for iRECIST [28]. This moment also provides new information about the blood markers that could help predict if someone is responding to immunotherapy. Other factors may also hold predictive value. For example, tumours with a higher mutational load tend to be more sensitive to immunotherapy drugs, as they present a greater number of antigens recognizable by the immune system [21]. In NSCLC trials, smokers were found to have better responses compared to non-smokers. Another potential predictive marker studied before was the loss of muscle mass [41]. While clear recommendations are still lacking in this regard, in the absence of available biomarkers, these epidemiological findings may help guide challenging therapeutic decisions in clinical practice. This chapter aims to evaluate various blood markers and non-blood markers on their predictive and prognostic value in NSCLC patients undergoing immunotherapy.

Table 3: Blood values that will be tested on their possible predictive value in response to immunotherapy.

Potential markers literature	**Commonly determined blood values**
Baseline absolute lymphocyte count	ALAT
Absolute neutrophil count	ASAT
NLR	MCV
Absolute Eosinophil Count	TSH
Absolute monocyte Count	Creatinine
Systemic Immunoinflammatory Index (SII)	FT4
Serum LDH Level	Haemoglobin
Platelet-to-Lymphocyte Ratio (PLR)	Leukocytes
Lymphocyte-to-Monocyte Ratio (LMR)	Platelets
	Sodium
	Potassium
	Glucose
	Calcium
	LDH

4.2 Method

4.2.1 Patient Population & Data Collection

This study focused on patients with stage IV NSCLC who received (the first) immunotherapy between January 3, 2017, and August 18, 2023. Inclusion criteria required baseline blood samples, CT images, and complete clinical data. Data were collected from medical records (HiX electronic patient file system) and included demographic

details, disease stage, comorbidities, treatment information, and blood sample results at baseline (before start of therapy: timepoint 0) and during therapy (1 & 3 months, after start: timepoint 1 and 3).

4.2.2 Statistical Analysis

Primary outcome measures were OS and PFS. OS was defined as the duration from treatment initiation to death from any cause, while PFS was the time from treatment initiation to disease progression as determined by imaging. Response to therapy was assessed based on the best response observed. Blood markers analysed included routinely determined values as well as potential other markers like the NLR, PLR, and SII. Table 3 shows all assessed blood markers. Descriptive statistics like the mean, median, standard deviation and range were used to provide a first insight into all variables. Kaplan-Meier method with log-rank tests were employed to identify significant biomarkers in time till death or time till progression, when indicated as in the normal or abnormal range of the blood value. When one or more out of three time points was abnormal it was set to abnormal. The multinomial regression analysis was performed to understand the influence of various features on different therapy outcomes (CR, PR, mixed response (MR), stable disease (SD) and PD). A positive coefficient indicates an increased likelihood of a better response (CR, PR, or SD) relative to PD, while a negative coefficient indicates a decreased likelihood. The blood values were further analysed using the Cox (time dependent) regression model to estimate their association with OS and PFS. Missing values were accounted for by excluding the patient solely for the test it has a missing value for. All statistical analyses were performed in Python and R. Benjamini-Hochberg correction was performed to control for the false discovery rate.

4.3 Results

Descriptive statistics can be found in table 4, more can be found in appendix D. The mean overall survival time is 637 days and the mean progression free time is 538 days.

Table 4: Descriptive Statistics for non-blood value group parameters

Variable	Level	Amount (% of total)
n		216
Survival time	0 (till 365 days)	121
	1 (past 365 days)	95
Progression free time	0 (till 365 days)	147
	1 (past 365 days)	69
Sex M/V (%)	0 (V)	85 (39.4)
	1 (M)	131 (60.6)
Smoking (At diagnosis) (%)	FALSE	139 (64.4)
	TRUE	77 (35.6)
Smoking (In the past) (%)	FALSE	86 (38.3)
	TRUE	140 (61.7)
Tumour type (0: adenocarcinoma, 1: squamous cell carcinoma, 2: large cell carcinoma, 3: other) (%)	0	145 (67.1)
	1	52 (24.1)
	2	9 (4.2)
	3	1 (4.6)
PD-L1 positive (%)	FALSE	139 (64.4)
	TRUE	77 (35.6)
T (%)	0	3 (1.4)
	1	25 (11.7)
	2	23 (10.7)
	3	39 (18.2)
	4	111 (51.9)

	x	13 (6.1)
N (%)	0	31 (14.4)
	1	10 (4.7)
	2	57 (26.5)
	3	109 (50.7)
	x	8 (3.7)
M (%)	0	5 (2.3)
	1	205 (95.3)
	x	5 (2.3)
Malignancy in medical history (%)	FALSE	170 (78.7)
	TRUE	46 (21.3)
Comorbidities (%)	FALSE	17 (7.9)
	TRUE	199 (92.1)

4.3.1 Kaplan-Meier and log-rank results

The p-values of the different blood values for the different time points included are shown in table 5. The combined analysis of all three time points revealed the following significant associations:

- Sodium and MCV showed a significant association in OS but did not significantly correlate with PFS.
- Creatinine levels and eosinophils indicated a significant association with PFS while showing no significant correlation with OS.

For the analysis of the different time points the following blood values were significant. Before the start of therapy, CRP, LDH, haemoglobin and ASAT showed significant association with OS. For PFS, this was LDH, Free T4 (FT4), ASAT and ALAT. One Month After Treatment, this changed to CRP, TSH, haemoglobin, sodium, calcium, glucose, leukocytes, Neutrophils, NLR and SII for OS. eGFR (CKD-EPI) and HB were significant in PFS. Three months after treatment start, CRP, eGFR (CKD-EPI), MCV, thrombocytes, calcium, glucose, sodium, leukocytes, lymphocytes, neutrophils and PLR are significant for OS and non are significantly associated with PFS. After correction at all three time points combined, only MCV was significant. CRP remained significant before therapy. At 1 month after, CRP, TSH, HB, Sodium, Leukocytes, glucose, Neutrophils and NLR remain significant for OS. At 3 months after, CRP, MCV, Platelets, Glucose, Sodium, Leukocytes, Lymphocytes, Neutrophils and PLR are significant. For PFS creatinine remained significant on all three timepoints and FT4 and ASAT at timepoint 0. In the non-blood values, table 6, Age was significant for OS and PD-L1 for PFS. After correction in the non-blood values both remained significant. PD-L1 showed a p-value of 0.02 for OS when the acquisition method was cytological. The P-value of PD-L1 for pathological acquisition was 0.90 in OS, see appendix E.

Table 5: Log-rank test results for OS and PFS, on all different combinations of timepoints. t:0 being before start of therapy, t:1 is 1 month after start of therapy and t:3 is 3 months after start of therapy.

	P-value OS (t:0+1+3)	P-value PFS (t:0+1+3)	P-value OS (t:0)	P-value PFS (t:0)	P-value OS (t:1)	P-value PFS (t:1)	P-value OS (t:3)	P-value PFS (t:3)
CRP	0.06	0.44	**0.00**	0.24	**0.01**	0.56	**0.01**	0.60
Bilirubin	0.86	0.46	0.87	0.20	0.77	0.86	0.91	0.61
eGFR (CKD-EPI)	0.95	0.72	0.79	0.14	0.15	**0.05**	**0.02**	0.14
TSH	0.96	0.56	0.62	0.14	**0.01**	0.79	0.72	0.54
LDH	0.47	0.28	**0.05**	**0.04**	0.62	0.53	0.94	0.08
MCV	**0.00**	0.28	0.59	0.75	0.86	0.16	**0.01**	0.96
Creatinine	0.78	**0.00**	0.20	0.59	0.30	0.20	0.13	0.68
Free T4 (FT4)	0.92	0.24	0.69	**0.01**	0.14	0.38	0.93	0.77
Haemoglobin	0.52	0.79	**0.05**	0.78	**0.00**	**0.03**	0.33	0.97
Platelets	0.86	0.19	0.63	0.21	0.10	0.70	**0.01**	0.32
Sodium	**0.03**	0.06	0.17	0.25	**0.00**	0.36	**0.00**	0.15
Potassium	0.45	0.15	0.10	0.82	0.06	0.40	0.31	0.09
Calcium	0.18	0.34	0.17	0.16	**0.02**	0.21	**0.04**	0.32
Glucose	0.46	0.93	0.29	0.55	**0.00**	0.41	**0.01**	0.96
Leukocytes	0.45	0.14	0.19	0.30	**0.00**	0.53	**0.00**	0.64
Lymphocytes	0.44	0.81	0.52	0.23	0.15	0.54	**0.01**	0.99
Neutrophils	0.31	0.06	0.71	0.72	**0.00**	0.48	**0.00**	0.54
Monocytes	0.59	0.38	0.99	0.45	0.54	0.07	0.86	0.40
Eosinophils	0.51	**0.04**	0.48	0.39	1.00	0.15	0.51	0.85
ALAT	0.88	0.61	0.10	**0.02**	0.76	0.61	0.77	0.14
ASAT	0.21	0.86	**0.04**	**0.01**	0.91	0.73	0.82	0.45
NLR	0.40	0.36	0.15	0.40	**0.00**	0.88	0.09	0.26
PLR	0.98	0.36	0.14	0.13	0.41	0.95	**0.01**	0.35
LMR	0.97	0.82	0.20	0.56	0.06	0.51	0.08	0.31
SII	0.38	0.79	0.40	0.83	**0.03**	0.33	0.08	0.45

Table 6: Log-rank test results for OS and PFS for non-blood values.

	p-value OS	p-value PFS
Age at diagnosis (below or above 70)	**0.01**	0.19
BMI	0.33	0.90
Packyears(0-1 or more)	0.19	0.06
FEV1 (% pred, below 80)	0.84	0.64
DLCO (% pred below 80)	0.74	0.75
DLCO/VA (% pred below 80)	0.47	0.62
ECOG-score (higher then 1)	0.38	0.21
Sex M/V	0.11	0.21
Smoking (At diagnosis)	0.39	0.66
Smoking (In the past)	0.56	0.37
Tumour type	0.34	0.29
PD-L1 positive	0.16	**0.00**
Metastases at diagnosis	0.48	0.24
T	0.78	0.65
N	0.22	0.43
M	0.92	0.68
(Other) malignancy in medical history	0.66	0.29
Comorbidities	0.58	0.28

4.3.2 Multinomial Regression Analysis

The coefficients that represent the log-odds of being in a response category compared to the reference category: progression, are shown in table 7. The intercept for CR(-0.26), SD(-0.35) & MR (-0.01) are negative. Sodium shows mixed effects across response categories, with a positive association for SD and PR and negative associations for CR and Mixed Response. Potassium demonstrates a negative effect in CR and SD and a positive association for Mixed Response. Calcium shows a positive association for SD (45.39) and a negative effect for the other categories. Among immune markers, lymphocytes, monocytes exhibit mixed association. Total bilirubin exhibits negative associations for all four response to PD, as well as HB, neutrophils, eosinophils, ASAT, PLR, LMR, SII. CRP, LDH, and MCV show small positive associations in most categories, and glucose is positively associated with all categories. Age and BMI exhibit positive effects across categories, such as PR (1.57 for BMI) and Mixed Response (1.62 for BMI). Packyears, FEV1, DLCO/VA, and ECOG-score show predominantly negative effects. Smoking history shows mixed effects, smoking at diagnosis shows negative associations with CR and SD. Past smoking is positively associated with CR. PD-L1 positivity correlates with CR (17.66) and Mixed Response (1.86). Comorbidities are negatively associated with CR but positively associated with SD.

Table 7: Multinomial Regression Coefficients (std. errors) for Response to Immunotherapy.

	Complete Response	**Partial Response**	**Stable Disease**	**Mixed Response**
(Intercept)	-0.26 (0.00)	0.35 (0.02)	-0.35 (0.02)	-0.01 (0.00)
Sodium	-0.46 (0.34)	0.15 (0.41)	0.58 (0.42)	-0.42 (0.55)
LDH	0.04 (0.26)	0.05 (0.11)	-0.04 (0.11)	0.02 (0.25)
CRP	0.07 (0.54)	0.11 (0.30)	0.01 (0.30)	-0.01 (0.48)
Total Bilirubin	-0.38 (0.06)	-1.42 (0.62)	-1.10 (0.59)	-0.48 (0.31)
eGFR	0.49 (0.48)	0.09 (0.40	-0.51 (0.39)	0.21 (0.63)
TSH	0.17 (0.80)	0.09 (0.80)	-0.84 (0.81)	-0.04 (0.14)
MCV	0.33 (0.30)	0.60 (0.40)	0.46 (0.39)	0.30 (0.31)
Creatinine	0.16 (0.56)	-0.19 (0.18)	-0.72 (0.18)	-0.13 (0.47)
Free T4 (FT4)	0.10 (0.46)	0.10 (0.44)	0.14 (0.45)	0.46 (0.62)
Haemoglobin	-4.00 (0.07)	-1.13 (0.69)	-8.01 (0.74)	-3.33 (0.10)
Platelets	0.09 (0.18)	0.07 (0.10)	0.05 (0.10)	0.10 (0.30)
Potassium	-1.39 (0.03)	-0.96 (0.37)	-7.99 (0.38)	2.04 (0.04)
Leukocytes	0.83 (0.27)	-1.17 (0.59)	0.65 (0.53)	-0.41 (0.12)
Glucose	1.94 (0.24)	0.96 (0.34)	2.96 (0.33)	1.42 (0.08)
Calcium	-5.26 (0.01)	-48.47 (0.19)	45.39 (0.18)	-1.33 (0.01)
Lymphocytes	-2.99 (0.09)	5.29 (0.52)	2.04 (0.50)	1.25 (0.03
Neutrophils	-1.81 (0.18)	-0.10 (0.64)	-3.09 (0.67)	-2.19 (0.15)
Monocytes	-0.07 (0.03)	-8.20 (0.29)	5.24 (0.29)	3.59 (0.02)
Eosinophils	-2.66 (0.01)	-1.78 (0.40)	-2.47 (0.39)	-0.55 (0.02)
ALAT	0.33 (0.37)	0.42 (0.41)	0.18 (0.42)	0.20 (0.50)

ASAT	-0.47 (0.29)	-0.61 (0.54)	-0.33 (0.55)	-0.32 (0.62)
NLR	4.12 (0.14)	4.00 (0.72)	3.51 (0.71)	3.90 (0.14)
PLR	-0.05 (0.15)	-0.06 (0.09)	-0.06 (0.08)	-0.08 (0.27)
LMR	-1.68 (0.08)	-6.66 (0.57)	-1.95 (0.56)	-2.56 (0.5)
SII	-0.01 (0.02)	-0.01 (0.01)	-0.00 (0.01)	-0.00 (0.03)
Age at diagnosis	0.20 (0.48)	0.73 (0.41)	0.23 (0.41)	0.57 (0.67)
BMI	0.95 (0.33)	1.57 (0.38)	1.36 (0.33)	1.62 (0.21)
Packyears	-0.13 (0.84)	-0.45 (0.48)	-0.75 (0.48)	-0.44 (0.74)
FEV1 (% pred)	-0.17 (0.88)	-0.30 (0.30)	-0.55 (0.31)	-0.31 (0.840
DLCO (% pred)	-0.08 (0.39)	0.14 (0.54)	0.99 (0.53)	-0.20 (0.41)
DLCO/VA (% pred)	-0.13 (0.66)	-0.38 (0.38)	-1.28 (0.38)	-0.23 (0.51)
ECOG-score	-7.95 (0.03)	-3.50 (0.54)	-9.20 (0.56)	-0.77 (0.06)
Sex M/V	-5.73 (0.04)	7.53 (0.29)	14.42 (0.30)	5.3 (0.6)
Smoking (At diagnosis)	-4.22 (0.02)	35.3 (0.38)	-49.99 (0.39)	-0.46 (0.04)
Smoking (In the past)	2.27 (0.02)	48.35 (0.42)	-33.90 (0.43)	0.97 (0.04)
PD-L1 positive	17.66 (0.02)	4.44 (0.22)	9.21 (0.21)	1.86 (0.02)
(Other) malignancy in medical history	6.26 (0.03)	-7.51 (0.20)	-6.01 (0.21)	-1.11 (0.04)
Comorbidities	-8.35 (0.01)	-6.91 (0.42)	12.53 (0.42)	0.70 (0.00)
1/Patientnr	-0.26 (0.00)	0.35 (0.02)	-0.35 (0.02)	-0.01 (0.00)

4.3.3 Cox Regression Analysis

Figure 7 shows all significant blood values of the different models in vulcano plots. The first analysis of all three timepoints was conducted on the data with 427 observations, with 291 events (death). A total of 122 observations were deleted due to missing data. LDH turned out significant with a p-value of <0.001, CRP had a p-value of 0.05. Each normalized unit increase in LDH was associated with a 15.2% higher risk. At baseline, the analysis included 159 patients, with 121 events, and 59 observations deleted due to missing data. Glucose showed a statistically significant HR of 1.337 with a p-value of 0.03, as ASAT was also significant. Other variables showed no significant associations ($p > 0.05$). At Month 1, the analysis included 153 observations, with 102 events. A total of 35 observations were deleted due to missing data. The HR for CRP was 1.442 with a p-value of 0.04, as well as a positive HR for SII. Calcium and platelets showed a significant HR of 0.686 and 0.499 respectively. At month 3, the analysis included 123 patients with 75 events, and 30 observations were deleted due to missing data. LDH and CRP showed up significant. The analysis identified several blood values with associations to increased risk. LDH, glucose, CRP, PLR, SII, and ASAT levels provided a positive HR at various time points. Calcium, platelets, NLR and sodium levels were associated with negative HR. Performing Lasso feature selection or forward backward selection and adding the markers that show up in these selections, showed a higher AIC and a lower concordance index, therefore it was decided not to include these results.

