Gezamenlijke opleiding UHasselt en KU Leuven

Universiteit Hasselt | Campus Diepenbeek | Faculteit Industriële Ingenieurswetenschappen | Agoralaan Gebouw H - Gebouw B | BE 3590 Diepenbeek

Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE 3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE 3500 Hasselt

Faculteit Industriële Ingenieurswetenschappen

master in de industriële wetenschappen: nucleaire technologie

Masterthesis

The effect of different segmentation methods on the performance of prognostic models for early-stage non-small cell lung cancer

Katleen Claesen

Scriptie ingediend tot het behalen van de graad van master in de industriële wetenschappen: nucleaire technologie, afstudeerrichting nucleair en medisch

PROMOTOR :
Prof. dr. Brigitte RENIERS

PROMOTOR :
Prof. dr. Liesbet MESOTTEN

BEGELEIDER :
drs. Jill MEYNEN

2024
2025

Faculteit Industriële Ingenieurswetenschappen

master in de industriële wetenschappen: nucleaire technologie

Masterthesis

The effect of different segmentation methods on the performance of prognostic models for early-stage non-small cell lung cancer

Katleen Claesen

Scriptie ingediend tot het behalen van de graad van master in de industriële wetenschappen: nucleaire technologie, afstudeerrichting nucleair en medisch

PROMOTOR :

Prof. dr. Brigitte RENIERS

PROMOTOR :

Prof. dr. Liesbet MESOTTEN

BEGELEIDER :

drs. Jill MEYNEN

Preface

This master’s thesis represents the final step in obtaining the degree of Master of Nuclear Engineering Technology at UHasselt and KU Leuven. The choice of this topic comes from my interest in the medical applications within the nuclear sector. Thanks to the collaboration between Ziekenhuis Oost-Limburg (ZOL) and the University of Hasselt, I gained insight into radiomics and prognostic models for lung cancer. Lung cancer is the most common cause of cancer-related death worldwide. Therefore, I am proud to have contributed to research aimed at improving prognostic models. This project has also made me realize the importance of being a multidisciplinary engineer. My background in nuclear technology helped me interpret nuclear imaging, while my knowledge of various programming languages facilitated learning a new programming language, R, necessary for understanding the prognostic models. Additionally, I broadened my skill set by learning about the biological aspects of cancer, especially lung cancer, and radiomics.

First and foremost, I would like to thank my external supervisor, Prof. Dr. Liesbet Mesotten from ZOL, for her indispensable expertise and valuable advice. Thanks to her, I was able to deepen my knowledge. Her feedback and innovative ideas made this master’s thesis a success.

Next, I would like to thank my internal supervisor, Prof. Dr. Brigitte Reniers, for providing me with essential knowledge about the medical nuclear field and for igniting my interest in this discipline.

I would also like to express my gratitude to my supervisor, PhD student Jill Meynen. She was always available with feedback, advice, and positivity. I especially want to thank her for joining me on the trip to Amsterdam, where we enhanced our radiomics know-how. It was a pleasure working with her.

Furthermore, I want to thank Prof. Dr. Ronald Boellaard for welcoming me to Amsterdam UMC and guiding me during my first steps in the world of radiomics. I am also grateful for his provision of the ‘Accurate’ and ‘Radiomics’ tools and for supplying the prognostic models.

I also want to thank ChatGPT-4o for helping me write code in R and rephrase sentences in this thesis to make them sound more academic.

Finally, I wish to thank my family and partner for their unconditional support throughout my entire education. They have inspired me to pursue an engineering degree and strengthened my interest in the medical field. Their encouragement and support have made it possible for me to be where I am today.

Table of contents

List of tables

List of figures

List of abbreviations

^{18}F-FDG	2-[18]F fluorodeoxyglucose
β^+	positron
AI	artificial intelligence
ATP	adenosine triphosphate
AUC	area under the curve
BMI	body mass index
CT	computed tomography
DNA	deoxyribonucleic acid
FN	false negatives
FP	false positives
FPR	false positive rate, 1-specificity
LOR	line of response
MV	majority vote
NK	natural killer cells
NSCLC	non-small cell lung cancer
OS	overall survival
OXYPHOS	oxidative phosphorylation
PET	positron emission tomography
PET/CT	positron emission tomography-computed tomography
PFS	progression-free survival
PMT	photomultiplier tube
pRB	retinoblastoma protein
pw	pairwise elimination
rf	random forest
rfe	recursive feature elimination
ROC	receiver operating characteristics
SCLC	small cell lung cancer
SUV	standardized uptake value
TLG	total lesion glycolysis
TN	true negatives
TP	true positives
TPR	true positive rate, sensitivity
TTP	time to progression
UMC	universitair medisch centra
uni	univariate selection
VOI	volume of interest
ZOL	Ziekenhuis Oost-Limburg

Abstract

Lung cancer is the leading cause of cancer-related death worldwide. Early diagnosis and accurate staging are essential to improve patient survival. Radiomics, a method for extracting features from medical images, can improve this process by identifying patterns invisible to the human eye. Segmentation defines the analyzed region and may influence radiomic features. This master's thesis examines whether the choice of PET segmentation method affects the performance of prognostic models for non-small cell lung cancer (NSCLC).

A total of 121 patients with stage I-IIIA NSCLC were selected from the NCT03736993 trial (ProLUNG study) and the NCT02024113 trial. PET/CT images were segmented using three methods (SUV4, MV2, and an AI-based method LIONZ). Radiomic features were subsequently extracted from all the different segmentations. Preselected features were used as input for different prognostic models that were either based on logistic regression or random forest.

Model performance was compared across the segmentation methods, as well as between the models, using the area under the curve (AUC) values of the receiver operating characteristic curves. A Friedman test did not find any statistically significant differences between the different segmentation methods. Differences in model performance were mainly caused by the applied feature selection. The TLG-only model showed the highest overall reliability (AUC 0.76 ± 0.14), while the most optimal combination included TLG, SUVmax, and DmaxBulk with MV2 segmentation (AUC 0.77 ± 0.16).

Abstract in Dutch

Longkanker is wereldwijd de dodelijkste vorm van kanker. Voor een betere overleving is een vroege diagnose en nauwkeurige stadiëring essentieel. Radiomics, een methode om parameters uit medische beelden te halen, kan dit proces verbeteren door patronen te identificeren die voor het menselijk oog onzichtbaar zijn. Segmentatie bepaalt het geanalyseerde gebied, wat invloed kan hebben op de radiomic parameters. Deze masterproef onderzoekt of de keuze van de PET-segmentatiemethode de prestaties van prognostische modellen voor niet-kleincellige longkanker (NSCLC) beïnvloedt.

In totaal werden 121 patiënten met stadium I-IIIA NSCLC geselecteerd uit de NCT03736993-trial (ProLUNG-studie) en de NCT02024113-trial. PET/CT-beelden werden gesegmenteerd met drie methoden (SUV4, MV2 en een AI-gebaseerde methode LIONZ). Radiomic parameters werden geëxtraheerd uit alle segmentaties. Vooraf geselecteerde parameters werden gebruikt als input voor prognostische modellen gebaseerd op logistische regressie of random forest.

Modelprestaties werden vergeleken tussen de segmentatiemethoden en modellen met behulp van de area under the curve (AUC)-waarden van de receiver operating characteristic curves. Een Friedman test toonde geen statistisch significante verschillen tussen de verschillende segmentatiemethoden. Verder toonde het model met alleen TLG de hoogste globale betrouwbaarheid (AUC 0,76 ± 0,14), terwijl de meest optimale combinatie het model met TLG, SUVmax en DmaxBulk voor MV2-segmentatie is (AUC 0,77 ± 0,16).