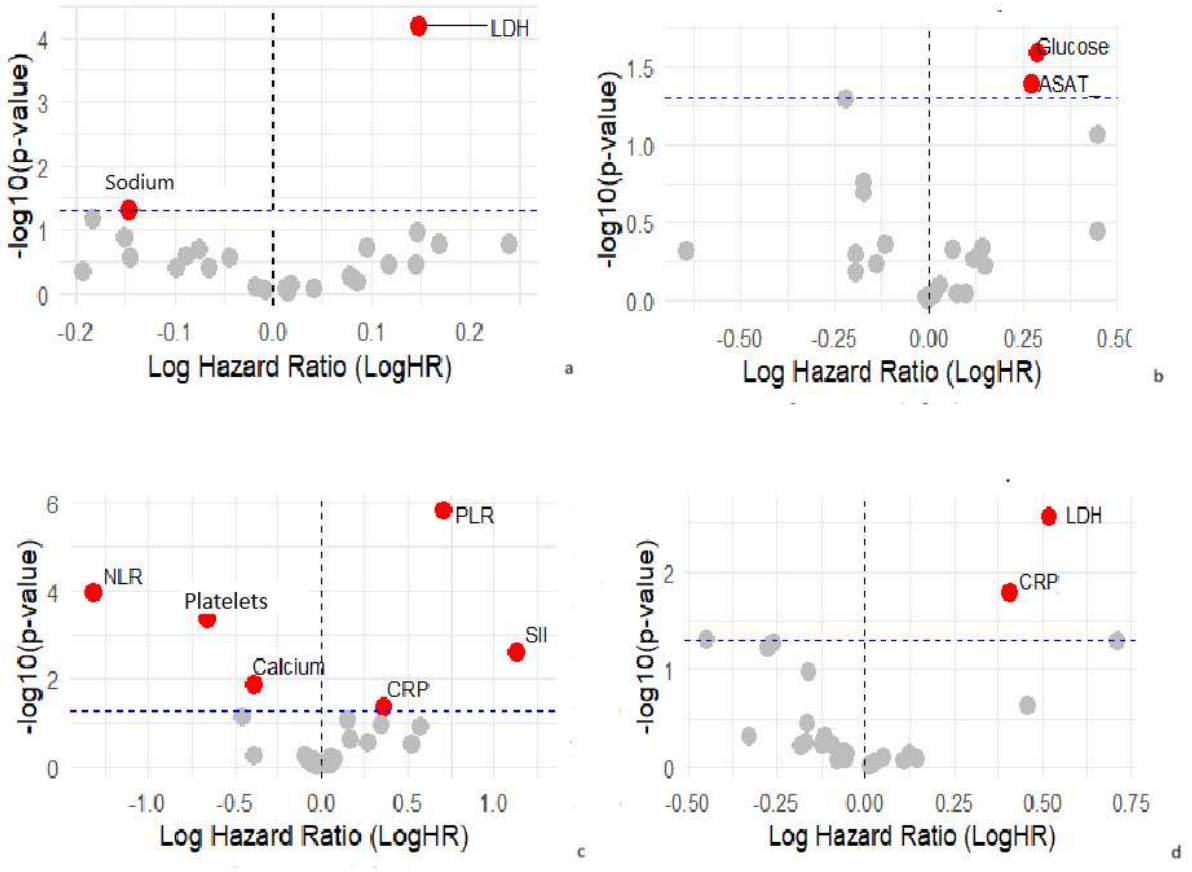


Figure 5: Vulcano plots from the cox regression models with a) all three time points, b) time point 0, c) timepoint 1, d) time point 3. .

4.4 Discussion

This study examined a range of clinical parameters to evaluate their prognostic significance in relation to OS and PFS of patients. A broad spectrum of biomarkers: LDH, CRP, MCV, glucose, ASAT, platelets, calcium, NLR, PLR, SII, TSH, sodium, FT4, leukocytes, neutrophils, were found to be significant predictors at different time points for OS. Age and PD-L1 status showed significance in OS and PFS, respectively. More biomarkers were significant after treatment initiation than before. CRP emerged as a recurrent indicator associated with OS. Ratios such as NLR, PLR, and SII were also significant at more than one timepoint. For PFS after correction for multiple testing, PD-L1 was significant, no blood value remained significant. Multinomial regression further emphasized the critical roles of specific markers. PD-L1 positivity strongly correlated with the favourable outcomes CR and mixed response, but not with stable disease. Smoking history and ECOG score, also revealed patterns. Smoking at diagnosis had a strong negative association with CR and SD, while a history of past smoking was positively associated with CR, suggesting a complex relationship between smoking status and treatment outcomes.

One of the most consistent findings was the significance of CRP in predicting OS across multiple time points. CRP is a well-known marker of systemic inflammation [72] and has been implicated in cancer progression and poor outcomes [73]. Elevated CRP levels may reflect an inflammatory tumour microenvironment, which can foster tumour growth and suppress immune responses. Interestingly, it shows up significant consistently in both tests at 1 month or 3 months after initiation of therapy. This may indicate that immunotherapy-related inflammation or immune activation contributes to patient outcomes, highlighting the dynamic relationship between treatment and systemic inflammation. The immune and inflammatory ratios, including NLR, PLR, and SII, also emerge as potential predictors at more than one time-point for OS. These ratios capture the balance between immune suppression and immune surveillance. A lower ratio indicates a more active immune system, which could potentially be because of less immune suppression by the tumour. This aligns with the therapeutic goal of immunotherapy ICI, which targets immune system activation.

An important observation was the shift in biomarker significance before and after treatment initiation. Prior to therapy, CRP was significant for OS, likely reflecting baseline tumour burden. Post-treatment, a broader array of markers gained significance, which could indicate the effects of therapy itself or indicate progression of the disease. This temporal change could suggest that therapy induces systemic changes that influence survival. An observed shift was the significance of calcium levels detected only at the one-month post-therapy time point. There are multiple reasons for calcium to be high but a potential explanation for this phenomenon in being predictive for OS is tumour lysis syndrome (TLS) [74], a metabolic complication resulting from the rapid and extensive breakdown of tumour cells. This process releases intracellular components, such as electrolytes and nucleic acids, into the bloodstream. This disrupts homeostasis, which can lead to complications, like acute kidney injury, cardiac arrhythmias, seizures, or sometimes death. Another possible reason for calcium to be significant is that calcium can also indicate bone metastases, only it would be expected to show up significant at more than the time point one month after start of therapy.

The predictors for PFS showed notable variability across time points. Markers such as sodium, neutrophils, and creatinine were significant, particularly at earlier time points, suggesting they might reflect acute physiological changes in response to therapy. The absence of significant biomarkers at the 3-month mark may indicate that PFS reflects an interplay of factors, including tumour biology and response to treatment, which are harder to encapsulate through static blood values. Which is reinforced by the fact that after correction, nothing remains significant.

PD-L1 status, a known predictive biomarker [15, 22, 33], is significant for CR but shows limited association with PR and SD. It correlates with PFS but not OS. Interestingly, when PD-L1 is assessed cytologically, it becomes significant for OS, despite cytology typically being less accurate than pathology. Cytologically acquired samples are likely influenced more by immune cell infiltration and tumour heterogeneity [34]. This inconsistency highlights the complexity of PD-L1's role, but it remains the most reliable clinical predictor of treatment outcomes.

Lifestyle and clinical factors, such as smoking history and ECOG score, added prognostic information. Active smoking at diagnosis was strongly associated with worse outcomes like lower CR and SD rates, possibly due to smoking-induced systemic inflammation, immune suppression, and enhanced tumour progression [75]. Interestingly, a history of past smoking showed a positive association with CR, potentially reflecting the higher mutational burden seen in smoking-related cancers, which might increase responsiveness to immunotherapy [21]. The ECOG score's strong negative correlation with favourable outcomes shows the important role of baseline functional status in determining prognosis, showing its relevance alongside blood-based biomarkers.

These findings demonstrate the complex nature of prognostic biomarkers in immunotherapy. The prominence of inflammatory markers and immune ratios highlights the crucial role of systemic immune regulation in shaping patient outcomes. The temporal variability in biomarker significance suggests that dynamic monitoring could provide an option to predict response, offering potential for treatment adjustments and more accurate long-term predictions.

4.4.1 Comparison with Existing Literature

The findings from this study provide new insights into the role of blood-based biomarkers in predicting the response to immunotherapy in NSCLC patients. While previous research has pointed to specific biomarkers like CRP [76], eosinophils, NLR, PLR, and SII [4, 35, 36] as potentially predictive markers, our results were mixed for timepoints and did not fully support all the previous findings, but do show overlap.

Eosinophils: Prior studies have suggested a role of eosinophils in predicting response to immunotherapy for melanoma and colorectal cancer [25, 37, 38]. In NSCLC, eosinophilia has been associated with improved outcomes in some cases, but the findings remain inconsistent [39]. This study found that eosinophil counts were not significant predictors of therapy response in NSCLC patients, even though their predictive value has been recognized in other cancers.

Blood-Based Markers: This study analysed a range of blood-based markers and revealed significant associations for some, but not all. Specifically, markers like CRP, and NLR, PLR, SII were found to be predictive of OS. For classic markers such as NLR, PLR, and SII, which have been previously highlighted in the literature as predictive of immunotherapy response [4, 35, 36], this is consistent with our findings. Only these studies mainly find them indicative before therapy, where this study does not show any predictive values at this point. In the

mentioned studies the blood markers that were significant, show significance for response in no progression. In this area none show up significant in this study.

The differences between our results and prior studies may stem from several factors. First, patient population heterogeneity, along with variations in the timing and methods of blood sample collection, likely contributed to these discrepancies. Second, the retrospective design of this study limits control over confounding variables, such as concurrent medications, treatment regimens, and clinical trajectories, all of which could affect biomarker levels.

4.4.2 Limitations

This study had several limitations that should be considered when interpreting the findings. First, the retrospective design introduced potential biases, including variability in the timing of blood sample collection and the influence of unmeasured confounders such as therapy dosage, combination treatments, or comorbidities. Second, the analysis relied on data from a single cohort at a specific medical centre, limiting the generalizability of the results to broader populations. The relatively small sample size also reduced the statistical power to detect significant associations for some markers, and missing data posed a significant challenge. To address missing values, excluding patients entirely was avoided, but this made it difficult to compare groups across different time points, since the groups often included different individuals. This variability complicated efforts to draw definitive conclusions about treatment response. Moreover, interpreting blood marker levels was challenging due to variability unrelated to treatment response, such as the presence of comorbidities, infections, or other factors. For instance, some patients who died early may have responded to immunotherapy but succumbed to unrelated causes, such as infections, before disease progression could be checked or confirmed. Similarly, radiological evaluations were not always definitive. Early scans might show tumour growth due to pre-treatment progression rather than a lack of response, leading to potential misclassification. Lastly, differences in analytical methods, such as Kaplan-Meier survival analysis and Cox proportional hazards regression, posed interpretive challenges. Kaplan-Meier provides descriptive survival estimates without adjusting for covariates, whereas Cox models account for covariates but assume proportional hazards over time. These approaches sometimes gave conflicting results for the same covariates, complicating predictor selection. While both methods are valuable, careful integration of their insights is necessary to ensure robust conclusions. Future studies should address these limitations by using larger, more diverse cohorts, minimizing missing data, incorporating additional confounders, and employing longitudinal designs to enhance causal inference and improve the reliability of blood biomarkers as predictive tools for immunotherapy response.

4.4.3 Future Directions

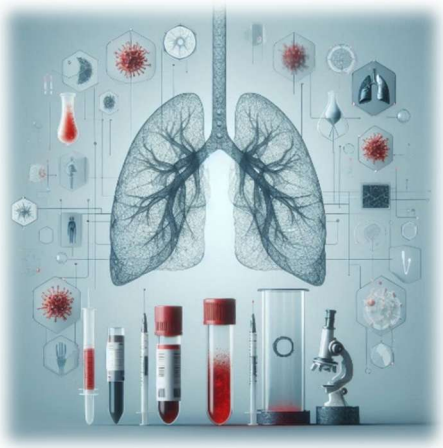
These findings highlight the complexity of predicting treatment outcomes in NSCLC immunotherapy and emphasize the need for more sophisticated analytical approaches. Although these findings can help develop more advanced models to predict treatment response, incorporating the variables identified in this chapter as key inputs to enhance model accuracy and clinical applicability.

Future research should explore optimal timing for biomarker assessments in NSCLC immunotherapy, as markers showed more significance at one- followed by the three-month intervals than at baseline, likely reflecting the patient's evolving biological response to treatment. To address study limitations, prospective designs are a potential solution to minimize missing data and heterogeneity in treatment course. Adjusting for therapy variations (e.g., monotherapy vs. combination therapy) and validating findings in larger, multi-centre cohorts will enhance reliability and generalizability. Incorporating additional data, such as CT imaging and genomic profiling [76, 77], could improve insights into tumour progression and treatment efficacy. Longitudinal studies tracking dynamic blood marker changes could provide stronger evidence for their prognostic value. Additionally, integrating multiple biomarker types may further refine predictive models, supporting personalized immunotherapy strategies. Distinct predictive profiles for OS and PFS underscore the need for improved evaluation measures. OS can be influenced by factors beyond disease progression, including unrelated causes of death. PFS, however, poses unique challenges due to its dependency on accurate detection of tumour progression. Progression might remain undetected until after death, or patients may survive despite recorded progression in the first weeks. While iRECIST tries to account for this by requiring confirmation of progressive disease (PD) after a set period, there can still be a significant time gap before confirmation. As a

result, treatment may be discontinued prematurely, or cases of pseudoprogression can complicate decision-making, making early PFS assessments challenging. Refining definitions and outcome measures is crucial for accurate prognostic evaluations. Finally, studying the biological mechanisms through which blood markers affect immunotherapy response could deepen understanding of NSCLC pathology and reveal new therapeutic targets.

4.5 Conclusion

In conclusion, this study highlights the prognostic significance of a wide range of clinical and biochemical parameters in predicting outcomes such as OS and PFS in patients undergoing immunotherapy. The findings underscore the additional value of using a combination of several markers and also reveal possible patterns regarding their temporal significance and their relationship with different outcomes.



CHAPTER 5:
MIXED EFFECT MODELS

5.1 Introduction

This chapter aims to complement the previous chapter and answer the question: which patient will respond favourably to immunotherapy. Various biomarkers, including blood values and clinical characteristics may influence treatment outcomes [21, 41], as seen in the previous chapter but integrating these diverse data sources into a unified predictive model has shown to be difficult. MEMs are suited for analysing these complex data structures, particularly when repeated measures are taken over time, and for accounting for both fixed effects (such as blood values and patient characteristics) and random effects (such as inter-patient variability) [50-52]. In addition to MEMs, linear models have been explored to investigate relationships between specific groups of variables and treatment outcomes. This preliminary analysis serves several purposes: it allows for an assessment of simpler, more interpretable models; it facilitates group-level exploration of related predictors (e.g., infection parameters, inflammation markers) to identify trends or potential interactions; and it provides a complementary approach to validate findings and refine variable selection for a future more complex MEM framework. This chapter details the development part of a MEM aimed at predicting NSCLC patient response to immunotherapy by incorporating routinely collected clinical and imaging data. By using this model, the ultimate goal is to create a personalized prediction tool for treatment outcomes which can assist in clinical decision-making in the future.

5.2 Method

To analyse the association between blood biomarkers and survival outcomes, a Generalized Linear Mixed Effects Model (GLMM) was made, using R(version 2024.04.2) as well as a Linear Mixed effects model (LMM). Survival time was converted to surviving past 365 days, as a binary outcome for the GLMM and as continuous days for the LMM. The general dataset was used in which values of before, 1 month after and 3 months after start of treatment were included. All blood biomarkers were standardized to a mean of 0 and a standard deviation of 1 to ensure comparability across variables. Polynomial transformations (quadratic and cubic terms) were created for the biomarkers to capture potential other non-linear relationships. The dataset was divided in a training, validation and test set (0.7:0.15:0.15). A GLMM was fitted on the training set with survival as the (binary) dependent variable and all standardized blood biomarkers and clinical values, including their transformed terms, as fixed effects. The optimal threshold was determined on the validation set. Multicollinearity was assessed using Variance Inflation Factor (VIF) analysis. Biomarkers with high VIF values were removed, which ensured that the remaining predictors were independent and contributed unique information to the model. After this, new models were made with a subset of the blood values. Models were also built with blood values selected based on the prior analyses (Chapter 4), biomarkers significantly associated with OS were isolated and included in the MEM. This was done including all three time points but also only including the timepoints after initiation of therapy. These models were compared to the full MEMs using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) to assess the complexity of the models. Area under the curve (AUC), negative predictive value (NPV), sensitivity and specificity was determined for evaluating the function of the model using the test set. Lasso regression was attempted to refine the set of predictors and identify important biomarkers for survival.

To complement to building of the MEM, Linear models were employed to explore group-level effects of related biomarkers. Groups of blood values were formed based on their biological functions or roles in immune response:

- Inflammatory / immune markers: CRP, NLR, PLR, LMR, SII
- Liver function markers: Bilirubin, ALAT, ASAT
- Kidney function markers: Creatinine, eGFR (CKD-EPI)
- Electrolytes: sodium, potassium, calcium
- Haematological markers: haemoglobin, platelets, leukocytes, lymphocytes, neutrophils, monocytes, eosinophils
- White blood cels: lymphocytes, monocytes, eosinophils, neutrophils
- Endocrine and metabolic markers: TSH, FT4, glucose

Linear models were used to investigate the relationships between these biomarker groups and survival outcomes. This step helped identify potential trends and interactions that could inform subsequent MEM development.

5.3 Results

The initial GLMMs and LMMs were fitted using all three time points and included all blood values along with their quadratic and cubic transformations. However, these models faced significant convergence issues due to the large number of variables relative to the dataset size, resulting in a degenerate Hessian matrix and failure of VIF computations. After removing quadratic and cubic terms, similar numerical issues persisted. After further refinement as mentioned in the methodology was performed some models no longer showed these issues. Table 8 shows the data used for every model, all specific blood values included in the different models can be found in appendix F. The results of all GLMM models can be found in table 9. Models 1, 1.1, 2, 6 and 6.1 failed to converge. In model 2 and 6 all included blood values were significant. Model 3, 4, 4.1, 5 and 5.1 did not fail to converge. Model 3 contained solely CRP at timepoints one and three based on its relevance found in the previous chapter. CRP showed significancy. In model 4 CRP, platelets and PLR were significant. In model 4.1 blood values with a high VIF in model 4 were deleted. In which CRP remained significant. In model 5 CRP showed a p value below 0.05. Model 5.1 was adjusted for VIF in which CRP remained significant.

Table 8: Variables per model.

Model	Time points	Variables included
1	0, 1, 3	All possible blood values and transformations
1.1	0, 1, 3	Reduced set of model 1 based on VIF
2	0, 1, 3	All blood values without transformations
3	1, 3	CRP only
4	1, 3	All inflammatory/immune markers based on chapter 4
4.1	1, 3	Reduced set of model 4 based on VIF
5	1, 3	Ratios, CRP and calcium, based on chapter 4
5.1	1, 3	Reduced set of model 5 based on VIF
6	1, 3	Same as model 2
6.1	1, 3	Reduced set of model 6 based on VIF

Lasso regression did not provide lower AIC, BIC, or higher AUC, sensitivity, specificity.

The full results of the linear model can be found in appendix E. All residual standard errors ranged between 0.467 and 0.5013. Linear models that were statistically significant were: inflammatory, electrolytes and haematology. Blood values that were statistically significant were, CRP, Sodium, HB and glucose.

Table 9: Results for the different MEMs.

	AIC	BIC	logLik	AUC	NPV	Sensitivity	Specificity
Model 1	332.6	620	-88.3	0.71	0.47	0.6	0.60
Model 1.1	276	364.4	-114	0.82	0.55	0.55	1.00
Model 2	249.1	352.3	-96.6	0.82	0.63	0.7	0.91
Model 3	150.6	163.6	-71.3	0.74	0.75	0.88	0.67
Model 4	157.1	195.8	-66.6	0.66	0.57	0.82	0.44
Model 4.1	154.3	173.7	-71.2	0.77	0.67	0.82	0.67
Model 5	156.8	185.9	-69.4	0.73	0.6	0.76	0.67
Model 5.1	187	211.3	-86.5	0.71	0.75	0.88	0.67
Model 6	156.5	246.8	-50.2	0.59	0.45	0.65	0.56
Model 6.1	153.3	237.1	-50.6	0.63	0.38	0.53	0.56

5.4 Discussion

The development of a predictive model to assess the response of NSCLC patients to immunotherapy is complex. This study aimed to use MEMs to integrate blood-based data to advance personalized treatment predictions for outcomes such as OS. While MEMs have demonstrated their use in other fields, the application of this framework in this context showed limitations. These issues show both the potential and the complex challenges of employing a MEM in this situation.

The primary obstacle in developing a MEM was convergence failure, which can be caused by several factors, mainly overfitting, multicollinearity, and insufficient data. Initial attempts to include all biomarkers across three time points with polynomial transformations resulted in degenerate Hessian matrices and high variance inflation factors. This reflected numerical instability, likely due to the excessive complexity of the model relative to the dataset size. Despite removing quadratic and cubic transformations, filtering out collinear variables, and testing reduced sets of biomarkers, many models continued to show convergence problems. The performance of simpler models (e.g., those including only CRP or CRP and LMR) demonstrated no convergence failures, lower AIC and BIC values and also some higher sensitivity and NPV. However, these gains also came with the cost of lower AUC and specificity.

Model 1.1 showed a high specificity which ensures that identified responders are accurate. Only at the cost of identifying many responders as non-responders. This would not provide any clinical benefit, as therapy cannot be withheld from the identified non-responders, as there is a big change they could be a responder. And a Non-converging model, it raises concerns about the reliability and interpretability of the result. From the remaining models that did converge, model 5.1 shows the highest sensitivity and NPV, although it has the lowest AUC. With the found NPV of 0.75 on the test set, this still results in a one in four patients being predicted as non-responder when they are a responder, limiting clinical use of the predictive model. The other models (models 3, 4, 4.1 and 5) trained on the found values in chapter 4 show their predictive power remains discussable because of their lower sensitivity, specificity and NPV compared to model 5.1. The majority of blood values used in all these models were not significant, further limiting their ability to provide meaningful insights into survival outcomes.

CRP repeatedly emerged as the significant predictor. While CRP is a biologically plausible biomarker, relying on a single biomarker possibly limits the robustness of the model as other (not converging models) showed higher AUCs. The trade-off between model complexity and interpretability was evident throughout this study. Reducing the number of predictors lowered AIC and BIC values as expected and also solved convergence issues, but it also led to a decrease in AUCs. While reducing predictors can stabilize a model, it may inadvertently exclude variables that hold subtle but important predictive value.

Linear models provided complementary insights and were used to explore group-level effects. These models identified significant associations between survival and inflammatory markers, electrolytes, and hematologic parameters. Statistically significant predictors included CRP, sodium, haemoglobin, and glucose, as previously noted in Chapter 4. However, the residual standard errors (ranging from 0.467 to 0.5013) showed their limited predictive power for individual outcomes (e.g., binary classification of survival). These results illustrate the trade-off between interpretability and complexity: simpler models offer clearer insights but lack the granularity needed for personalized predictions.

5.4.1 Comparison with literature

Developing these models aligns with the growing interest in personalized medicine and the use of explainable statistical models to predict treatment response [78-80]. MEMs have been successfully employed in other areas of medical research, particularly when longitudinal, hierarchical, or repeated measures data are available. Examples of researches where MEMs where employed in predicting effect are; Anti-VEGF Therapy in Ophthalmology [79], Psychiatry and Psycholinguistics [78, 80-82], and Language Testing [50] as well as analysis of tumour size dynamics in clinical settings [83]. In contrast to those studies, the dataset used in this NSCLC study has several limitations that impact the performance of MEMs. One of the challenges can be the small sample size, as the previous studies included for example 793 [82], up until 3398 [83], and 193 with 13 datapoints a patient [79], where in this research around 150 patients were included with 2 or 3 datapoints per patient. This reduces the statistical power of the model and limits the ability to detect patterns or relationships within all the different blood values tested. MEMs require a sufficient number of observations to estimate both fixed and random effects accurately. With fewer data points, the models are more prone to overfitting, or failing to converge, as there was insufficient information to reliably estimate parameters, which is one of the things most likely to have happened here.

Additionally, the dataset has missing or infrequently collected measurements. This creates gaps in the data structure, which does not help the MEMs' ability to effectively capture longitudinal trends. MEMs are suited to repeated measures and longitudinal data, but when data points are missing, their power to track changes over time or make reliable predictions is compromised. In this line, MEMs need data that is collected at regular intervals, comparable points, providing sufficient information to describe longitudinal trends. In this study the time points can vary about 6 weeks, which can affect the blood value measurement a lot. A measurement two days after (chemo)immunotherapy can provide different values compared to measured measurement just before the next therapy.

Another possible issue is noise in the data, partially caused by variables that are not included in the effects. High noise levels distract from the true signals in the dataset, making it difficult for the model to differentiate between meaningful trends and random fluctuations. Adding patient-specific random effects caused challenges as well. These random effects are meant to account for individual differences, like varying baseline biomarker levels or unique responses to treatment and are important for making personalized predictions. However, in this study, while the random effects worked well for the training data, the models had trouble making accurate predictions for new, unseen patients. This problem is especially relevant in clinical settings, where models need to reliably predict outcomes for patients who were not part of the training data. Hence extending the dataset is desirable to limit overfitting.

5.4.2 Future directions

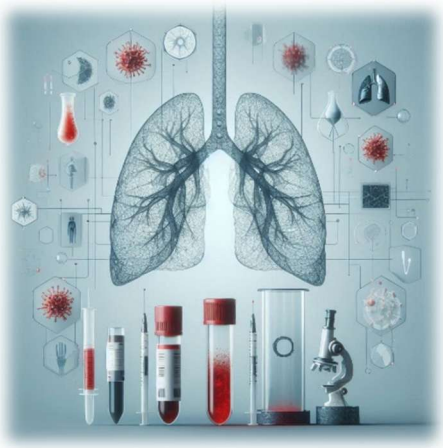
The MEMs in other fields, such as ophthalmology, psychiatry, and language testing as aforementioned, show the importance of data characteristics and study design in achieving reliable outcomes. A primary reason for this is the availability of more and high-quality data. Large datasets with frequent and consistent measurements enable accurate estimation of model parameters, which reduces the likelihood of overfitting and enhances the reliability of the models and can better filter out noise. These datasets often show low variability and high repeatability, which is important to identify consistent trends within the data. What also comes into play is the focused scope of the studies. They target specific outcomes with well-defined predictors (found as predictors before), researchers are able to simplify the models, reduce complexity, and minimize instability. This focus makes sure that the models remain stable and predictive while maintaining a manageable level of detail. This provides several changes for future studies.

It would be beneficial to expand the dataset by including more patients and collecting data at additional time points. Increasing the size and scope of the dataset can help filling missing values, providing accurate measure points and with the addition of timepoints it is less likely to miss fluctuations in the data. Which could allow for the model to capture more meaningful patterns and interactions. However, care must be taken to ensure that the added complexity of a larger dataset does not lead to overfitting. Balancing the amount of data with the complexity of the model is essential for achieving reliable results. More standardization of the data collection also belongs to this suggestion, collecting data at consistent intervals and ensuring frequent measurements.

Finally, conducting a more thorough and extensive exploratory data analysis, along with careful specification of the MEM structures, is crucial for future studies. As demonstrated in this MEM development, the model based on the findings from Chapter 4 shows better accuracy, precision, sensitivity, and specificity compared to other models, leading to improved performance. This shows the importance of identifying the correct random and fixed effects early in the process to create a well-functioning MEM, as these values were not found in the building of the MEM itself. Therefore, a recommendation is to do more extensive research to determine which values belong to fixed and random effects and understanding their influence beforehand. This can offer valuable guidance in building a robust MEM that fits the data more effectively.

5.5 Conclusion

This chapter shows the potential and challenges of using MEMs to predict NSCLC patient response to immunotherapy, with convergence failure issues, multicollinearity, and insufficient data. While the preliminary models demonstrated some predictive capabilities, particularly with CRP as a significant biomarker, more exploratory data analysis is needed to define better models.



CHAPTER 6:
RANDOM FOREST

6. 1 Introduction in the random forest

When MEMs fail to produce reliable predictions, alternative methods can be explored to identify meaningful variables and improve predictive performance. One such approach is the use of Random Forests (RF), a supervised learning method that has a robustness in handling high-dimensional datasets, ability to integrate multivariate biomarkers, and more resilience to overfitting than MEM. RF is a learning method that constructs multiple decision trees during training and combines their outputs [30, 84]. For classification tasks, RF outputs the mode of the individual tree predictions, while for regression tasks, it calculates the mean of the outputs [85, 86]. By combining the results of multiple trees, RF reduces the risk of overfitting and is robust against noise in the data. This makes it particularly suited for complex datasets with multivariate biomarkers [30, 84] In contrast with for example SVM, that can struggle with high-dimensional feature spaces or require extensive tuning to avoid overfitting, particularly when the data is noisy, has a lot of missing values or when there are many irrelevant features. In the context of this thesis, RF can serve two purposes: First, RF can help identify important variables, such as blood or clinical biomarkers, that significantly influence outcomes, as indicated by a metrics like Gini and permutation importance. These variables can then be used to refine MEMs. Second, RF itself is a robust predictive tool that may outperform MEMs when applied to the prediction of OS, both as a binary outcome and as a continuous variable. However, while RF can achieve strong predictive performance, a limitation is its lack of interpretability. Unlike MEMs, that provide insight into how specific variables contribute to the model, RF operates as a "black box" making it difficult to understand why a patient is classified as a responder or non-responder. Despite this drawback, RF can still be a valuable tool for predicting outcomes, particularly when traditional methods like MEMs struggle to deliver meaningful results. In this chapter, the application of RF is explored to predict OS and identify potential biomarker sets. By using RF's ability to handle complex, high-dimensional data, this approach also aims to improve our understanding of which variables are most relevant to treatment outcomes and improve predictive performance in clinical decision-making.