1 Introduction

Lung cancer is the most common type of cancer and the leading cause of cancer-related death worldwide [1]. Based on the location, size, and metastasis of the cancer, a stage is given [2]. Staging is essential in order to make an accurate prognosis. To detect and stage tumors in a non-invasive way, positron emission tomography (PET) together with computed tomography (CT) can be used.

In recent years, radiomics has gained popularity in cancer research. Radiomics is a type of quantitative image analysis where a large number of features are extracted from medical images. These features can be used to identify patterns that cannot be observed by the human eye, with the help of a machine and/or deep learning model. In this way, radiomics has the potential to offer insights into tumor biology and improve the decision-making processes related to cancer detection, staging, prognosis, and treatment [3].

The radiomic workflow consists of five main steps. The first step is image acquisition. High-quality images can be obtained through commonly used scans available in the hospital, such as CT and PET/CT scans. After image acquisition, tumor segmentation, the outlining of the tumor, is performed to obtain a volume of interest (VOI). The third step in the process entails the extraction of radiomic features from the VOI. As approximately 500 different radiomic features are extracted, the fourth step involves the selection of the most significant features for further analysis. Feature selection can either be performed before or simultaneously with step five (performance testing) [4], [5]. In this last step, different models are used to analyze the radiomic features and develop prognostic cancer models.

This master's thesis will focus on the development of prognostic models for early-stage non-small cell lung cancer (NSCLC). The data used in this study comes from the ProLUNG study at Ziekenhuis Oost-Limburg (ZOL), located in Genk. The ProLUNG study is funded by 'Kom op tegen Kanker'. Additional data were obtained from the NCT02024113 trial, performed at the Limburg PET-Center in Hasselt. In total, 121 patients from the two studies combined were selected and included in this study for radiomic analysis. Data from the ProLUNG trial was used for prognostic model training, while data from the NCT02024113 trial was used for external validation of the trained models. All patients were diagnosed with stage I–IIIA NSCLC, or early-stage NSCLC, and underwent a lobectomy as part of their standard-of-care treatment, along with a PET/CT scan.

Using the ACCURATE tool, tumor segmentation was performed on the patient's PET/CT scans with three methods: standardized uptake volume (SUV) and majority vote (MV), two threshold-based methods; and LIONZ, an artificial intelligence (AI)-based segmentation method. Following segmentation, around 500 radiomic features were extracted using the RADIOMICS tool. Both tools are developed by the research team of Prof. Dr. Ronald Boellaard (Amsterdam, UMC). Seven different prognostic models, both simple and more complex, were tested for the three segmentation methods. The complexity of each model is determined by the feature selection approach. The simple models only included conventional PET metrics like total lesion glycolysis (TLG), a combination of TLG and SUVmax, and a combination of TLG, SUVmax, and DmaxBulk. The intermediate models included features using different feature elimination methods, such as pairwise elimination (pw), univariate selection (uni), or recursive feature elimination (rfe), but remained quite simple because of the use of logistic regression. The most complex model tested was a machine learning model called random forest. Cross-validation was used to test the accuracy of the different models across the different segmentation methods. This methodology is described in more detail in Chapter eight.

The first objective of this study is to investigate whether the choice of tumor segmentation method impacts the performance of different prognostic models for early-stage NSCLC. A second objective is to investigate which model, and consequently which features, are preferred to create the most accurate prognostic model for early-stage NSCLC.

The present thesis is divided into chapters, each dedicated to a different aspect of the subject of this study. Chapter two explains the basics of cancer and the cancer hallmarks, while Chapter three provides more detail about the grading and staging of tumors, which play an important role in making predictions. Chapter four provides more in-depth information about lung cancer.

Chapter five explains the most commonly used imaging techniques to diagnose lung cancer. X-ray imaging is explained in this chapter, as this technique is used in CT, which is also discussed in depth. Further, Chapter five provides information about PET scans. This includes information about the radioactive tracer ^{18}F-FDG, the PET camera, and the use of PET/CT scans.

Chapter six provides a deeper understanding of radiomics. This chapter includes radiomic applications as well as a thorough discussion of all the different steps of the radiomic workflow, including segmentation, feature selection, and performance testing of the prognostic models.

The last chapter that provides theoretical information is Chapter seven, which elaborates on the different analysis methods that were used, such as the Friedman and the DeLong tests.

While Chapter eight discusses the methodology, Chapter nine focuses on the results. Chapter ten provides a thorough discussion of these results and reflects on limitations and future research possibilities. At last, the conclusion on both objectives can be found in Chapter eleven.

2 Cancer

After cardiovascular diseases, cancer is worldwide the most common cause of death. Cancer is a disease that can manifest in different ways, but the underlying principle remains the same: uncontrolled cell division. Cells grow and divide in the body. Abnormal tissue growth can occur when cells develop abnormalities due to genetic changes. These genetic changes can be caused by a variety of factors such as ionizing radiation, asbestos, ultraviolet light, and tobacco. An overview of the impact of gene mutation is shown in Figure 1 [5].

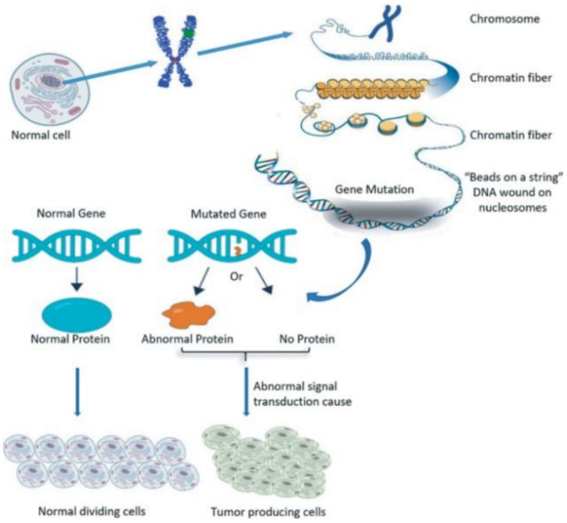


Figure 1: An overview of the cell cycle for normal dividing cells compared to tumor-producing cells, showing the impact of gene mutation [5, p. 4].

As shown in Figure 1, deoxyribonucleic acid (DNA) is packaged in a highly condensed manner, forming chromosomes composed of chromatin fibers. DNA contains genes that encode proteins essential for regulating growth and cell division, involving complex signaling networks that include extracellular signaling molecules such as growth factors. These molecules bind to specific receptors on the cell surface, initiating signal transduction pathways that control cellular division through the cell cycle [6]. Importantly, signaling only occurs in response to the right conditions, ensuring that cell proliferation is tightly coordinated to prevent uncontrolled growth. However, mutations, caused by various factors, can occur within the genes, resulting in the production of abnormal proteins or, in some cases, the complete loss of protein expression. This can interfere with key pathways, leading to the dysregulation of cell growth and division, and ultimately promote uncontrolled proliferation, transforming normal dividing cells into tumor-producing cells [5].

A tumor is defined as a group of cells, either benign or malignant, with dysregulated growth control, and thus, uncontrolled cell division [5]. Benign tumors, also called non-cancerous tumors, are considered innocuous as they are unlikely to spread to other parts of the body. Malignant tumors, also called cancer or cancerous tumors, are considered aggressive as these can spread (metastasize) to other parts of the body to eventually form secondary tumors [5], [7], [8].