6.2 Method

In this chapter three RF models were trained for survival prediction, progression prediction, and survival time regression. All datasets were loaded and processed using Python (version 2024.1.3). The data included patient characteristics, clinical outcomes, and longitudinal blood test values. Records were filtered to include only the first three intervals of blood testing. The blood values were set to a wide format and features such as the average and differences between the points were derived. The outcome variables varied by task. For survival prediction, "Time to death" was binarized to indicate whether a patient survived beyond 365 days but also retained as a continuous target for survival regression. For progression prediction, the presence of disease progression was used. The dataset was split into training, validation and testing sets (70:15:15), to determine hyperparameters and the optimal threshold. Hyperparameter optimization was performed using RandomizedSearchCV, testing 1000 combinations with 3-fold cross-validation. The best-performing Random Forest model was selected based on these results and retrained. The optimal decision threshold was determined using the Youden index from the ROC curve on the validation set, maximizing sensitivity and specificity. Performance was evaluated on the test set using AUC, NPV, sensitivity and specificity. Similar steps

were applied for progression prediction. The last training model applied was a random forest regressor to predict the continuous outcome of survival time. After splitting the dataset as mentioned before and trying different parameters, its performance was evaluated using the root mean squared error (RMSE). For all models, feature importance was extracted from the trained Random Forest models using Gini and permutation importance. This analysis quantified the contribution of each feature to the predictive tasks, ranking features to identify the most influential predictors.

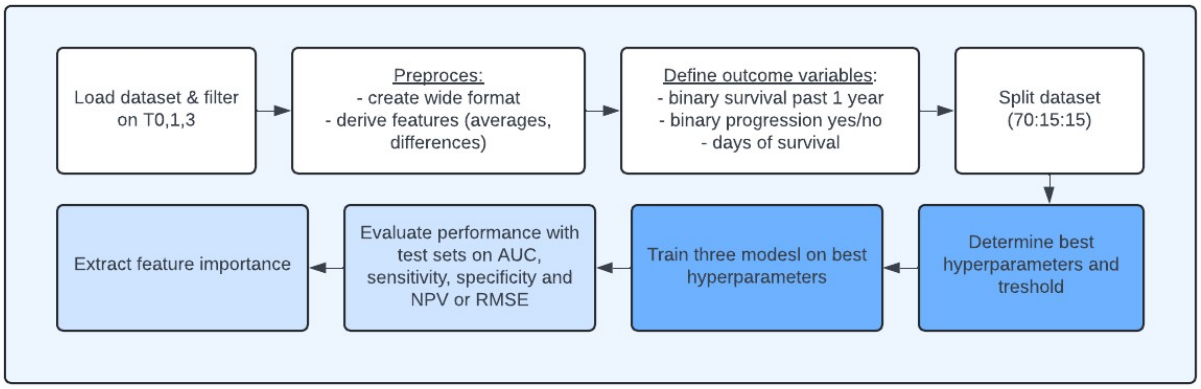


Figure 8: Flow chart of creation of the RF models, white is the preprocessing, dark blue the training of the model, and light blue the evaluation of the trained models.

6.3 Results

The best parameter settings for the models predicting progression can be found in Table 10. Table 11 presents the performance metrics of these models. The T1 model achieved the highest sensitivity (0.86), while the T0,3 model showed the highest specificity (1.00). The highest AUC was 0.71 for the T1 model.

Table 10: The found best parameters for the different inclusions of timepoints in PFS.

	T0,1,3	**T0**	**T1**	**T3**	**T0,3**	**T1,3**	**T0,1**
N estimators	100	50	500	100	200	100	200
Min sample split	5	5	2	10	2	2	2
Min sample leaf	1	1	4	4	2	2	1
Max features	None	Log2	Log2	Sqrt	Log2	Log2	Sqrt
Max depth	None	None	None	None	None	None	None

Table 11: The results of the models on the found best parameter settings for the different timepoints for PFS.

	T0,1,3	**T0**	**T1**	**T3**	**T0,3**	**T1,3**	**T0,1**
Sensitivity	0.24	0.72	0.86	1.00	0.00	0.73	0.38
Specificity	0.88	0.29	0.12	0.10	1.00	0.14	0.57
AUC	0.66	0.49	0.71	0.66	0.53	0.51	0.54
NPV	0.27	0.22	0.25	1.00	0.27	0.14	0.20

For binary OS prediction, the best parameters are shown in table 12, with model results shown in table 13. The T1,3 model had the highest sensitivity (0.88), while the T0 model had the highest specificity (0.97). The highest AUC was 0.86 for the T3 model. The highest NPV was model T0,1.

Table 12: The found best parameters for the different inclusions of timepoints in binary OS.

	T0,1,3	T0	T1	T3	T0,3	T1,3	T0,1
N estimators	50	50	200	200	50	100	50
Min sample split	10	10	10	10	10	10	5
Min sample leaf	2	2	4	4	4	1	1
Max features	Log2	Sqrt	Sqrt	Sqrt	Log2	Sqrt	Sqrt
Max depth	None	none	None	None	None	None	None

Table 13: The results of the models on the found best parameter settings for the different timepoints for binary OS.

	T0,1,3	T0	T1	T3	T0,3	T1,3	T0,1
Sensitivity	0.82	0.07	0.87	0.61	0.35	0.88	0.82
Specificity	0.67	0.97	0.43	0.80	0.83	0.33	0.47
AUC	0.72	0.53	0.77	0.86	0.69	0.64	0.73
NPV	0.57	0.53	0.75	0.36	0.31	0.50	0.80

For continuous OS prediction, the optimal hyperparameters are listed in Table 14, and Table 15 reports the RMSE values. The T0,1,3 model had the lowest RMSE (332.45), while the T3 model had the highest RMSE (515.31).

Table 14: The found best parameters for the different inclusions of timepoints in continuous OS.

	T0,1,3	T0	T1	T3	T0,3	T1,3	T0,1
N estimators	500	200	100	200	100	100	200
Min sample split	10	10	2	10	2	5	10
Min sample leaf	4	4	2	4	4	1	4
Max features	None	Log2	Log2	Log2	Log2	Log2	None
Max depth	None	None	None	None	None	None	10

Table 15: The RMSE of the models on the found best parameter settings for the different timepoints for continuous OS in days.

	T0,1,3	T0	T1	T3	T0,3	T1,3	T0,1
RMSE	332.45	374.47	412.95	515.31	348.49	489.12	390.62

Lastly, Table 16 displays the top 5 feature importances for each model, showing the variables with the greatest impact on prediction across different outcomes, such as progression, binary survival, and continuous survival.

The highest feature importance was 0.117 for CRP at T3 in binary survival prediction. Other features ranked highly across different models include leukocytes at T3, NLR at T3, and creatinine at T1

Table 16: The feature importance per model.

	Model Progression binary Gini	Model Progression Binary Permutation	Model Survival Binary Gini	Model Survival Binary Permutation	Model Survival Continuous Gini	Model Survival Continuous Permutation
1	Creatinine t1: 0.044	ECOG: 0.028	CRP t3: 0.117	Age: 0.000	Neutrophils t3: 0.037	CRP t3: 0.042
2	DLCO/VA: 0.042	FT4 t1: 0.024	Leukocytes t3: 0.073	Oligometastasis: 0.000	CRP t3: 0.029	Neutrophils t3: 0.028
3	Age: 0.041	PD-L1: 0.024	NLR t3: 0.068	Total bilirubin t3: 0.000	Calcium t3: 0.026	Creatinine t3: 0.017
4	PD-L1: 0.039	Sodium t1: 0.021	SII t3: 0.065	ALAT t3: 0.000	creatinine t3: 0.025	FT4 t1: 0.014
5	NLR t1: 0.036	DLCO: 0.017	LDH t3: 0.051	Tumour type: 0.000	SII t3: 0.022	ASAT t3: 0.013

6.4 Discussion

In this chapter, the application of RF algorithms was investigated to predict patient responses to immunotherapy, focusing on OS and PFS as outcomes. The highest AUC score was achieved in T3 for OS binary model, after which the best performance metrices were found by integrating data from all three time points (T0, T1, T3) or the timepoints after initiation of therapy (T1 or T0+1). Continuous outcome prediction, however, showed limited potential, as the RMSE exceeded one year. Among the predictors, CRP consistently emerged as the most influential feature. However, both Gini importance and permutation importance scores were low across all features. The findings imply that the predictive strength of RF lies in the combined contributions of multiple features.

6.4.1 Clinical implications

The feature importance analysis highlighted CRP [76] and ratios as significant predictors across multiple models. These biomarkers are indicators of systemic inflammation and immune response, making their prominence clinically meaningful, as discussed in chapter 4. However, the relatively low and closely spaced Gini and permutation importance scores suggest that no single feature alone is strongly predictive. This can occur in datasets where relationships between features and the target variable are weak or where many features are redundant or correlated. This aligns with the understanding that immunotherapy response is likely influenced by a combination of factors and cannot be captured using solely one or two blood values. This reinforces the need of using an algorithm like a RF for capturing complex multivariate interactions. Gini importance scores ranked features by their predictive contributions in this and other studies. For instance, breast cancer research [87], tuberculosis treatment failure [88] and a study on brain metastases in NSCLC [89] used Gini importance to identify important predictors. In a radiomics study on early-stage ground-glass opacity pulmonary adenocarcinoma, top feature importances were around 0.25 [90], whereas the top score found in this study only reached till 0.117 for 1 model and all the others reached no higher than 0.044. Feature importances vary across models and domains as well as their exact use. Studies can find high scores due to rich data structures, while a biomarker-based models, like this research, can show lower scores due to redundancy or weaker direct associations. Class imbalance influences results, as standard permutation struggle to differentiate associated predictors under imbalance [91]. These factors could explain differences in feature importances across studies.

The binary models trained on data from multiple time points and timepoints after start of therapy demonstrate the value of longitudinal data in capturing dynamic changes in biomarker levels, even though the Random Forest model itself does not explicitly account for a longitudinal data structure. Among these, the model incorporating data from all three time points (T0, T1, T3) achieved a solid sensitivity (82%) and AUC (0.72). This underscores the benefit of comprehensive longitudinal data in capturing critical patterns, however

its NPV (57%) and specificity (67%) suggest some limitations in correctly identifying non-responders. The T1 model shows the highest sensitivity (87%) and an AUC (0.77) and a relatively high NPV (0.75), demonstrating its use in identifying non-responders from a single time point. Its relatively low specificity (43%), however, implies a higher risk of misclassifying non-responders as responders. The T3 model achieved a notable balance between sensitivity (61%), specificity (80%), and the highest AUC (0.86). Although the high AUC suggests that T3 captures critical predictive information effectively, its NPV (36%) is lower compared to other models, therefore limiting its clinical use. The T0,1 model also performed with sensitivity (82%) equal to the T0,1,3 model and an AUC (0.73). Importantly, its NPV (80%) was the highest among all models. This indicates that when the model predicts a patient as a non-responder, it is correct more often. Although it might misclassify non-responders as responders more often, in a clinical context the high NPV is beneficial. Overtreatment poses risk of unnecessary side effects and increases healthcare costs. While misidentifying a responder as a non-responder would result in withholding a potentially effective therapy. This approach ensures that the patients that are excluded from therapy are all unlikely to benefit from therapy, optimizing resource use and sparing patients from unnecessary side effect, but assuring that all patients that might benefit from the therapy will get treated.

The regression task for predicting survival time showed limited utility, with RMSE values exceeding 350 days, far beyond a clinically meaningful threshold. This highlights the need for more sophisticated modelling techniques or richer datasets to improve continuous outcome predictions.

6.4.2 Limitations

While binary RF models demonstrate notable predictive capabilities, their clinical application is limited by a lack of interpretability. Unlike statistical models like MEMs, which provide clear coefficients indicating how variables influence outcomes, RF operates as a "black box." This makes it difficult to understand the relationships between predictors and clinical outcomes, complicating efforts to trust, validate, or justify treatment decisions based on RF predictions. The integration of AI model into clinical practice is therefore researched [92], where the biggest issues identified was adapting AI models to real-world situations, with incomplete or incorrect data in practice. However, creating and evaluating explanations of AI models can provide deeper insights into both the models themselves and the underlying subject matter. Explainable AI techniques can help identify key factors influencing the model's decisions, making it easier to test and refine hypotheses [93]. By implementing such techniques, not only can the transparency of RF models be improved, but also the trust of healthcare professionals in AI-driven decisions can be strengthened, as almost a forth of healthcare professionals in AI and cancer point out ethical problems and who is legally responsible for decisions made with the help of AI? [92]. These ethical and legal concerns add to the complexity. Can a model that lacks transparency be relied upon for potentially life-altering decisions in clinical settings? Without interpretability, there is a risk of implicit bias, making it essential to complement RF models with explainability tools or more transparent approaches [92, 93]. Clear rules and guidelines are needed to resolve this uncertainty and build more trust in the use of AI in healthcare, potentially with explainable AI.

The RF models faced additional hurdles due to class imbalances, with more people passing the one year than people not reaching the one-year mark in the dataset. This uneven distribution can cause the model to become biased toward predicting the majority class, reducing specificity in identifying true responders. Techniques such as oversampling, under sampling, or applying class-weight adjustments could help with this bias, though they were not explored in depth here. Missing data also affected the models' reliability and generalizability, although it will also be seen in real-life situations. Important biomarkers or patient characteristics might be incomplete due to inconsistent record-keeping or patient dropout. Advanced imputation methods or models that can handle missing data intrinsically might address this limitation.

PFS is challenging to use as a binary outcome in RF models due to its inherent clinical complexity. PFS indicates the time during which a patient’s disease does not worsen, but defining a binary PFS response can be problematic. A patient might respond well to immunotherapy for several years before experiencing disease progression. In a binary PFS model, such a patient would be labelled a "non-responder" as soon as progression occurs, ignoring the treatment's earlier effectiveness.

Using "days to PFS" as a continuous target also introduces complications , as progression depends on scan intervals and is not measured continuous. If patients who never show progression are included, their time

to PFS would be undefined. To address this, some models use the date of death as a proxy for PFS, assuming death indicates disease progression, even though this is not confirmed by scans. This assumption may be inaccurate, especially if patients die from unrelated causes. This could obscure the true relationship between biomarkers and treatment effectiveness, complicating attempts to identify reliable predictors.

Overall, these challenges highlight the need for a nuanced approach when defining outcomes like PFS and OS in prognostic models, emphasizing the importance of careful data preprocessing and clinically informed labelling strategies.

6.4.3 Recommendations for Future Research

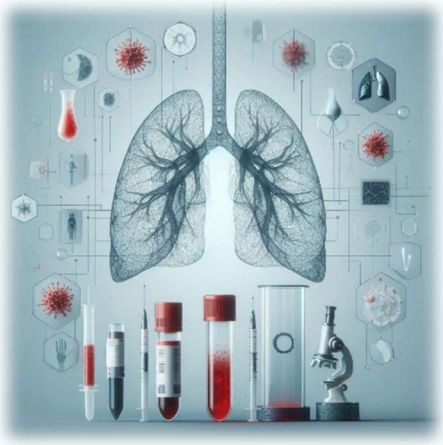
Future studies should enhance feature selection through advanced engineering techniques, incorporating interaction terms and biomarker trajectories. By capturing the relationships between variables and their changes over time, these methods could improve model performance and better reflect the dynamic nature of patient responses. Another important area to investigate is the application of explainability tools to improve the interpretability of RF models. Techniques such as SHAP (Shapley Additive Explanations) or feature importance visualization can provide insights into the decision-making process of RF models, enabling clinicians to understand the important drivers of predictions. This could bridge the gap between model complexity and clinical applicability, making RF models more transparent and actionable in practice. Future research can also focus on exploring more advanced predictive models. Given the superior performance of models integrating all three time points, temporal modelling approaches such as recurrent neural networks (RNNs) or time-series analyses could be explored. Causal ML could also provide a solution as it provides flexible, data-driven methods for predicting treatment outcomes, including efficacy and toxicity, thereby supporting the assessment and safety of treatments [94]. The main advantage of causal ML is its ability to estimate individualized treatment effects, allowing clinical decision-making to be personalized to individual patient profiles. Causal ML techniques can handle high-dimensional and unstructured data, including patient covariates, and estimate treatment effects from multimodal datasets containing images, text, time series, and genetic data. For instance, these methods can estimate treatment effects from CT scans or electronic health records. These models can predict personalized estimates of treatment effects for subpopulations or even predict outcomes for individual patients. They can identify patient subgroups for whom a treatment is effective. However, estimating treatment effects from data is challenging because individual patient outcomes under alternative treatments are not directly observable. Causal ML methods generate estimates, but these estimates must be carefully validated, as reliable decision-making in medical applications requires robust and well-calibrated evidence. Causal ML offers the possibility of predicting the efficacy and safety of treatments, and personalizing treatment strategies to improve patient health. Successful applications of causal ML in clinical use are still emerging, but they offer a good potential method in the future. Besides RF and causal Ml several other types of algorithms could still have potential like:

- Support Vector Machine (SVM) is a supervised learning model that can classify both linear and non-linear data. It works by mapping data points into an n-dimensional feature space and finding the hyperplane that best separates the data into two classes while maximizing the margin between them. The SVM can handle high-dimensional data and is effective in cases where the number of dimensions exceeds the number of samples and could therefore be useful on this dataset [56, 85, 95].
- Extreme Gradient Boosting (XGBoost) is an optimized gradient-boosting library designed to be highly efficient, flexible, and portable. It is used for supervised learning tasks and can handle regression, classification, ranking, and user-defined prediction problems. It is particularly known for its speed and performance [96].

6.5 Conclusion

RF modelling is a promising method for predicting patient responses to immunotherapy, outperforming traditional mixed effects models in clinical importance for NPV. By identifying significant biomarkers and improving predictive performance, RF can guide the refinement of clinical decision-making tools and models. However, future work should address limitations in interpretability and class imbalance while exploring complementary modelling approaches to enhance prediction accuracy and reliability. Integrating RF insights

with more transparent models, such as mixed effects models, could bridge the gap between predictive power and clinical applicability, ultimately improving patient outcomes.



CHAPTER 7: SUMMARY & RECOMMENDATIONS

This thesis explores the potential of routinely collected blood values and CT scans to predict the response of patients with stage IV NSCLC to immunotherapy. The findings provide a first insight into the relationship between clinical parameters and treatment outcomes, showing potential in predictive modelling in oncology, but also showing current complexities.

7.1 Summary

This research focused on improving the predictability of immunotherapy outcomes by leveraging routinely collected blood values and CT scan data. Various blood biomarkers were identified as significant predictors of OS at different stages of treatment, as detailed in Chapter 4. Among these, CRP emerged as a consistent and significant indicator of OS, underscoring the importance of monitoring inflammation during treatment. Significance of ratios such as NLR, PLR, and SII, which reflect the balance between immune cell populations, further emphasize the role of systemic immune balance in survival. This aligns with immunotherapy's mechanism of action, which activates the immune system to combat cancer. The predictive value of blood biomarkers for PFS was less consistent, possibly reflecting a more complex interplay of factors influencing PFS. PD-L1 status, a well-established biomarker, was a significant predictor of achieving CR but did not show a direct impact on OS. Blood values taken after the initiation of therapy were more strongly associated with OS, suggesting the importance of monitoring during treatment. While pre-treatment predictions are desirable, identifying early indicators of survival could still help clinicians avoid ineffective therapies, minimize side effects, and explore alternative treatments such as chemotherapy to prolong survival.

In this research, several networks were tested for their ability to automatically segment tumours from CT scans. Since accurate tumour segmentation is an important step in extracting radiomic features, this evaluation was performed to assess whether automated segmentation methods can support prognostic analysis. This aligns with the overarching goal of determining whether CT-based information can help predict long-term survival in patients undergoing immunotherapy. Among the tested networks, the nnU-Net achieved the highest Dice scores, outperforming the other tested networks. However, it currently lacks the robustness required for reliable analysis of all scans for the extraction of radiological features.

Developing MEMs to predict immunotherapy response was challenging due to factors such as convergence issues, multicollinearity, and limited dataset size. Despite these difficulties, MEMs demonstrated potential, with some models achieving high AUC values. However, integrating multiple variables often led to convergence failures, highlighting the need for careful model design, variable selection, and larger datasets.

RF models showed relatively good AUCs (~0.8) and demonstrated high NPV, which is particularly useful clinically. Although these models are less effective at predicting responders, their ability to reliably identify non-responders makes them valuable. The low Gini and permutation importance scores for individual variables suggest that the predictive power of RF lies in the combined contribution of multiple features, reflecting the multifactorial nature of immunotherapy response. The study emphasizes the trade-off between model complexity and interpretability. While simpler models may enhance clarity, they risk omitting critical variables. Conversely, more complex models, such as RF, function as "black boxes," potentially limiting their clinical applicability. Clear and interpretable models are essential for clinicians to make informed treatment decisions. Both MEM and RF models combining blood markers and clinical data demonstrated potential for more accurate

prediction of immunotherapy outcomes. However, challenges such as small datasets, generalizability, and interpretability must be addressed.

7.2 Clinical Relevance

This research shows potential implications for clinical practice in the future. Given the cost and potential side effects of immunotherapy, identifying patients most likely to benefit from the treatment is the attempted end point. While the study did not achieve its goal of identifying patients most likely to benefit from treatment, it demonstrated promise in developing models that can reliably identify non-responders. This represents an important first step, as excluding patients unlikely to benefit from therapy could reduce unnecessary exposure to treatment-related toxicity and improve resource allocation in clinical settings. However, implementing such models in practice remains challenging. For instance, while a model with an NPV of 0.80 is promising in a research context, the clinical decision to withhold therapy from a patient with a 20% chance of responding requires careful ethical consideration. Additionally, these models need further validation of these scores, before they can be integrated into routine clinical workflows. Some blood markers, as explored in this study, show potential as less invasive tools for identifying responders or non-responders. However, further research is required to refine these markers, including more extensive exploratory data analysis. Improved patient selection strategies could eventually have a significant clinical impact by sparing non-responders from unnecessary toxicity while maximizing the benefits of immunotherapy for responders. Validated predictive models, incorporating both blood biomarkers and other clinical data and CT scans, could support personalized treatment strategies in NSCLC, enhancing the overall effectiveness and efficiency of immunotherapy.

7.3 Recommendations for Future Research

Based on the findings, several recommendations are proposed to guide future research aimed at improving the predictability of immunotherapy outcomes and enhancing clinical applicability:
In the future incorporating CT image data in the models also provides valuable information about the tumour that could help predict response. The first step to achieve this would be to optimise the CT segmentation models, to be able to incorporate radiological features. This would include finding segmentation networks that can show higher performance scores, or retraining for example the current nnU-net on the dataset used in this research. The dataset worked with in this research also includes per/post-treatment scans, therefore it could improve the network to function on this dataset. Including data per/post-treatment is also very valuable as recent research showed potential of using the radiological features in distinguishing progression from pseudo progression [97]. Once segmentation models are sufficiently robust, radiological features extracted from CT scans can be integrated into predictive models to determine if they provide additional value in predicting therapy response. This step will help evaluate the combined impact of imaging and blood biomarkers. Developing models capable of reliably making this distinction would also provide immediate clinical value, even without integrating radiological features with blood biomarkers for response prediction. This could already assist doctors in accurately determining whether a patient is truly experiencing progression on a CT scan.

Collaborations between data collectors and developers are essential to create larger, more diverse datasets and improve the robustness of segmentation models. Such collaborations can support the development of networks that function reliably across different institutions and patient populations.

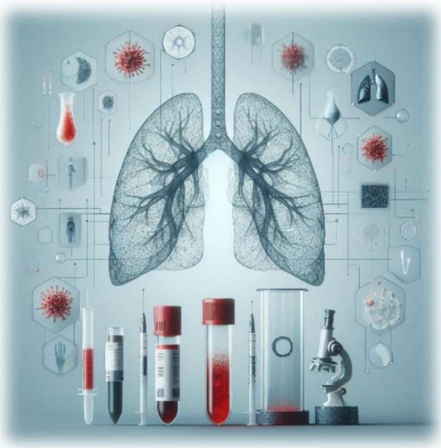
A second recommendation would be to perform more exploratory data analysis on more homogeneous and larger dataset. Creating a dataset where patients receive uniform treatment regimens (e.g., exclusively immunotherapy or a combination of chemotherapy and immunotherapy), ensuring consistent follow-up intervals, such as standardized time points for blood draws and imaging, will reduce variability and improve the reliability of findings. Efforts to include larger, multi-institutional datasets can enhance generalizability while maintaining homogeneity in critical variables like therapy type, dosage, and metastasis profiles. This approach will also address part of the missing values and enable tracking biomarker fluctuations over time better, leading to more robust models. With taking this step “back” research can focus on a better identification of both fixed and random effects in mixed-effect models to improve model performance. By refining these models, researchers can better understand the complex interactions between various biomarkers and treatment outcomes.

As a last recommendation, once more comprehensive and homogeneous datasets are available, the current AI models could be retrained/updated and be externally validated. Ensuring external validation of these

models is crucial to prevent overfitting and establish reliability across diverse patient populations. The potential use of neural networks and other advanced machine learning models should be explored for their potential to improve predictive performance. Although these models may lack interpretability, if they provide a higher predictive accuracy or NPV it could justify their use for specific clinical applications. If “black box” models, such as random forests or neural networks, demonstrate superior performance, efforts should be made to develop methods that enhance their interpretability. This could involve creating visualization tools or hybrid models that provide clinicians with actionable insights while retaining predictive strength.

7.4 Conclusion

This thesis provides first insights into the potential of blood values to predict immunotherapy response in NSCLC patients. The identified biomarkers, particularly CRP, and the promising results of random forest models in identifying non responders offer a foundation for future research. By expanding datasets, employing segmentation techniques, and improving model interpretability, this research paves the way for personalized medicine in NSCLC, ultimately leading to improved patient care.



REFERENCES

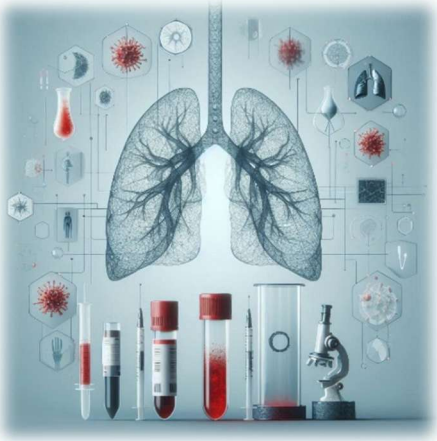
1. Wu, Y., et al., *Neoadjuvant immunotherapy for advanced, resectable non–small cell lung cancer: A systematic review and meta-analysis.* Cancer, 2023. **129**(13): p. 1969-1985.
2. Waser, N., et al., *Real-world treatment patterns in resectable (stages I–III) non-small-cell lung cancer: a systematic literature review.* Future Oncology, 2022. **18**(12): p. 1519-1530.
3. specialisten, F.m. *Niet kleincellig longcarcinoom*. 2020 25-04-2024]; Available from: https://richtlijnendatabase.nl/richtlijn/niet_kleincellig_longcarcinoom/startpagina_-_niet-kleincellig_longcarcinoom.html.
4. Li, S., et al., *Emerging Blood-Based Biomarkers for Predicting Response to Checkpoint Immunotherapy in Non-Small-Cell Lung Cancer.* (1664-3224 (Electronic)).
5. Putzu, C., et al., *Duration of Immunotherapy in Non-Small Cell Lung Cancer Survivors: A Lifelong Commitment?* Cancers, 2023. **15**(3): p. 689.
6. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* (1542-4863 (Electronic)).
7. Mouronte-Roibás, C., et al., *COPD, emphysema and the onset of lung cancer. A systematic review.* Cancer Letters, 2016. **382**(2): p. 240-244.
8. Schmidt, K., et al., *Preferences of lung cancer patients for treatment and decision-making: a systematic literature review.* European Journal of Cancer Care, 2016. **25**(4): p. 580-591.
9. Cui, Q., et al., *Prognostic significance of blood-based PD-L1 analysis in patients with non-small cell lung cancer undergoing immune checkpoint inhibitor therapy: a systematic review and meta-analysis.* (1477-7819 (Electronic)).
10. Society, A.C. *What Is Lung Cancer?* 29 January 2024; Available from: https://www.cancer.org/cancer/types/lung-cancer/about/what-is.html.
11. Wu, Z., et al., *Lung cancer risk prediction models based on pulmonary nodules: A systematic review.* Thoracic Cancer, 2022. **13**(5): p. 664-677.
12. Duma, N., R. Santana-Davila, and J.R. Molina, *Non–Small Cell Lung Cancer: Epidemiology, Screening, Diagnosis, and Treatment.* Mayo Clinic Proceedings, 2019. **94**(8): p. 1623-1640.
13. Ganti, A.K., et al., *Update of incidence, prevalence, survival, and initial treatment in patients with non–small cell lung cancer in the US.* JAMA oncology, 2021. **7**(12): p. 1824-1832.
14. van Delft, F.A.-O., et al., *The Validity and Predictive Value of Blood-Based Biomarkers in Prediction of Response in the Treatment of Metastatic Non-Small Cell Lung Cancer: A Systematic Review. LID - 10.3390/cancers12051120 [doi] LID - 1120.* (2072-6694 (Print)).
15. P.E. Postmus, K.M.K., M. Oudkerk, S. Senan, D.A. Waller, J. Vansteenkiste, C. Escriu, S. Peters. *Early and locally advanced non-small-cell lung cancer (NSCLC) guideline*. 2017; Available from: https://interactiveguidelines.esmo.org/esmo-web-app/gl_toc/index.php?GL_id=46.
16. Ettinger, D.S., et al., *Non-Small Cell Lung Cancer, Version 3.2022, NCCN Clinical Practice Guidelines in Oncology.* (1540-1413 (Electronic)).

17. Rescigno, M., F. Avogadri, and G. Curigliano, *Challenges and prospects of immunotherapy as cancer treatment.* Biochimica et Biophysica Acta (BBA) - Reviews on Cancer, 2007. **1776**(1): p. 108-123.
18. Disis, M.L., *Mechanism of Action of Immunotherapy.* Seminars in Oncology, 2014. **41**: p. S3-S13.
19. Kirkwood, J.M., et al., *Immunotherapy of cancer in 2012.* CA: A Cancer Journal for Clinicians, 2012. **62**(5): p. 309-335.
20. Scirocchi, F.A.-O., et al., *Soluble PD-L1 as a Prognostic Factor for Immunotherapy Treatment in Solid Tumors: Systematic Review and Meta-Analysis. LID - 10.3390/ijms232214496 [doi] LID - 14496.* (1422-0067 (Electronic)).
21. Sánchez de Cos Escuín, J., *New Immunotherapy and Lung Cancer.* (1579-2129 (Electronic)).
22. Haanen, J., et al., *Management of toxicities from immunotherapy.* ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up, 2018. **2018**: p. 29Suppl.
23. Bruyère, C.L.D.L., et al., *Investigating the Impact of Immune-Related Adverse Events, Glucocorticoid Use and Immunotherapy Interruption on Long-Term Survival Outcomes.* Cancers, 2021. **13**(10): p. 2365.
24. Zhang, Q., et al., *The Predictive Value of Pretreatment Lactate Dehydrogenase and Derived Neutrophil-to-Lymphocyte Ratio in Advanced Non-Small Cell Lung Cancer Patients Treated With PD-1/PD-L1 Inhibitors: A Meta-Analysis.* (2234-943X (Print)).
25. Simon, S.C.S., et al., *Eosinophil accumulation predicts response to melanoma treatment with immune checkpoint inhibitors.* (2162-4011 (Print)).
26. Tsim, S., et al., *Staging of non-small cell lung cancer (NSCLC): A review.* Respiratory Medicine, 2010. **104**(12): p. 1767-1774.
27. Costelloe, C.M., et al., *Cancer Response Criteria and Bone Metastases: RECIST 1.1, MDA and PERCIST.* (1837-9664 (Electronic)).
28. Seymour, L., et al., *iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics.* The Lancet Oncology, 2017. **18**(3): p. e143-e152.
29. Liu, Y., et al., *Radiological Image Traits Predictive of Cancer Status in Pulmonary Nodules.* Clinical Cancer Research, 2017. **23**(6): p. 1442-1449.
30. Van Griethuysen, J.J., et al., *Computational radiomics system to decode the radiographic phenotype.* Cancer research, 2017. **77**(21): p. e104-e107.
31. Ready, N., et al., *First-line nivolumab plus ipilimumab in advanced non–small-cell lung cancer (CheckMate 568): outcomes by programmed death ligand 1 and tumor mutational burden as biomarkers.* Journal of Clinical Oncology, 2019. **37**(12): p. 992.
32. Hellmann, M.D., et al., *Nivolumab plus ipilimumab in lung cancer with a high tumor mutational burden.* New England Journal of Medicine, 2018. **378**(22): p. 2093-2104.
33. Paz-Ares, L., et al., *Pembrolizumab (pembro) plus platinum-based chemotherapy (chemo) for metastatic NSCLC: tissue TMB (tTMB) and outcomes in KEYNOTE-021, 189, and 407.* Annals of Oncology, 2019. **30**: p. v917-v918.
34. Gerlinger, M., et al., *Intratumor heterogeneity and branched evolution revealed by multiregion sequencing.* New England journal of medicine, 2012. **366**(10): p. 883-892.
35. Mei, P., et al., *Prognostic value of lymphocyte-to-monocyte ratio in gastric cancer patients treated with immune checkpoint inhibitors: a systematic review and meta-analysis.* (1664-3224 (Electronic)).
36. Tanizaki, J., et al., *Peripheral blood biomarkers associated with clinical outcome in non–small cell lung cancer patients treated with nivolumab.* Journal of thoracic oncology, 2018. **13**(1): p. 97-105.
37. Moreira, A., et al., *Eosinophilic count as a biomarker for prognosis of melanoma patients and its importance in the response to immunotherapy.* (1750-7448 (Electronic)).
38. Prizment, A.E., et al., *Tumor eosinophil infiltration and improved survival of colorectal cancer patients: Iowa Women's Health Study.* Modern pathology, 2016. **29**(5): p. 516-527.