2.1 Cancer hallmarks

In 2000, Hanahan and Weinberg described 'The Hallmarks of Cancer' [9]. These six functional capabilities, acquired by human cells, are crucial for to develop malignant tumors [10]. However, in 2011, two emerging hallmark capabilities and two enabling characteristics were added [11]. The emerging hallmarks are referred to as such because, at the time they were published, they were still considered provisional. However, at this time, the emerging hallmarks are also considered core hallmarks [10]. Enabling characteristics are molecular and cellular mechanisms that can lead to the acquisition of the hallmarks. The eight hallmarks and two enabling characteristics of cancer are shown in Figure 2.

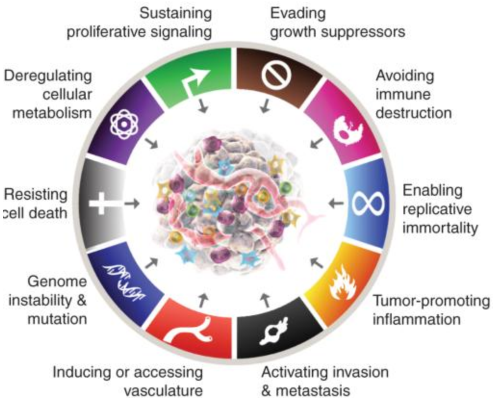

Sustaining proliferative signaling
Evading growth suppressors
Avoiding immune destruction
Enabling replicative immortality
Tumor-promoting inflammation
Activating invasion & metastasis
Inducing or accessing vasculature
Genome instability & mutation
Resisting cell death
Deregulating cellular metabolism

Figure 2: The eight hallmarks of cancer and two enabling characteristics [10, p. 32].

2.1.1 Initial hallmarks

The capacity for **sustaining proliferative signaling** represents the most fundamental hallmark of cancer. In normal cells, cell division is controlled and occurs in response to specific extracellular signals, such as growth factors, extracellular matrix components, or signals from neighboring cells [9]. These signals activate receptors on the cell surface and trigger pathways that allow the cell to proliferate. Malignant cells, however, acquire the ability to undermine these signals through various strategies. First, these cells often produce growth factors themselves, a process known as autocrine signaling, which makes the growth factors of other cells redundant [9]. Second, malignant cells might alter the signal receptors by increasing the number of receptors on the cell surface or by overexpression so that these will provoke a response even in the presence of minimal signals [9], [11]. Third, malignant cells frequently activate intracellular signals by mutating, for example, proteins that regulate growth signals [11]. These mutations disable the normal mechanisms that would prevent excessive signaling, allowing for uncontrolled cell growth. All these changes result in uncontrolled cell growth and increased cell division.

In order to stimulate cancer growth, malignant cells are not only able to induce and sustain growth-stimulating factors but also to **evade growth-inhibiting factors or growth suppressors**. Two of the most important tumor suppressor proteins are the retinoblastoma protein (pRb), encoded by the retinoblastoma gene, and the tumor protein p53, encoded by the tumor protein p53 gene. The pRb protein integrates signals from both extracellular and intracellular sources and can prevent the expression of genes required for DNA synthesis and cell cycle advancement [9], [11]. In contrast, p53 responds only to intracellular signals. If the levels of different factors, such as glucose, oxygen, or growth-promoting signals, are suboptimal or excessive genome damage is present, p53 will halt the cell cycle until normal conditions are restored [9], [11]. While pRb and p53 restrict cell proliferation, both operate via distinct yet complementary mechanisms, forming a robust barrier against malignant transformation [11]. Functional loss or inhibition of pRb or p53, through mutation or pathway disruption, is frequently observed in cancers and subsequently results in uncontrolled cellular proliferation [9], [12], [13].

A third hallmark is **enabling replicative immortality**. Normal cells only pass a limited number of complete cell cycles due to senescence and apoptosis. This is mainly regulated by telomeres. Telomeres are protective repeated segments at the ends of chromosomes that are added by the telomerase enzyme. When a cell undergoes division, the telomeres shorten. If the telomeres become too short, cell viability is compromised, ultimately resulting in cell death. Malignant cells, however, show increased telomerase levels, leading to less telomere shortening and improved cell viability [14].

Malignant cells can spread to other tissues in the body through the processes of **invasion and metastasis**, which is the fourth hallmark. While invasion is defined as the expansion of malignant cells to nearby tissue, metastasis is defined as the process by which a part of the primary tumor detaches and spreads to other parts of the body to establish a secondary tumor [15].

Inducing or accessing vasculature, also known as angiogenesis, is the fifth initial hallmark. Angiogenesis refers to the formation of new blood vessels originating from existing ones. In adults, angiogenesis is inactive if no wound healing is going on. However, malignant cells need new blood vessels to supply oxygen and nutrients, among other things. Consequently, when malignant cells start to develop, angiogenesis is induced [14].

The sixth initial hallmark of cancer is **resisting cell death**. Apoptosis, also known as cell death or cell suicide, is the cellular process that eliminates damaged or old cells. Apoptosis is one of the most powerful barriers against the development of cancer. However, malignant cells, along with the loss of tumor suppressor proteins and telomerase overexpression, have developed mechanisms to evade apoptosis, thereby allowing damaged cells to survive [14].

2.1.2 Emerging hallmarks

In addition to the six initial hallmarks, there are two emerging hallmarks. The first emerging hallmark is **avoiding immune destruction**. The human body has an immune system that serves as the body's primary defense. This system can detect transformed or damaged cells and prevent their proliferation. However, malignant cells can develop a mechanism to bypass the immune system, thereby facilitating tumor formation [14]. Malignant cells can, for example, avoid immune destruction by producing transforming growth factor – beta (TGF-β) [11]. In the tumor environment, this growth factor suppresses natural killer (NK) cells, which detect and kill abnormal cells. This allows cancer cells to evade immune surveillance, resulting in the survival and growth of variants that are weakly recognized by NK cells [16].

A secondary emerging hallmark of cancer is the **deregulation of cellular metabolism**. As previously mentioned, malignant cells are characterized by increased proliferation and cell division, thereby requiring significant amounts of energy and other nutrients. To meet these high energy demands, malignant cells must undergo metabolic reprogramming [11]. Figure 3 shows the difference between the metabolism of differentiated tissue and malignant cells.

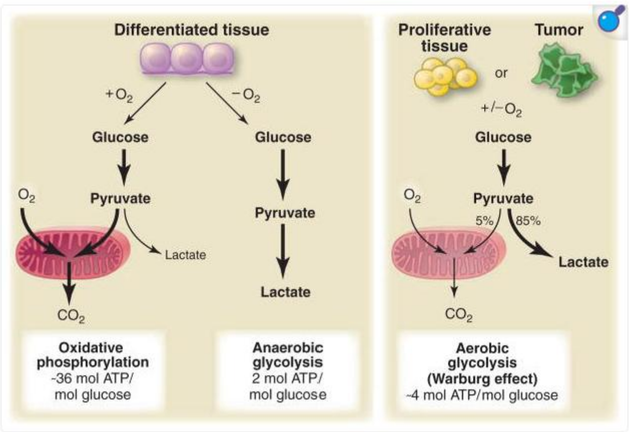


Figure 3: The difference between the metabolism of differentiated tissues and malignant cells. This figure illustrates different pathways to generate ATP, such as oxidative phosphorylation, anaerobic glycolysis, and the Warburg effect [17, p. 10]. ATP: adenosine triphosphate.

The energy production of non-proliferating, differentiated tissue depends on oxygen availability. In the presence of oxygen, cells mainly rely on oxidative phosphorylation (OXPHOS) to meet energy demands. During this process, glucose is converted to pyruvate via glycolysis, which is then oxidized in the mitochondria [11], [14]. The OXPHOS is the most efficient pathway to produce energy, as it produces 36 adenosine triphosphate (ATP) molecules per glucose molecule [17].