39. Sakkal, S., et al., *Eosinophils in cancer: favourable or unfavourable?* Current medicinal chemistry, 2016. **23**(7): p. 650-666.
40. Wankhede, D.A.-O., S. Grover, and P.A.-O. Hofman, *Circulating Tumor Cells as a Predictive Biomarker in Resectable Lung Cancer: A Systematic Review and Meta-Analysis. LID - 10.3390/cancers14246112 [doi] LID - 6112.* (2072-6694 (Print)).
41. Ouwerkerk, W., et al., *Biomarkers, measured during therapy, for response of melanoma patients to immune checkpoint inhibitors: a systematic review.* (1473-5636 (Electronic)).
42. Schober, P. and T.R. Vetter, *Kaplan-Meier Curves, Log-Rank Tests, and Cox Regression for Time-to-Event Data.* Anesthesia & Analgesia, 2021. **132**(4).
43. Kleinbaum, D.G. and M. Klein, *Kaplan-Meier Survival Curves and the Log-Rank Test*, in *Survival Analysis: A Self-Learning Text*, D.G. Kleinbaum and M. Klein, Editors. 2012, Springer New York: New York, NY. p. 55-96.
44. Rousseaux, C.G. and S.C. Gad, *Statistical assessment of toxicologic pathology studies.* Haschek and Rousseaux's Handbook of Toxicologic Pathology, 2013: p. 893-988.
45. Therneau, T.M., et al., *The cox model*. 2000: Springer.
46. Abd ElHafeez, S.A.-O., et al., *Methods to Analyze Time-to-Event Data: The Cox Regression Analysis.* (1942-0994 (Electronic)).
47. Hosmer Jr, D.W., S. Lemeshow, and S. May, *Applied survival analysis: regression modeling of time-to-event data*. Vol. 618. 2008: John Wiley & Sons.
48. Bayaga, A., *Multinomial Logistic Regression: Usage and Application in Risk Analysis.* Journal of applied quantitative methods, 2010. **5**(2).
49. Meloun, M. and J. Militký, *4 - Statistical analysis of multivariate data*, in *Statistical Data Analysis*, M. Meloun and J. Militký, Editors. 2011, Woodhead Publishing India. p. 151-403.
50. Cunnings, I., *An overview of mixed-effects statistical models for second language researchers.* Second Language Research, 2012. **28**(3): p. 369-382.
51. Edwards, L.J., et al., *An R² statistic for fixed effects in the linear mixed model.* Statistics in Medicine, 2008. **27**(29): p. 6137-6157.
52. Seltman, H., *Mixed models. A flexible approach to correlated data.* Experimental design and analysis, 2009: p. 357-377.
53. StatsTest.com. *Mixed Effects Model*. Available from: https://www.statstest.com/mixed-effects-model/#:~:text=A%20Mixed%20Effects%20Model%20is,the%20other%20assumptions%20listed%20below.
54. Harrison, X.A.-O., et al., *A brief introduction to mixed effects modelling and multi-model inference in ecology.* (2167-8359 (Print)).
55. Janiesch, C., P. Zschech, and K. Heinrich, *Machine learning and deep learning.* Electronic Markets, 2021. **31**(3): p. 685-695.
56. Mahesh, B., *Machine learning algorithms-a review.* International Journal of Science and Research (IJSR).[Internet], 2020. **9**(1): p. 381-386.
57. El Naqa, I. and M.J. Murphy, *What Is Machine Learning?*, in *Machine Learning in Radiation Oncology: Theory and Applications*, I. El Naqa, R. Li, and M.J. Murphy, Editors. 2015, Springer International Publishing: Cham. p. 3-11.
58. van Dijk, L.V., et al., *Head and neck cancer predictive risk estimator to determine control and therapeutic outcomes of radiotherapy (HNC-PREDICTOR): development, international multi-institutional validation, and web implementation of clinic-ready model-based risk stratification for head and neck cancer.* European Journal of Cancer, 2023. **178**: p. 150-161.
59. Habehh, H. and S. Gohel, *Machine Learning in Healthcare.* (1389-2029 (Print)).
60. Ait Skourt, B., A. El Hassani, and A. Majda, *Lung CT Image Segmentation Using Deep Neural Networks.* Procedia Computer Science, 2018. **127**: p. 109-113.
61. Primakov, S.P., et al., *Automated detection and segmentation of non-small cell lung cancer computed tomography images.* Nature Communications, 2022. **13**(1): p. 3423.

62. Tribuana, D., Hazriani, and A.L. Arda, *Image Preprocessing Approaches Toward Better Learning Performance with CNN.* Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 2024. **8**(1): p. 1-9.
63. De Raad, K.B., et al. *The Effect of Preprocessing on Convolutional Neural Networks for Medical Image Segmentation*. IEEE.
64. Ronneberger, O., P. Fischer, and T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015, Springer International Publishing. p. 234-241.
65. Jiang, H., Z. Diao, and Y.-D. Yao, *Deep learning techniques for tumor segmentation: a review.* The Journal of Supercomputing, 2022. **78**(2): p. 1807-1851.
66. Razavi, S., *Deep learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling.* Environmental Modelling & Software, 2021. **144**: p. 105159.
67. Lecun, Y., Y. Bengio, and G. Hinton, *Deep learning.* Nature, 2015. **521**(7553): p. 436-444.
68. Riaz, Z.A.-O., et al., *Lung Tumor Image Segmentation from Computer Tomography Images Using MobileNetV2 and Transfer Learning. LID - 10.3390/bioengineering10080981 [doi] LID - 981.* (2306-5354 (Print)).
69. Fredriksen, V., et al., *Teacher-student approach for lung tumor segmentation from mixed-supervised datasets.* PLOS ONE, 2022. **17**(4): p. e0266147.
70. Isensee, F., et al., *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation.* Nature methods, 2021. **18**(2): p. 203-211.
71. Aerts, H.J.W.L., et al., *Data From NSCLC-Radiomics*. 2019, The Cancer Imaging Archive.
72. Veldman, C.S., Romée. Smit, Veerle, *Pocket Longgeneeskunde*. 2022: Synopsis BV. 320.
73. 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, 2020. **12**(8): p. 2319.
74. Mirrakhimov, A.E., et al., *Tumor lysis syndrome: a clinical review.* World Journal of Critical Care Medicine, 2015. **4**(2): p. 130.
75. Liu, Y., et al., *Dysregulation of immunity by cigarette smoking promotes inflammation and cancer: A review.* Environmental Pollution, 2023. **339**: p. 122730.
76. Yeghaian, M., et al., *Can blood-based markers predict RECIST progression in non-small cell lung cancer treated with immunotherapy?* Journal of Cancer Research and Clinical Oncology, 2024. **150**(6).
77. Tsoulos, N., et al., *Tumor molecular profiling of NSCLC patients using next generation sequencing.* Oncology reports, 2017. **38**(6): p. 3419-3429.
78. Stern, S., et al., *Prediction of response to drug therapy in psychiatric disorders.* Open biology, 2018. **8**(5): p. 180031.
79. Vogl, W.-D., et al., *Analyzing and predicting visual acuity outcomes of anti-VEGF therapy by a longitudinal mixed effects model of imaging and clinical data.* Investigative Ophthalmology & Visual Science, 2017. **58**(10): p. 4173-4181.
80. Philip, N.S., et al., *Theta-burst transcranial magnetic stimulation for posttraumatic stress disorder.* American Journal of Psychiatry, 2019. **176**(11): p. 939-948.
81. McInnes, A.N., et al., *Trajectory Modeling and Response Prediction in Transcranial Magnetic Stimulation for Depression. LID - 100135 [pii] LID - 10.1016/j.pmip.2024.100135 [doi].* (2468-1725 (Print)).
82. Iniesta, R., et al., *Combining clinical variables to optimize prediction of antidepressant treatment outcomes.* Journal of psychiatric research, 2016. **78**: p. 94-102.
83. Ribba, B., et al., *A review of mixed-effects models of tumor growth and effects of anticancer drug treatment used in population analysis.* CPT: pharmacometrics & systems pharmacology, 2014. **3**(5): p. 1-10.
84. Giger, M.L., *Machine Learning in Medical Imaging.* Journal of the American College of Radiology, 2018. **15**(3, Part B): p. 512-520.
85. Uddin, S., et al., *Comparing different supervised machine learning algorithms for disease prediction.* BMC Medical Informatics and Decision Making, 2019. **19**(1).

86. Parmar, A., R. Katariya, and V. Patel, *A Review on Random Forest: An Ensemble Classifier*. 2019, Springer International Publishing. p. 758-763.
87. Algehyne, E.A., et al. *Fuzzy Neural Network Expert System with an Improved Gini Index Random Forest-Based Feature Importance Measure Algorithm for Early Diagnosis of Breast Cancer in Saudi Arabia*. Big Data and Cognitive Computing, 2022. **6**, DOI: 10.3390/bdcc6010013.
88. Sauer, C.M., et al., *Feature selection and prediction of treatment failure in tuberculosis.* PLOS ONE, 2018. **13**(11): p. e0207491.
89. Visonà, G., et al., *Machine-Learning-Aided Prediction of Brain Metastases Development in Non–Small-Cell Lung Cancers.* Clinical Lung Cancer, 2023. **24**(8): p. e311-e322.
90. Bin, J., et al., *Predicting invasion in early-stage ground-glass opacity pulmonary adenocarcinoma: a radiomics-based machine learning approach.* (1471-2342 (Electronic)).
91. Janitza, S., C. Strobl, and A.-L. Boulesteix, *An AUC-based permutation variable importance measure for random forests.* BMC bioinformatics, 2013. **14**: p. 1-11.
92. Cabral, B.A.-O., et al., *Future of Artificial Intelligence Applications in Cancer Care: A Global Cross-Sectional Survey of Researchers.* (1718-7729 (Electronic)).
93. Marcus, E. and J. Teuwen, *Artificial intelligence and explanation: How, why, and when to explain black boxes.* European Journal of Radiology, 2024. **173**: p. 111393.
94. Feuerriegel, S., et al., *Causal machine learning for predicting treatment outcomes.* Nature Medicine, 2024. **30**(4): p. 958-968.
95. Zhang, Z., *Introduction to machine learning: k-nearest neighbors.* Annals of Translational Medicine, 2016. **4**(11): p. 218-218.
96. Arif Ali, Z., et al., *eXtreme Gradient Boosting Algorithm with Machine Learning: a Review.* Academic Journal of Nawroz University, 2023. **12**(2): p. 320-334.
97. Li, Y., et al., *Noninvasive radiomic biomarkers for predicting pseudoprogression and hyperprogression in patients with non-small cell lung cancer treated with immune checkpoint inhibition.* OncoImmunology, 2024. **13**(1): p. 2312628.
98. Harrison, X.A., et al., *A brief introduction to mixed effects modelling and multi-model inference in ecology.* PeerJ, 2018. **6**: p. e4794.
99. Lindstrom, M.J. and D.M. Bates, *Nonlinear mixed effects models for repeated measures data.* Biometrics, 1990: p. 673-687.
100. Edwards, L.J., et al., *An R2 statistic for fixed effects in the linear mixed model.* Statistics in medicine, 2008. **27**(29): p. 6137-6157.
101. Isensee, F., et al., *nnu-net: Self-adapting framework for u-net-based medical image segmentation.* arXiv preprint arXiv:1809.10486, 2018.



Appendix A:
Mixed Effect Models explained

MEMs, also known as multilevel models, are a powerful class of statistical models increasingly used across various scientific disciplines, including biology [98] & medicine [83, 99, 100]. Their popularity stems from their ability to analyse complex data structures, particularly those involving repeated measurements or clustered data [98]. They offer several advantages over traditional statistical methods, most notably their ability to account for correlations between repeated measurements and individual variability in therapy response. [50, 99] These models provide a flexible framework for studying the relationships between variables while accounting for the hierarchical structure of the data and the variability among individuals or groups [98, 99]. Most important in MEMs is the concept of fixed and random effects [99]. Fixed effects represent the impact of variables that are central to the research question, such as treatment or experimental condition. These effects are assumed to be constant across all individuals or groups in the sample [98]. Random effects, in contrast, model the variability among individuals or groups in their response to the fixed effects [50, 98]. MEMs build on traditional linear models by integrating a combination of fixed and random effects as predictor variables. By including random effects, these models explicitly account for the correlation between repeated measurements within an individual while simultaneously acknowledging heterogeneity in individual responses to therapy. [50, 79, 99]

MEM also offer several other advantages over traditional methods [50, 79, 99]:

- Handling Missing Data. MEMs can accommodate missing data under the assumption of "missing at random" (MAR), meaning the likelihood of missing data does not depend on the unobserved values.
- Identifying Predictors of Therapy Response: MEMs enable researchers to investigate factors (e.g., age, gender, baseline disease activity) that predict therapy response, paving the way for personalized treatment plans.
- Predicting Future Outcomes: Based on modelled individual trajectories and identified predictors, MEMs can forecast future therapy responses, aiding decisions on optimal treatment duration or alternative therapies.

Within the family of MEMs, various models are tailored to the specific nature of the data and research questions: [50]

LMMs are the simplest form of MEMs and are used when the outcome variable is continuous and follows a normal distribution [50, 100]. An example is the analysis of plant growth under different treatments, where the height of each plant is measured at multiple time points. In this case, the treatment would be a fixed effect, while the individual plant would be a random effect. The mathematical formulation of an LMM is:

$$y_i = X_i b + Z_i b_i + e_i$$

- *y_i: Vector of observations for individual i*
- *X_i: Design matrix for fixed effects for individual i*
- *b: Vector of fixed effect parameters*
- *Z_i: Design matrix for random effects for individual i*

- *b_i: Vector of random effects for individual i*
- *e_i: Residual errors for individual i*

This equation shows that the response of each individual is determined by a combination of fixed effects, random effects, and residual errors [100].