When oxygen is limited, energy production in differentiated tissue depends on anaerobic glycolysis instead of the OXPHOS. In this case, pyruvate is converted into lactate rather than being transported to the mitochondria for oxidation. This pathway, however, only yields two ATP molecules per glucose molecule and is, therefore, less efficient [17].

As shown in Figure 3, malignant cells predominantly convert pyruvate into lactate despite the presence of oxygen. This phenomenon is called the aerobic glycolysis or the Warburg effect [11]. Nevertheless, this pathway will only yield around four ATP molecules per glucose molecule, which is much lower compared to the OXPHOS. Therefore, to sustain the elevated energy demands of malignant cells, an increased glucose uptake is required to enable proliferation [11], [17].

2.1.3 Enabling characteristics

The acquisition of the hallmarks is facilitated by two enabling characteristics. The first enabling characteristic is **tumor-promoting inflammation**. This means that tumors are infiltrated with immune cells, in large or small amounts, in a manner analogous to the body's response to inflammation triggered by infections. The presence of immune cells could be an indication that the body attempts to eliminate the tumor. However, evidence suggests that cancer has developed mechanisms to escape the body's immune response. In contrast to the expected immune response, whereby infiltrated immune cells attack the tumor, malignant cells manipulate immune cells to support tumor growth. This phenomenon occurs as malignant cells are able to take over the inflammatory response, thereby converting immune cells into tumor-promoting cells. These cells secrete growth factors that facilitate the proliferation, survival, and metastasis of the tumor rather than inducing apoptosis. Consequently, while inflammation is generally considered a defense mechanism, in this context, it can facilitate the progression of the cancer by supporting it rather than eliminating it [11], [18].

The second, perhaps most well-known, enabling characteristic is **genome instability and mutation**. Genomic instability is defined as the increased tendency of genomic changes in genes that contribute to tumor growth control. There are different types of mutations, such as substitutions, deletions, insertions, and translocations. Most mutations are non-cancer-related and can be resolved by genome surveillance and DNA repair mechanisms. Defects in these systems can, therefore, lead to an increased risk of cancer formation. These mutations create an environment where genetic instability accelerates the tumor progression rate. As cells acquire advantageous traits, they gain the ability to survive in conditions where normal cells would die, enabling them to proliferate uncontrollably. This uncontrolled cell division is a hallmark of cancer, and over time, the accumulation of mutations increases the chances of acquiring further genetic changes. These additional changes can enhance the aggressive behavior of the cancer, such as its ability to invade surrounding tissues and resist signals that would normally induce cell death [11], [19].

3 Grading and staging of tumors

Grading is performed to determine if a tumor is benign or malignant. Grading, indicated by the letter ‘G’, describes the degree to which malignant cells differ in appearance from normal cells. If ‘G’ is followed by a low number, it indicates that malignant cells bear a resemblance to normal cells. The higher the number, the more malignant cells will differentiate from normal cells and the quicker they grow. If ‘GX’ is indicated, it signifies that the grade remains undetermined [14], [20], [21].

To determine the severity of a malignant tumor, staging is used. Cancer staging is based on the location, size, and the presence of metastasis [2]. One of the most commonly used staging methods is the TNM classification. The letter T provides information about the size and extent of the original, or primary, tumor. The letter N indicates whether malignant cells have spread to nearby lymph nodes, while the letter M denotes the presence of metastasis in other parts of the body. Each letter is always followed by a number to provide more details about the cancer. A value of zero always indicates the absence of a specific event, while a value of one or higher indicates the presence of an event. The higher the value, the higher the degree of expansion. For example, N0 indicates no spread to the lymph nodes, N1 indicates spread to a small number of nearby lymph nodes, and N2 indicates spread to a larger number of lymph nodes or more distant nodes. If an X follows one of the letters, it indicates the absence of information due to its inability to be quantified or assessed [7], [20], [21].

The TNM values can be combined to determine an overall stage. These are described with Roman numbers from I to IV. Stage I is characterized as the most treatable stage of the disease, while stage IV is characterized as the most severe stage, as the cancer is widely spread, resulting in a poor prognosis [2], [14], [21].

4 Lung cancer

Lung cancer is the most common type of cancer and the leading cause of cancer-related death worldwide [1]. One of the main reasons for this high mortality rate is that the early stages of lung cancer often appear without noticeable symptoms. When symptoms, such as coughing or fatigue, do occur, they are often mistaken for other medical conditions [22]. Subsequently, this leads to a delayed diagnosis and treatment, resulting in a low survival rate. To improve current survival rates, early lung cancer detection is essential [23]. Table 1 shows that the 5-year survival rate increases rapidly when NSCLC is detected in an early stage without metastasis.

Table 1: 5-year relative survival rates for different stages of non-small-cell lung cancer (2012-2018) [24].

Stage	5-year relative survival rate
Localized	65%
Regional	37%
Distant	9%

This table illustrates a substantial decline in the survival rate between localized tumors and regional metastasized tumors. This decrease is just as significant between regional and distant metastasized tumors. The need for good detection, staging, and prognosis prediction is thus inevitable.

Lung cancer is categorized into two main types: small cell lung cancer (SCLC) and NSCLC. The main part, about 80%, is classified as NSCLC. NSCLC can be further subdivided into three distinct types based on tumor histology: adenocarcinoma, squamous cell carcinoma, and large-cell carcinoma [25].

Lung cancer detection can be performed in several ways. Chest X-rays, CT scans, and tumor biopsies are tests that can be used to both detect and stage lung cancer [26], [27]. NSCLC is usually staged using the TNM classification as explained in Chapter 3. Staging is very important in order to formulate an accurate prognosis. Early-stage lung cancer includes stage I-IIIA, where, in the severest case, the tumor has spread no further than the respiratory system itself or the nearby lymph nodes. For these stages of cancer, surgery is always included in the treatment plan, possibly in combination with radio- or chemotherapy, depending upon the extent of cancer development [28].

5 Imaging techniques

Imaging techniques are fundamental to the diagnosis and staging of lung cancer. This chapter provides an overview of three commonly used modalities in clinical practice: X-ray imaging, computed tomography (CT), and positron emission tomography (PET).

5.1 X-ray imaging

X-rays are a type of electromagnetic radiation, which are most often described as packages of electromagnetic energy called photons. Generally, X-rays are produced by an external electron that interacts near the nucleus [29]. There are two types of X-rays: Bremsstrahlung and characteristic radiation. Both types have different energy spectra, as shown in Figure 4.

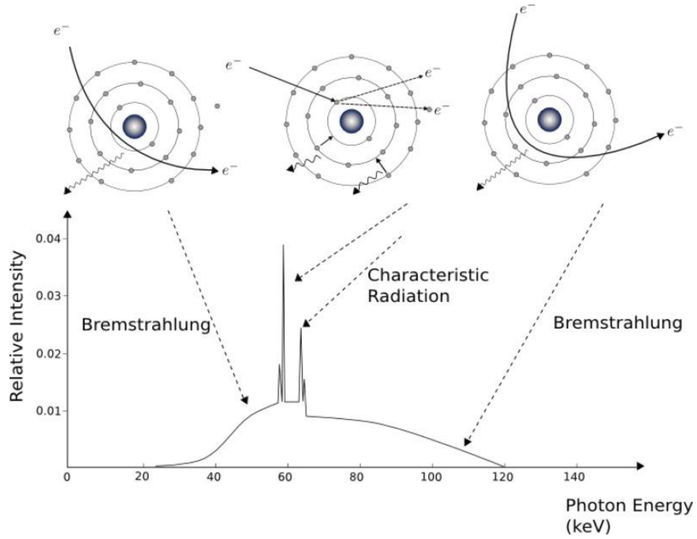


Figure 4: X-ray spectrum produced by Bremsstrahlung and characteristic radiation. This figure shows the process to produce both types of X-rays near the nucleus [30, p. 126]. keV: kiloelectron volt.

Around 80% of all X-rays are Bremsstrahlung X-rays [31]. These are produced when a high-energy electron passes near the nucleus of an atom and is slowed down as its path is deflected. The kinetic energy lost during this process is eventually emitted as an X-ray. The energy spectrum of Bremsstrahlung X-rays exhibits a continuous spectrum as it can range from approximately zero to the initial electron energy. The remaining 20% are characteristic X-rays. These are produced by the collision of an incoming high-energy electron with an inner-shell electron. This collision results in the ejection of the inner-shell electron and the creation of a vacancy. Subsequently, the vacancy is filled by an electron from an outer shell, resulting in the loss of binding energy for the orbital electron. This energy difference is emitted in the form of a characteristic X-ray. Due to the presence of a fixed energy difference between shells, X-rays exhibit a specific energy value for each material type, resulting in a discrete spectrum [29], [30].

5.1.1 X-ray tube

For X-ray-based imaging, controlled X-ray production is required. This production happens in an X-ray tube, or X-ray source, which is shown in Figure 5.

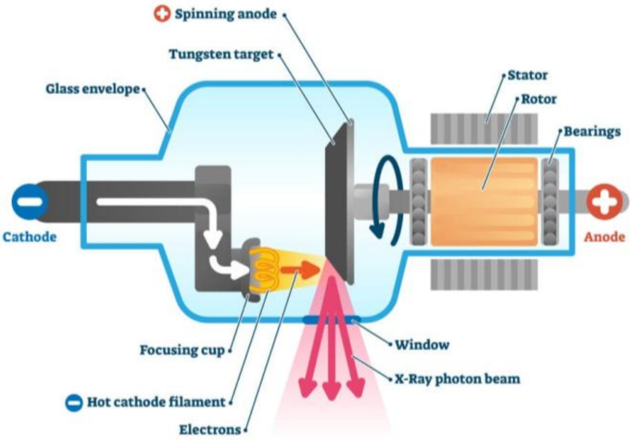


Figure 5: The different components of an X-ray tube [32].

The generation of the required electrons is facilitated by the application of an electric current to a cathode filament within the X-ray tube, thereby heating the filament. When the filament reaches an elevated temperature, electrons start to break free. The free electrons are accelerated to the anode side of the X-ray tube by an electric potential difference. To ensure that these electrons reach the target and do not interact with air molecules, a vacuum chamber is placed around the filament and the target. Upon reaching the anode, the free electrons interact with the target's atoms, producing heat and X-rays. The high degree of conversion of kinetic electron energy into heat necessitates the rotation of the target to dissipate the heat. The resulting X-rays are emitted as a beam that traverses the chamber through the window towards the patient [30], [33].

5.1.2 Imaging

To obtain images, X-rays are transmitted through the patient's body. Depending on the density that the X-rays pass, different degrees of attenuation occur. Attenuation is defined as the reduction in intensity of an X-ray beam as it passes through tissue [34]. Every tissue exhibits a different density, resulting in different levels of beam attenuation. The different types of tissues and corresponding colors are shown in Figure 6.

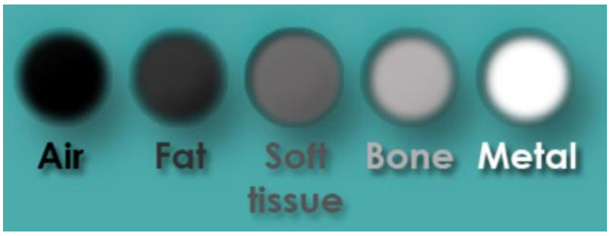


Figure 6: Different types of tissues and their representation on X-ray images [35].

As shown in Figure 6, the lower the density, the darker the color. A dark color indicates that most X-rays pass through the tissue. Vice versa, the higher the density, the lighter the color, as more X-rays are absorbed by the tissue, resulting in a lower number that ultimately reaches the detector. After passing through the body, the X-rays are captured by detectors. Based on the remaining intensity of the X-rays, an attenuation map can be reconstructed [36]. In the beginning, silver halide crystal films were used as detectors [37]. The parts of the film that were irradiated would turn darker after processing. Nowadays, most hospitals use flat panel detectors, which are a planar array of electronic detectors. The use of these digital detectors offers several advantages, including the direct readout of images and an improved spatial resolution [30], [35], [38].

5.2 Computed tomography

CT is a medical imaging technique that provides anatomical information [39]. A visual representation of a CT scan is shown in Figure 7.

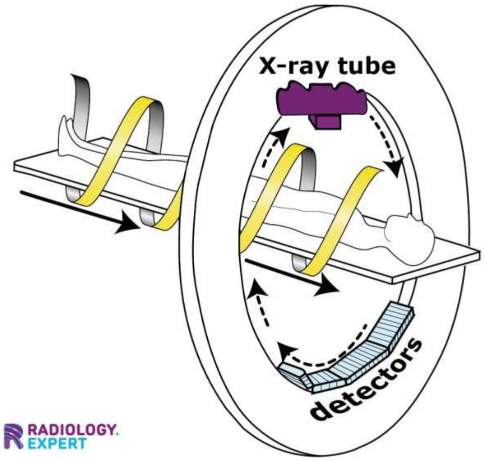


Figure 7: Visual representation of the main components and movements of a computed tomography (CT) scan [40].

The principle of a CT is similar to X-ray imaging, as explained in *5.1 X-ray imaging*; however, there are two major differences with traditional X-ray imaging. The first difference is that the X-ray tube and digital detectors of a CT scan quickly rotate 360°, thereby generating projections at various angles [34]. The part that contains the X-ray tube and the detectors is called the gantry. The second difference is that simultaneously with the rotation, the patient bed is translated through the gantry. This ensures that each rotation is performed on a different part of the body [36].

5.2.1 Reconstruction

After acquiring all the different projections, readable CT slices must be reconstructed. This process is shown in Figure 8.

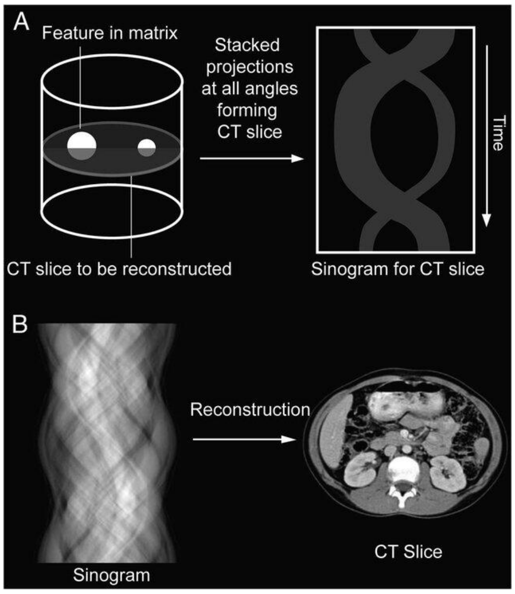

A
Feature in matrix
Stacked projections at all angles forming CT slice
Time
CT slice to be reconstructed
Sinogram for CT slice
B
Reconstruction
Sinogram
CT Slice

Figure 8: Process of the reconstruction of a CT scan, starting from a single projection to a whole CT slice. (A) The projection of scanned material and its corresponding sinogram; (B) A whole CT slice reconstructed by backprojection, starting from a sinogram [41]. CT: computed tomography.