GLMMs extend the applicability of LMMs to non-normal outcome variables, such as binary (yes/no), count, or categorical data[50, 98]. For instance, a study on animal survival in different habitats could use a GLMM with survival (alive or dead) as the binary outcome variable and habitat as a fixed effect. GLMMs use link functions to model the relationship between the linear predictor and the non-normal outcome variable [99]. Examples of link functions include the logit, probit, and log functions [99]. The mathematical formulation will then become:

$$g\big(E(y_i)\big) = X_i b + Z_i b_i$$

- *g: link function (e.g., logit, probit, log)*
- *$E(y_i)$: expected value of the outcome variable for individual i*

The choice of the link function depends on the distribution of the outcome variable. For example, the logit link function is used for binary data, while the log link function is commonly used for count data.

Generalized Additive Mixed Models (GAMMs) combine the flexibility of generalized additive models (GAMs) with the strength of MEMs [98]. GAMs allow for non-linear relationships between predictors and the outcome variable, making them suitable for complex datasets [98]. GAMMs introduce non-linear relationships between predictors and the outcome variable using smoothing functions (f):

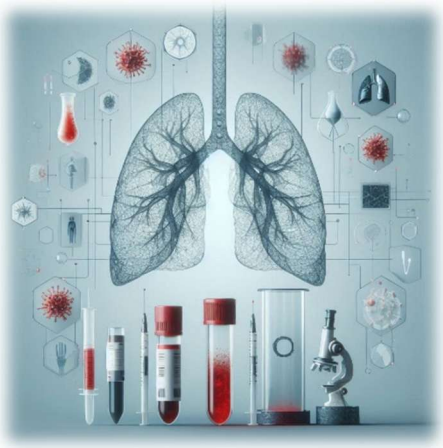
$$g(E(y_i)) = b_0 + f_1(x_{1i}) + f2(x_{2i}) + \ldots + Z_i b_i$$

- *f_1, f_2, ...: smoothing functions for predictors x_{1i}, x_{2i}, ...*

These smoothing functions are estimated from the data and allow for complex non-linear relationships.

The key difference between these models lies in how they model the relationship between predictors and the outcome variable: LMMs: Assume a linear relationship and a normal distribution for the outcome variable. GLMMs: Handle non-normal outcome variables using link functions while retaining the linear relationship between predictors and the transformed outcome variable. GAMMs: Allow for non-linear relationships through the use of smoothing functions. In summary, LMMs are the most constrained, while GAMMs are the most flexible. The choice of model depends on the specific research question and the characteristics of the data. While MEMs are a powerful tool, careful implementation is essential. The choice of fixed and random effects is critical for reliable results and should be based on the research question, data structure, and underlying mechanisms [98]. Statistical criteria like AIC and BIC can guide the selection of the best-fitting model. Comparing models with different combinations of fixed and random effects is important to improve and find a fitting model[98]. Residual plots and predictive accuracy should also be assessed to ensure the model adequately describes the data.

MEMs provide a robust and flexible framework for analysing longitudinal data and predicting therapy response. By accounting for correlations between repeated measurements within individuals and heterogeneity across individuals, MEMs yield valuable insights into therapy effectiveness and the factors influencing response.



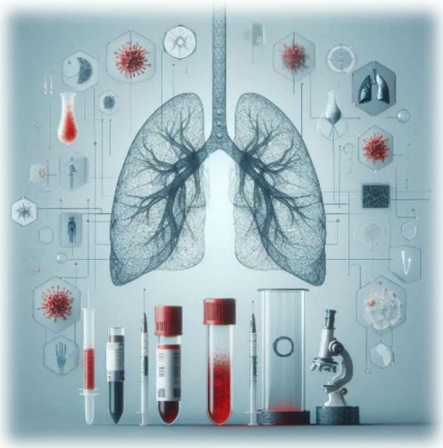
Appendix B:

Preprocessing types

Table 17: Preprocessing possibilities [62-64]

Preprocessing type	Options	What	How
Normalization	Min-max	Rescales pixel values to a consistent range (e.g., [0, 1] or [-1, 1]), ensuring uniform input	$x' = \frac{x - min(x)}{max(x) - min(x)}$ Rescales original pixel value x based on its minimum and maximum values in the dataset.
	Mean	Centers data around 0, eliminating the effect of lighting conditions or sensor variations and improving convergence	$x' = \frac{x - \mu}{\sigma}$ Subtracts the dataset's mean μ from each pixel value x and divides by the standard deviation σ
Noise Reduction	Gaussian blur	Smooths the image by averaging pixel values with neighbors, reducing sharp noise without distorting the image	$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$ Applies a Gaussian filter where $G(x,y)$ represents the Gaussian kernel and σ controls the smoothing. Pixels are averaged with neighboring pixels weighted by the Gaussian distribution.
	Median Filtering	Removes dotted noise while preserving edges	Replaces each pixel value with the median value of neighboring pixels in a defined window (e.g., 3x3 or 5x5). This approach is effective in reducing isolated noise points.
Standardization of Input Size	Resizing	Ensures uniform input size for all images to meet the fixed dimensional requirements of neural networks.	Resizes all images to a target size (e.g., 224x224 pixels) using interpolation techniques like bilinear or bicubic interpolation. However, resizing can distort the image's aspect ratio.
	Cropping and Padding	Crops or pads images to focus on the region of interest or preserve aspect ratio.	Crops images to a fixed size, focusing on key regions, or pads images (with values like 0 or border reflection) to fit target

			dimensions without distorting the image's content.
Data Augmentation	Rotation	Simulates different orientations of objects in the image.	Rotates the image by small random angles (e.g., ±10°) to create variations.
	Flipping	Introduces variations in image orientation	Flips images horizontally or vertically
	Translation	Shifting the position of the object in the image	Shifts the image horizontally or vertically by a small amount (e.g., 10-20 pixels)
	Brightness/Contrast Adjustment	Adjustment of contrast and brightness	Brightness is adjusted by adding or subtracting a constant from all pixel values, and contrast by scaling pixel values by a factor
	Adding noise		Adds small amounts of random noise (e.g., Gaussian noise) to images



Appendix C: Neural Networks explained

Deep learning [55, 66], a subset of machine learning within AI, is designed to model complex patterns and relationships in large datasets. It employs deep neural network, structures with multiple layers of artificial neuron, to map input data to meaningful outputs. The term deep refers to these networks' many hidden layers, each of which learns a different level of abstraction. As data passes through the layers, the network transforms it progressively into higher-level representations, enabling the model to learn complex patterns. A neural network mimics the way the human brain processes information. It consists of layers of artificial neurons, connected in a network-like structure, see figure 9. These neurons receive input, perform simple calculations, and pass the output to the next layer. Neural networks typically have three types of layers: an input layer, where data enters the network; hidden layers, where the bulk of the processing occurs and patterns are learned; and an output layer, which generates the final prediction. The network learns through training, where it adjusts the connections and weights, calculations in its neurons based on examples it has seen, enabling it to make accurate predictions on new data. In the context of image processing, deep learning amplifies features in an image that distinguish between different classes, while suppressing irrelevant details. Early layers of a deep neural network might detect basic elements like edges and textures, while deeper layers identify more complex structures. This is achieved through a chain of mathematical functions, where each layer processes the data and refines the network's output to match the desired target.

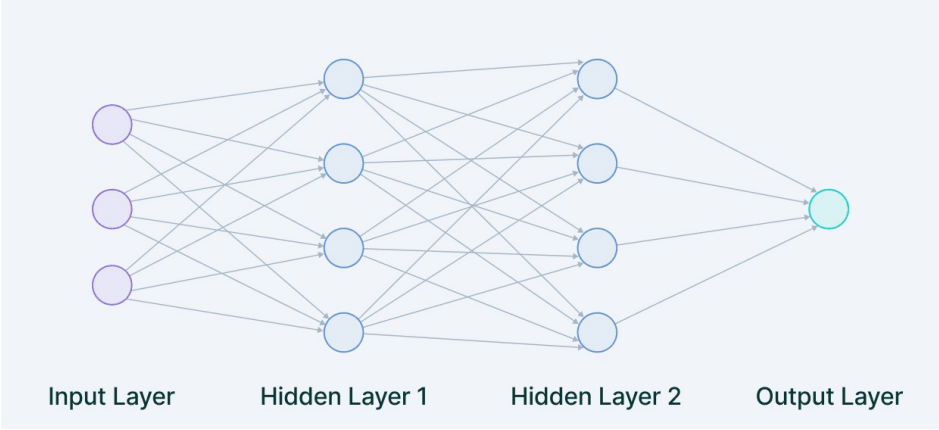


Figure 6: Neural network architecture, retrieved from: https://www.v7labs.com/blog/neural-network-architectures-guide

Convolutional Neural Networks

CNNs are a specialized type of artificial neural network designed for processing grid-structured data, such as images [55, 60, 66, 67]. They are particularly effective for tasks like image classification, object detection, and image segmentation because they can automatically learn meaningful patterns from raw image data. CNNs process images through a series of layers, each designed to transform and refine the data to extract features and make predictions. The first and most fundamental layer in a CNN is the convolutional layer, which is responsible for detecting patterns in the image. This layer applies a small grid, typically 3x3 or 5x5, containing

numbers (weights) called a filter or kernel to the input image. A kernel can be thought of as a sliding window that moves across the image, analyzing small regions at a time.

[1, 0, -1]
[1, 0, -1]
[1, 0, -1]

Figure 7: Example of a kernel, that detects edges in the images.

When this kernel moves over the image, it looks for changes in intensity from left to right, which would indicate an edge. This process of moving the kernel over the image is called convolution. At each position, the kernel performs an operation called a convolution: it multiplies the pixel values of the region it covers with the corresponding values in the kernel, then sums the results into a single number. This process creates a new representation of the image, known as a feature map, which highlights where specific patterns (like edges or shapes) are present. Different kernels can detect different types of patterns, such as edges, corners, or textures. For example, one kernel might be sensitive to horizontal edges, while another might detect vertical edges. By stacking multiple convolutional layers, a CNN can learn increasingly complex features, from simple edges to high-level structures like shapes or objects. After each convolutional layer, an activation function is applied to introduce non-linearity into the network. The most common activation function used in CNNs is the Rectified Linear Unit (ReLU) function, which replaces all negative values in the feature map with zeros. This allows the network to learn more complex patterns compared to linear transformations alone. To reduce the size of the feature maps and make the network more efficient, CNNs use a pooling layer. This layer reduces the spatial dimensions of the feature map while retaining its most important information. A common type is Max Pooling, which divides the feature map into small regions (for example, 2x2) and takes the maximum value from each region. Another type is Average Pooling, which takes the average value of each region. Pooling reduces the number of parameters and computations, making the model more efficient and less likely to overfit. It also helps the network focus on the most prominent features. Once the feature maps have been processed through several convolutional and pooling layers, they are flattened into a one-dimensional vector and passed to a fully connected layer. This layer comes after the convolutional and pooling layers. It flattens the feature maps into a single vector and connects every neuron in this vector to every neuron in the next layer. Fully connected layers are typically used for making final predictions, such as class probabilities in classification tasks.

The U-Net

The U-Net [64] architecture is a specific type of CNN designed for the task of image segmentation, where the goal is to classify each pixel in an image. Unlike traditional classification tasks that assign a single label to an entire image, segmentation involves creating a detailed map that identifies objects or regions at the pixel level. The U-Net is named after its U-shaped structure, which consists of two main components: an encoder and a decoder, see image ... These two parts are connected by a bottleneck and utilize skip connections to improve accuracy. The encoder, also known as the contracting path, is responsible for extracting features from the input image. It begins with the raw image and is responsible for extracting features from the input image. It progressively reduces the spatial dimensions (height and width) of the image while increasing the number of feature maps (depth). Each step in the encoder consists of convolutional layers, which extract spatial features such as edges and textures, possibly a activation function, followed by a pooling layer, which halves the dimensions of the feature maps. For example, an input image of size 256x256 might be reduced to 128x128 after one pooling step and to 64x64 after another. As the spatial size decreases, the depth of the feature maps increases, allowing the network to capture more complex features. At the deepest point of the network, known as the bottleneck, the spatial dimensions of the image are small, but the feature maps are rich with abstract features. The bottleneck consists of convolutional layers that capture high-level patterns in the data, providing a compact representation of the input image.

The decoder, or expanding path, reconstructs the image to its original size while using the features extracted by the encoder to make pixel-wise predictions. This is done through upsampling layers, which increase the spatial dimensions of the feature maps. At each step, the decoder uses skip connections to combine the feature maps from the encoder with those in the decoder. These skip connections ensure that fine details lost during the

downsampling process are preserved and incorporated into the final output. For instance, if the encoder detects fine edges or textures, the decoder can use this information to produce more accurate segmentation results. After upsampling and combining features, the decoder uses convolutional layers to refine the upsampled feature maps to produce precise predictions. The decoder outputs a segmentation mask of the same size as the original image, with each pixel assigned to a specific class. The U-Net architecture is specifically designed for image segmentation tasks where precise localization is critical. For example, in medical imaging, U-Net can be used to identify tumours in an MRI scan or segment organs in a CT image. Given an input image, U-Net produces a segmentation mask where each pixel belongs to a specific class, such as tumour, healthy tissue, or background. This pixel-wise classification capability makes U-Net suitable for applications where fine-grained detail is required. The nnU-Net [101], or "No New U-Net," is an advanced version of the U-Net designed to simplify and optimize image segmentation without requiring manual adjustments. It automatically adapts its architecture to different datasets and tasks, selecting the best configuration—such as layer numbers and kernel sizes—based on the input data. Like U-Net, it uses an encoder-decoder structure with skip connections, but its strength lies in this dynamic adaptation. Particularly effective in medical image segmentation, nnU-Net delivers high accuracy, outperforming many custom networks, while reducing the need for human intervention.

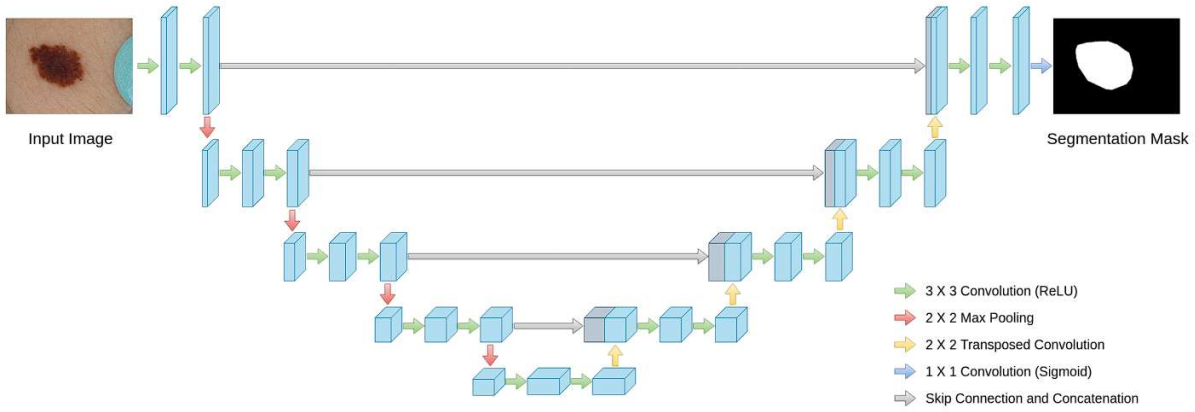
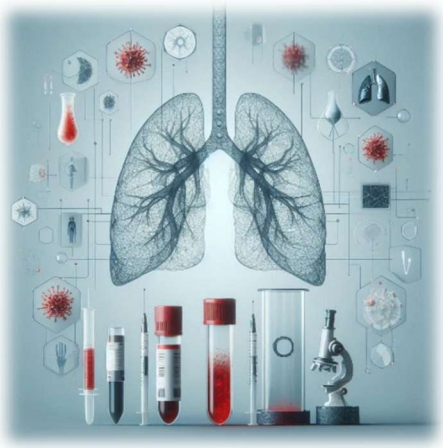


Figure 8: The U-net architecture example, retrieved from: https://viso.ai/deep-learning/u-net-a-comprehensive-guide-to-its-architecture-and-applications/



Appendix D:

Descriptive statistics

Table 18: Descriptive Statistics for blood value continuous parameters

Variable	n	miss	p_miss	mean	sd	median	p25	p75	min	max
Sodium	754	1	0.13	138.64	3.26	139	137	141	125	150
LDH	750	5	0.67	273.1	295.6	238	198	288	102	5791
CRP	754	1	0.13	28.91	40.5	12	4	38	0.5	297
Bilirubine Total	693	62	8.95	6.4	3.47	6	4	8	1	32
eGFR (CKD-EPI)	755	0	0	78.81	19.14	82	66	100	15	114
TSH	704	51	7.24	2.95	6.48	1.6	1	2.6	0	95
MCV	755	0	0	93.59	6.84	93	89	97	74	122
Creatinine	755	0	0	81.10	29.06	76	63	92	34	309
Free T4 (FT4)	711	44	6.19	17.02	4.16	16	15	19	1	56
Haemoglobin	755	0	0	7.78	1.16	7.8	7	8.6	4.6	10.7
Trombocyten	754	1	0.13	306.12	116.57	292	230	365	29	911
Potassium	754	1	0.13	4.35	0.46	4.3	4.1	4.6	2.9	6.3
Calcium	749	6	0.80	2.41	0.13	2.41	2.33	2.48	1.64	2.99
Leukocytes	755	0	0	8.85	3.94	8.2	6	10.75	1.4	37
Lymfocytes	739	16	2.17	1.75	0.82	1.6	1.2	2.1	0.19	6.8
Glucose	625	130	20.8	7.34	3.01	6.4	5.6	8	3.1	33.2
Neutrophils	736	19	2.58	5.79	3.42	5	3.4	7.23	0.04	29.1
Monocytes	739	16	2.17	0.85	0.38	0.81	0.62	1	0.05	5.1
Eosinophils	739	16	2.17	0.22	0.25	0.15	0.07	0.28	0	2.3
ALAT	752	3	0.40	30.62	38.07	21	14	33	4	513
ASAT	753	2	0.27	33.89	28.01	28	22	37	11	378
NLR	734	21	2.86	4.29	5.11	2.92	1.93	5	0.03	81.58
PLR	738	17	2.30	213.81	147.20	174.83	122.26	255.76	26.36	1463.16
LMR	738	17	2.31	2.33	1.32	2.06	1.45	2.91	0.19	11.19
SII	734	21	2.86	1388.08	1871.14	838.73	497.97	1527.93	2.46	21210

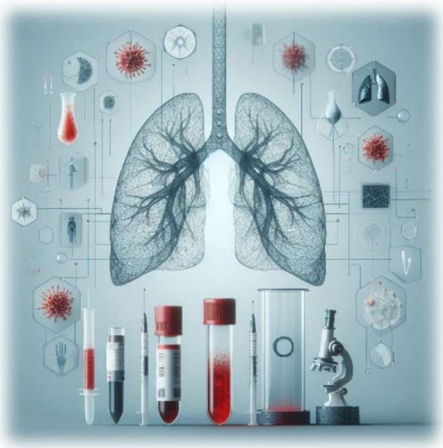
Table 19: Descriptive Statistics for blood value binary parameters, with false meaning a patient had a value outside of the normal ranges at any time during therapy measured.

Parameter	FALSE_Count	TRUE_Count
Sodium_normal	52 (24.1)	164 (75.9)
LDH_normal	145 (67.1)	71 (32.9)

CRP_normal	185 (85.6)	31 (14.4)
Bilirubine Total_normal	57 (26.4)	159 (73.6)
eGFR (CKD-EPI)_normal	98 (45.4)	118 (54.6)
TSH_normal	78 (36.1)	138 (63.9)
MCV_normal	52 (24.1)	164 (75.9)
Kreatinine_normal	93 (43.1)	123 (56.9)
Glucose_normal	111 (51.4)	105 (48.6)
Free T4 (FT4)_normal	53 (24.5)	163 (75.5)
Hemoglobine_normal	143 (66.2)	73 (33.8)
Trombocytes_normal	97 (44.9)	119 (55.1)
Potassium_normal	46 (21.3)	170 (78.7)
Calcium_normal	38 (17.6)	178 (82.4)
Leukocytes_normal	145 (67.1)	71 (32.9)
Lymfocytes_normal	78 (36.1)	138 (63.9)
Neutrophils_normal	131 (60.6)	85 (39.4)
Monocytes_normal	94 (43.5)	122 (56.5)
Eosinophils_normal	48 (22.2)	168 (77.8)
ALAT_normal	68 (31.5)	148 (68.5)
ASAT_normal	101 (46.8)	115 (53.2)
NLR_normal	186 (86.1)	30 (13.9)
PLR_normal	125 (57.9)	91 (42.1)
LMR_normal	167 (77.3)	49 (22.7)
SII_normal	157 (72.7)	59 (27.3)

Table 20: Descriptive Statistics for non blood value continuous parameters

Variable	n	miss	p.miss	mean	sd	median	p25	p75	min	max
Age at time of diagnosis	216	0	0	67.09	9.64	67.73	60.41	74.00	37.92	89.77
BMI	215	1	0.47	25.41	4.83	24.97	22.38	27.84	13.52	41.1
Pack years	172	44	25.58	37.75	18.22	40	30	48	0	150
FEV1 (L)	151	65	30.09	2.14	0.81	2.03	1.52	2.53	0.65	5
FEV1 (% pred)	151	65	30.09	71.78	19.45	72	60	83	24.5	113
DLCO	139	77	35.65	5.48	1.97	5.06	4.02	6.68	1.68	11.52
DLCO (% pred)	114	102	47.22	64.79	18.67	64	50	77	24	110
DLCO/VA	134	82	37.96	1.12	0.29	1.13	0.94	1.28	0.4	1.97
DLCO/VA (% pred)	102	114	85.78	79.70	21.30	79	64.04	92.75	28.45	144
ECOG-score at start of therapy	173	43	19.91	1.61	1.61	1	0	1	0	3



Appendix E:
Survival function PD-L1

Table 21: P-values of the log rank for OS for PD-L1 and the method used to acquire the samples.

	Log-rank p-value
All	0.16
Pathologisch	0.90
Cytologisch	0.02

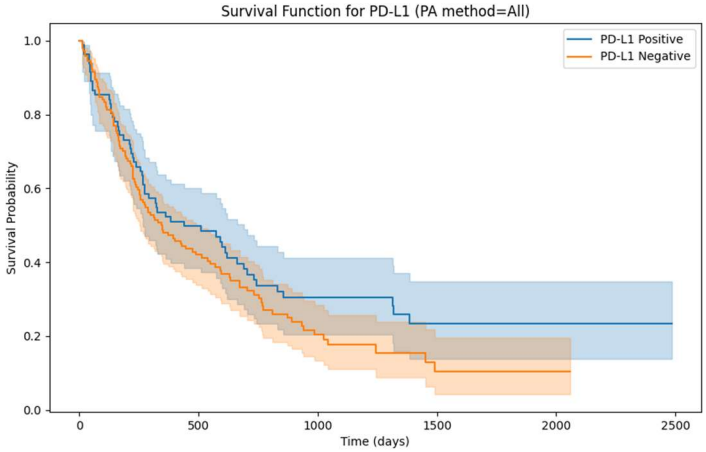


Figure 12: Kaplan Meier curve of OS for PD-L1.

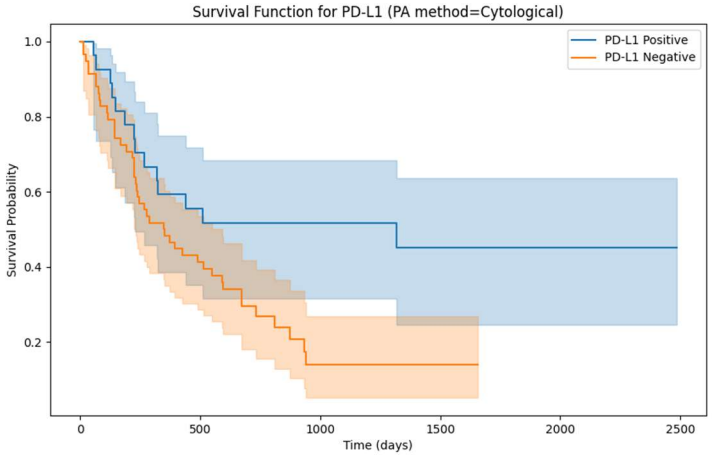


Figure 13: Kaplan Meier curve of OS for PD-L1 cytological determined

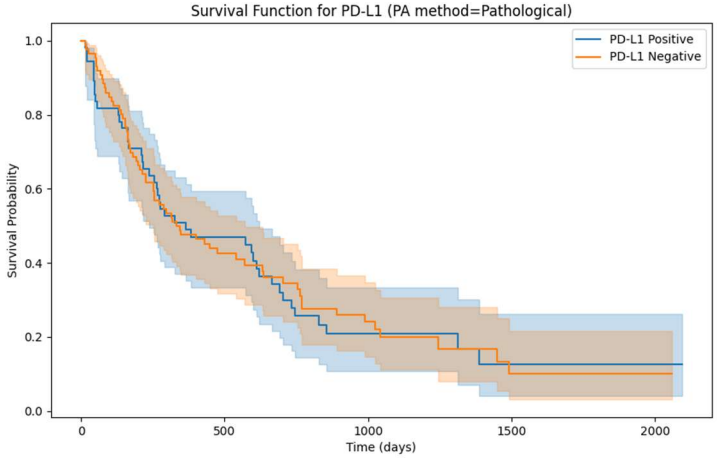
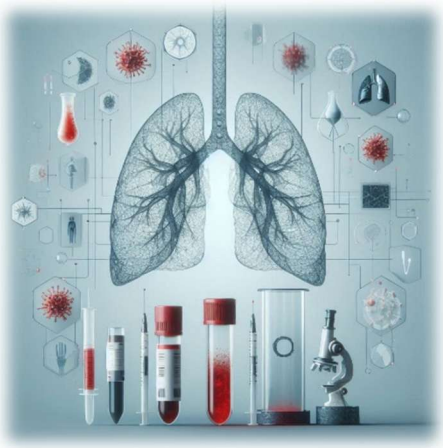


Figure 14: Kaplan Meier curve of OS for PD-L1 pathological determined

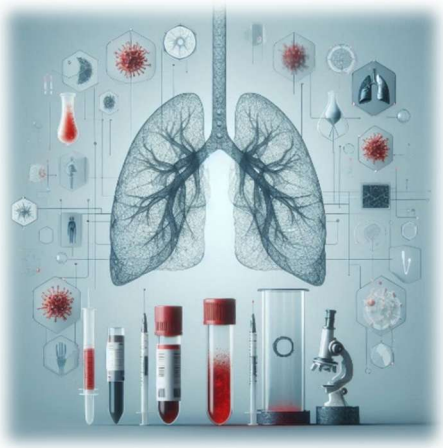


Appendix F:
Blood variables per model

Table 22: Variables included per MEM.

Model 1 t0+1+3	Sodium + Sodium_quad + Sodium_cube + LDH + LDH_quad + LDH_cube + CRP + CRP_quad + CRP_cube + Total_Bilirubin + Total_Bilirubin_quad + Total_Bilirubin_cube + Hemoglobin + Hemoglobin_quad + Hemoglobin_cube + Monocytes + Monocytes_quad + Monocytes_cube + eGFR_CKD_EPI + eGFR_CKD_EPI_quad + eGFR_CKD_EPI_cube + TSH + TSH_quad + TSH_cube + MCV + MCV_quad + MCV_cube + Creatinine + Creatinine_quad + Creatinine_cube + Free_T4 + Free_T4_quad + Free_T4_cube + Platelets + Platelets_quad + Platelets_cube + Glucose + Glucose_quad + Glucose_cube + Potassium + Potassium_quad + Potassium_cube + Leukocytes + Leukocytes_quad + Leukocytes_cube + Calcium + Calcium_quad + Calcium_cube + Lymphocytes + Lymphocytes_quad + Lymphocytes_cube + Neutrophils + Neutrophils_quad + Neutrophils_cube + ALAT + ALAT_quad + ALAT_cube + NLR + NLR_quad + NLR_cube + PLR + PLR_quad + PLR_cube + LMR + LMR_quad + LMR_cube + Eosinophils + Eosinophils_quad + Eosinophils_cube + SII + SII_quad + SII_cube + ASAT + ASAT_quad + ASAT_cube
Model 1.1 t0+1+3	Sodium + Sodium_quad + Sodium_cube + LDH + Bilirubine_Total + Bilirubine_Total_quad + Hemoglobine_quad + Hemoglobine_cube + Monocytes_cube + FT4 + FT4_quad + Potassium + Potassium_quad + Potassium_cube + Calcium + Calcium_quad + Calcium_cube + ASAT
Model 2 t0+1+3	Sodium + LDH + CRP + Bilirubine_Total + Haemoglobin + Monocytes + eGFR_CKD_EPI + TSH + MCV + Creatinine + FT4 + Platelets + Glucose + potassium + Leukocytes + Calcium + Lymfocytes + Neutrophils + ALAT + NLR + PLR + LMR + Eosinophils + SII + ASAT
Model 3 t1+3	CRP
Model 4 t1+3	CRP + Leukocytes + Lymphocytes + Neutrophils + platelets + LMR + PLR + NLR + SII
Model 4.1 t1+3	CRP + Lymphocytes + LMR

Model 5 t1+3	CRP + NLR + PLR + LMR + SII + Calcium
Model 5.1 t1+3	CRP + PLR + LMR + Calcium
Model 6 t1+3	Sodium + LDH + CRP + Bilirubine_Total + Haemoglobin + Monocytes + eGFR_CKD_EPI + TSH + MCV + Creatinine + FT4 + Platelets + Glucose + potassium + Leukocytes + Calcium + Lymfocytes + Neutrophils + ALAT + NLR + PLR + LMR + Eosinophils + SII + ASAT
Model 6.1 t1+3	Sodium + LDH + CRP + Bilirubine_Total + Haemoglobin + Monocytes + eGFR_CKD_EPI + MCV + Creatinine + Platelets + Glucose + potassium + Leukocytes + Calcium + Lymfocytes + Neutrophils + ALAT + NLR + PLR + LMR + Eosinophils + SII + ASAT



Appendix G:
Results linear models

Group: *Inflammatory*

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.55	0.03	15.92	0.00
CRP	-0.11	0.04	-2.90	0.00
NLR	-0.07	0.09	-0.74	0.46
PLR	-0.01	0.05	-0.28	0.78
LMR	0.02	0.04	0.45	0.65
SII	0.02	0.09	0.24	0.81

Residual standard error: 0.48

p-value: 0.00

Group: *Liver Function*

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.56	0.03	15.51	0.00
Bilirubine total	-0.01	0.04	-0.37	0.71
ALAT	0.00	0.07	0.07	0.95
ASAT	0.01	0.07	0.17	0.87

Residual standard error: 0.50

p-value: 0.97

Group: *Kidney Function*

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.56	0.03	15.70	0.00
Creatinine	-0.10	0.07	-1.54	0.13
eGFR (CKD-EPI)	-0.09	0.07	-1.35	0.18

Residual standard error: 0.50

p-value: 0.31

Group: *Electrolytes*

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.55	0.03	15.82	0.00
Sodium	0.10	0.04	2.78	0.01
Potassium	0.02	0.04	0.46	0.65
Calcium	0.02	0.03	0.57	0.57

Residual standard error: 0.49

p-value: 0.03

Group: *Haematology*

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.55	0.03	16.36	0.00
Haemoglobin	0.15	0.03	4.31	0.00
Platelets	0.02	0.04	0.62	0.53
Leukocytes	-0.06	0.07	-0.89	0.37
Lymphocytes	0.06	0.04	1.77	0.08
Neutrophils	-0.08	0.07	-1.14	0.26
Monocytes	-0.01	0.04	-0.15	0.88
Eosinophils	-0.02	0.04	-0.58	0.57

Residual standard error: 0.47

p-value: 0.00

Group: *Endocrine Metabolic*

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.57	0.04	15.06	0.00
TSH	-0.04	0.04	-1.04	0.30
FT4	0.01	0.04	0.23	0.82
Glucose	-0.10	0.04	-2.66	0.01

Residual standard error: 0.49

p-value: 0.05562