The acquired projections of one rotation can be put together in what is called a sinogram. A sinogram shows the different detector readings, with time progressing from the top to the bottom of the image [41]. It mainly consists of sinus-like lines put together. The acquired sinogram can then be reconstructed into a readable CT slice through different methods, with backprojection being the most commonly used [38]. These reconstructions eventually result in 2D images, which can be converted to 3D images by stacking different CT slices on top of each other [36].

5.3 Positron emission tomography

PET is a nuclear imaging modality that offers insights into the metabolic activity and molecular characteristics of tumor tissue, among other things [42]. It is a non-invasive method to detect and stage lung cancer. A positron (β^+)-emitting radioactive tracer is intravenously administered to the patient, where it will accumulate in areas it has an affinity for. Detectors surrounding the patient will eventually detect the emitted radiation from the injected tracer.

5.3.1 ^{18}F-FDG

A radioactive tracer in PET imaging consists of a carrier molecule combined with a radioactive nuclide. ^{18}F-labeled fluoro-2-deoxyglucose (^{18}F-FDG) is the most commonly used radiotracer for PET imaging. It is a glucose analog where the OH-group at the 2-C position is substituted by (radioactive) 18-fluorine, as shown in Figure 9Figure 9.

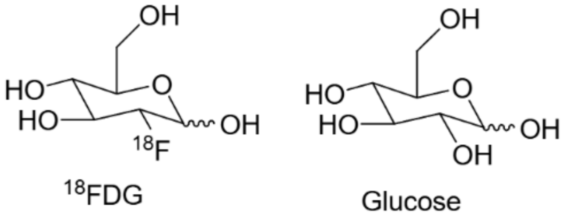


Figure 9: The atomic structures of both ^{18}F-FDG and glucose [43, p. 42].

As ^{18}F-FDG acts as a glucose analog, it will enter regions with a high glucose demand, such as the brain, bladder, inflammatory lesions, and tumors. As seen in Chapter *0*, tumors have a higher glucose uptake because of the Warburg effect. Once ^{18}F-FDG is taken up into a cell, it is subjected to phosphorylation, resulting in the formation of ^{18}F-FDG-6-phosphate. Due to the absence of an OH-group, processing of this compound through glycolysis is blocked, and the compound becomes metabolically trapped [44]. Due to this trapping, the glucose flux and the tumor's location can be determined [39].

Besides being a glucose analog, there are additional reasons why ^{18}F-FDG is a preferred radiotracer in PET imaging. For example, ^{18}F is a radionuclide with a half-life of 110 minutes. This duration is sufficient for the production of ^{18}F-FDG via a cyclotron, its subsequent transportation to the hospital, and its quantification in the patient. Moreover, it is not retained within the patient's body for an extended period, as it remains a radioactive substance [45].

5.3.2 Principle of a PET scan

All injected radioactive tracers emit positrons because they contain an atom with an unstable nucleus, typically due to an excess of protons, such as ^{18}F. In this process, a proton is converted into a neutron, resulting in the emission of a neutrino and a positron, as shown in (1) [46].

$$^{A}_{Z}X \rightarrow \ ^{A}_{Z-1}Y + \beta^{+} + \upsilon \tag{1}$$

Equation (1) represents the mother nucleus as *X*, with *A* the atomic mass and *Z* the number of protons. Y represents the daughter nucleus, while β^{+} represents a positron and υ represents a neutrino. A neutrino is uncharged and has no mass. Therefore, it doesn't interact with surrounding tissue after it is emitted. On the contrary, a positron has the same mass as an electron but has the opposite charge. As it travels through matter, a positron loses its kinetic energy and will combine with a free electron in a process called "annihilation".

The pathway from positron-emitting nuclear decay to annihilation is shown in Figure 10 [34].

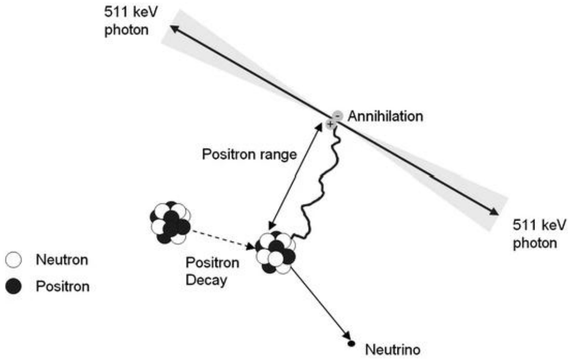


Figure 10: The process of positron-emitting decay and positron-electron annihilation [46, p. 5]. keV: kiloelectron volt.

Upon annihilation, two 511 keV photons are emitted back-to-back at a 180° angle. These photons will eventually be detected by a ring of PET detectors surrounding the patient, as shown in Figure 11.

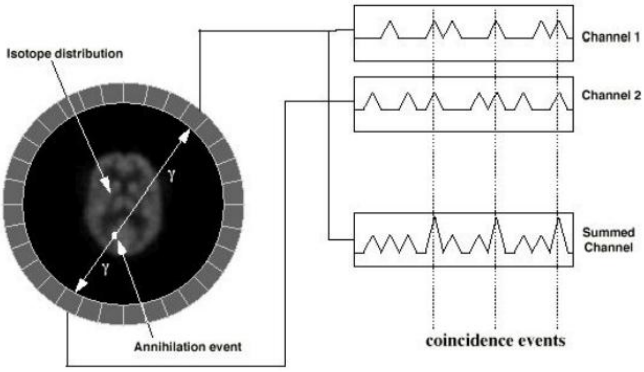


Figure 11: A visual representation of the capture of an annihilation event and the coincidence condition [34, p. 3].

When the free electron is not at rest before annihilation occurs, photons might be emitted at an angle of less than 180° [47]. To ensure that detection is limited to the true back-to-back photons, two conditions must be met. First, the two detectors on opposite sides of the detector ring must both detect a photon. The imaginary line between these two detectors is called the line of response (LOR). Second, the two photons must be detected within a coincidence window of about 5 ns [46]. The intersection of multiple LORs that also meet the coincidence window condition indicates the location of tracer accumulation and, thus, possibly a tumor.

5.3.3 PET camera

The working principle of the PET camera is shown in Figure 12.

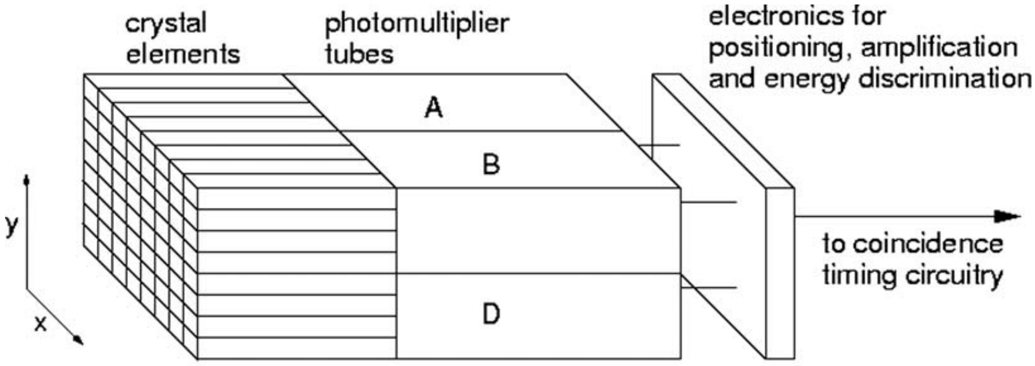

crystal elements
photomultiplier tubes
electronics for positioning, amplification and energy discrimination
A
B
D
y
x
to coincidence timing circuitry

Figure 12: A schematic representation of the main components of a PET camera [48, p. 30].

The initial step in a PET scan involves the emission of two back-to-back 511 keV photons, as explained in section *5.3.2 Principle of a PET scan.* Following Figure 12 from left to right, these photons are detected by crystal elements known as scintillators. The most commonly used scintillator materials in PET systems are inorganic, high-density crystals, such as bismuth germanate (BGO) or lutetium oxyorthosilicate (LSO) [49]. Upon interaction with the scintillator material, a photon converts into light photons. To ensure the collection of all emitted light, a reflector is placed around the scintillator. The emitted light is then amplified and converted into an electrical signal by the photomultiplier tube (PMT) array, which is optically coupled to the scintillator crystal [49]. As scintillator crystals are expensive, it is common practice to pair a single scintillator with multiple PMTs [46], [47]. The resulting electrical signals are eventually used to calculate the annihilation position and verify if the two detected photons are within the coincidence timing window. To accurately reconstruct the distribution of the radiotracer within the body, a substantial number of detected photon pairs and corresponding LORs is required [34].

5.4 PET/CT

Medical imaging with ^{18}F-FDG PET/CT is proven to be the most convenient method to diagnose and stage cancer due to its non-invasive character and its dual functionality. While a CT scan provides anatomical information, an ^{18}F-FDG PET scan provides metabolic information about a lesion of interest [39]. By fusing both images, a more accurate diagnosis and staging can be performed due to improved image interpretations [50].

This study will examine prognostic models for early-stage NSCLC using PET parameters extracted from ^{18}F-FDG PET/CT images.

6 Radiomics

Radiomics is a rapidly evolving discipline dedicated to extracting and analyzing high-dimensional quantitative features from medical images [51]. These features, often not visible to the human eye, provide a deeper and more objective understanding of tissue characteristics [52], [53]. By using AI models, radiomics is able to identify complex patterns within medical images [53]. This allows for more insights into tumor biology and improves the decision-making process regarding cancer detection, prognosis, and treatment [2]. Unlike biopsy, which samples only a limited tissue region and may lead to missing key information, radiomics can capture heterogeneity throughout the full tumor volume [51].

Another advantage of radiomics is its ability to integrate information from different imaging modalities, such as PET and CT [52], [53]. Radiomic data are mineable, meaning they can be used in large datasets to uncover hidden patterns and biomarkers related to disease progression, treatment response, or patient outcomes [52]. As the field advances, radiomics continues to enhance the ability to interpret medical images, not by replacing traditional diagnostics but by adding valuable layers of quantitative information that support more personalized and informed clinical decisions.

6.1 Radiomic applications

As radiomic data can be extracted from different widely available imaging modalities, there are also different applications. Although radiomics has demonstrated potential in a variety of non-oncological domains, including the cardiovascular, neurological, and respiratory sciences, it is most commonly employed within the domain of oncology [54]. Within this domain, radiomics has already been applied to different types of cancer, but this thesis will focus on the use of radiomics in lung cancer, specifically early-stage NSCLC.

6.1.1 Lung cancer

Radiomics has become a vital tool in the analysis of lung cancer. This technique takes advantage of the large amounts of data provided by imaging modalities such as CT, PET, and PET/CT.

One of the major clinical applications of radiomics in lung cancer is the differentiation between benign and malignant tissue. Different studies have demonstrated that CT- and PET-based texture features can significantly improve the accuracy of distinguishing malignant from benign lung lesions, in contrast with traditional diagnostic methods [55], [56]. Radiomics can thus facilitate early and more accurate lung cancer diagnosis [57].

Furthermore, radiomics is valuable in assessing tumor heterogeneity, which has significant prognostic implications. For example, texture features derived from CT scans have been shown to reflect the underlying biology of tumors [58]. Additionally, radiomic data can be used to identify tumor subtypes. This can be a non-invasive alternative to invasive biopsy procedures. This is especially the case for lung cancer, where tissue biopsy is not always feasible or may not provide enough information [4].

Another application of radiomics in lung cancer is the measurement of total lesion glycolysis (TLG) in PET scans. TLG combines both tumor volume and the intensity of glucose uptake, thus providing a comprehensive measure of metabolic tumor activity. Several studies have demonstrated that TLG is a significant predictor of survival outcomes [59]. This makes TLG a valuable imaging biomarker that could assist in determining the aggressiveness of the tumor and the most appropriate treatment option [51].

6.1.2 Prognosis

Radiomics has demonstrated significant potential in prognostic outcomes in lung cancer, focusing on predicting overall survival (OS) and progression-free survival (PFS). The ability to predict survival is critical for personalized treatment, as it helps in selecting the most appropriate treatment option. Radiomic features derived from CT and PET scans, such as tumor heterogeneity, shape, and texture, have been correlated with clinical outcomes such as recurrence and metastasis [58], [57].

Metrics, such as TLG, play a crucial role in prognosis. Higher TLG values often indicate increased metabolic tumor activity and larger tumor volumes, both of which are associated with worse prognostic outcomes, especially for advanced-stage lung cancer [59]. However, in early-stage lung cancer, the application of radiomic prognoses is mainly used to predict the response to treatments. Radiomic features have been demonstrated to predict the chance of recurrence, influencing decisions regarding the need for additional treatments such as chemotherapy [58].

Thus, radiomics is effecting a paradigm shift in the domain of lung cancer by furnishing more accurate and personalized prognostic outcomes. The analysis of radiomic features facilitates not only prognostic assessments but also optimization for individual patients.

6.2 Workflow of radiomics

The general radiomic workflow consists of five main parts, as shown in Figure 13. The workflow starts with image acquisition, followed by tumor segmentation. This step is followed by feature extraction. Out of the extracted features, the relevant features will be selected in the fourth step. The final step includes performance testing, where models are developed, tested, and validated.

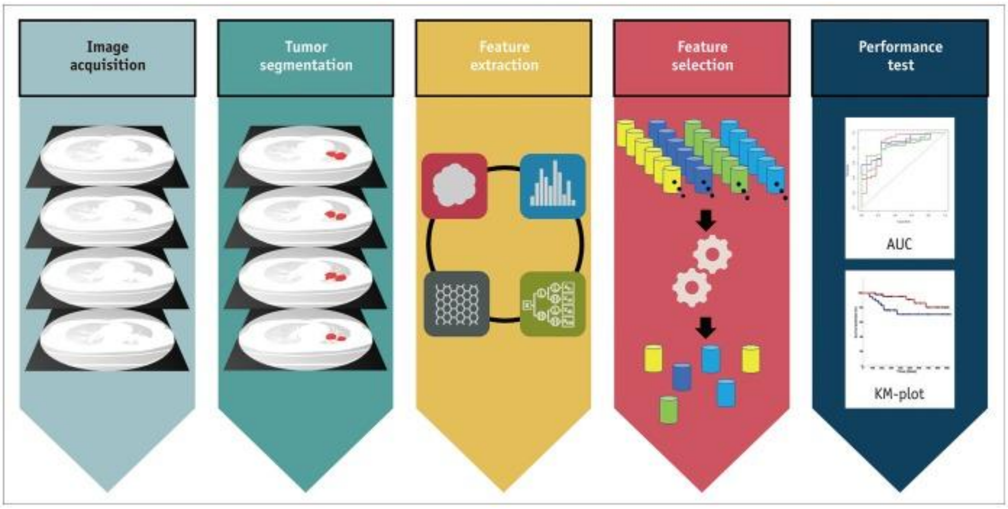


Figure 13: Workflow of radiomics [4, p. 160]. AUC: area under the curve, KM: Kaplan-Meier.

6.2.1 Image acquisition

The first step is image acquisition. High-quality images can be obtained through commonly used scans available in the hospital, such as CT and PET/CT scans. Extracted features can be influenced by the quality of the images [57]. Therefore, it is important to standardize the acquisition of images in order to use the same model for data derived from different machines [58]. More information about PET and CT images is provided in Chapter 5.

6.2.2 Tumor segmentation

Once a tumor is located on PET/CT, segmentation can be performed. This means the outlining of the metabolic tumor volume. Tumor segmentation can be performed in many different ways, such as manual segmentation, threshold-based methods, region-based methods, stochastic-based methods, boundary-based methods, and joint segmentation methods [60]. Each method uses specific criteria, such as intensity or spatial location, to determine if a voxel should be marked. A voxel can be defined as a 3D pixel, as PET/CT images are able to render 3D images. Segmentation produces a VOI created by the marked voxels. Because many tumors exhibit indistinct borders, which complicates their delineation, the selection of a segmentation method may significantly impact the final outcome of a model [61].

Most segmentation methods are semi-automated, meaning that the segmentation can be performed automatically with a segmentation algorithm, but using information provided by an observer. After the automatic segmentation, the observer can adjust selections, for example, by deleting regions with high physiological ^{18}F-FDG uptake, such as the brain or the bladder [62].

This study will focus on three different segmentation methods: two different threshold-based segmentation methods and one AI-based method. All three methods are discussed in detail below.

Threshold-based methods are the most commonly used segmentation methods. Because of its biologically significant evaluation of cellular metabolism, the **standardized uptake value (SUV)** is the most commonly used threshold-based segmentation method in PET imaging [60]. This value standardizes the intensity of a PET scan. It is defined as a ratio of radiotracer uptake in a region of interest to the total injected radioactive ^{18}F-FDG, normalized to the patient's bodyweight [63], [64]. If the SUV of a voxel is higher than the threshold, for example, 4 (SUV4), the voxel will be marked. When using SUV, a value of 2.5 or more typically indicates that the tissue might be malignant [65].

Another threshold-based segmentation method is based on the **majority vote (MV),** where the voxels in a VOI of the same object, segmented with different methods, will be compared. If a voxel is marked by the majority of the segmentation methods, the voxel is kept. Otherwise, the voxel is considered a segmentation error [66]. For example, MV2 marks a voxel if two or more segmentation methods, such as SUV2.5 or 41% of SUV_{max}, have marked the voxel [67].

Other than semi-automated methods, **AI-based segmentation methods, such as LIONZ**, exist. LIONZ is a deep learning-based automatic lesion segmentation tool for segmenting lesions in whole-body ^{18}F-FDG PET/CT scans of lung cancer patients. The tool was trained on the autoPET challenge dataset. LIONZ learned from over 1,000 ^{18}F-FDG -PET scans, including cases of melanoma, lymphoma, lung cancer, and cancer-free controls [68].

This study will conduct a comparative analysis of the SUV4, MV2, and LIONZ segmentation methods.

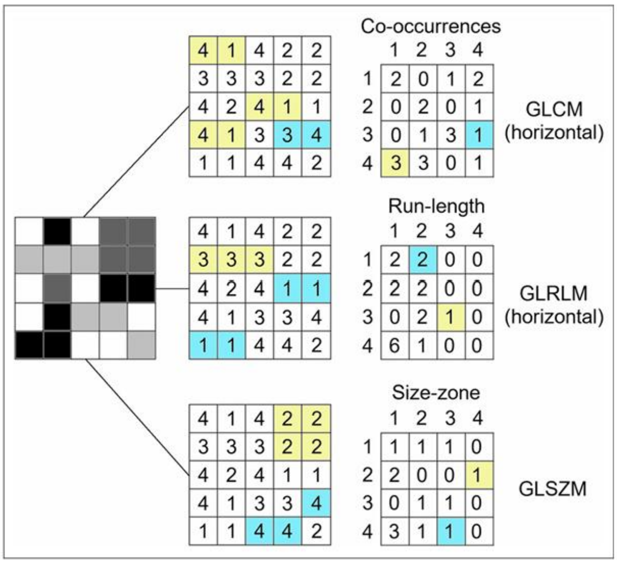
6.2.3 Feature extraction

Radiomic features are quantitative parameters that can be extracted from the VOI of medical images with the help of different programs such as LifeX or RADIOMICS. These features correlate with the outcome of clinical analysis [57]. To understand this thesis and other research, it is not mandatory to have knowledge about the different types of features, but it can aid in understanding some results. Therefore, a brief overview of the various types of radiomic features is included. [69] provides more detailed information about the different features.

Shape-based features characterize the geometry of the VOI and are conceptually the easiest radiomic features [52]. These features, such as volume, surface area, and surface-to-volume ratio, differentiate regular from irregular lesions. Because of the higher image resolution, shape features can be more accurately assessed using CT imaging compared to PET imaging [57].

Histogram features, also known as first-order statistical features, are based on single-voxel analyses, meaning that spatial connections are not taken into account [59]. These features include values that are relatively invariant to geometric transformation and thus based on the gray-level histogram, such as grey-level mean, range, and skewness [57]. Conventional PET metrics such as SUVmax, which indicates the maximum SUV in a region; DmaxBulk, which indicates the distance between lesions; or TLG are classified under this category [70], [71].

Texture features, also known as second-order statistical features, are acquired by the joint signal variation between neighboring voxels [57], [59]. These features are commonly categorized into five types. The Gray-Level Co-occurrence Matrix characterizes the spatial relationship between voxel pairs based on predefined intensity levels and spatial configurations. The Gray-Level Run-Length Matrix captures the distribution of consecutive voxels with identical intensities along specific directions. The Gray-Level Size Zone Matrix reflects the size and frequency of homogeneous zones formed by connected voxels with the same gray level. The three previously explained types are shown in Figure 14: Three types of texture features in radiomics. GLCM is based on the spatial relation between voxel pairs, GLRLM is based on the distribution of consecutive voxels, and GLSZM is based on homogenous zones with the same gray-level voxels Figure 14. The Neighborhood Gray-Tone Difference Matrix assesses the contrast between a voxel's intensity and the average intensity of its surrounding neighborhood. Lastly, the Neighborhood Gray-Level Dependence Matrix quantifies the extent to which voxels with similar intensities are spatially dependent within a defined neighborhood [52].



4	1	4	2	2
3	3	3	2	2
4	2	4	1	1
4	1	3	3	4
1	1	4	4	2

Co-occurrences

	1	2	3	4
1	2	0	1	2
2	0	2	0	1
3	0	1	3	1
4	3	3	0	1

Run-length

	1	2	3	4
1	2	2	0	0
2	2	2	0	0
3	0	2	1	0
4	6	1	0	0

Size-zone

	1	2	3	4
1	1	1	1	0
2	2	0	0	1
3	0	1	1	0
4	3	1	1	0

Figure 14: Three types of texture features in radiomics. GLCM is based on the spatial relation between voxel pairs, GLRLM is based on the distribution of consecutive voxels, and GLSZM is based on homogenous zones with the same gray-level voxels [52, p. 490]. GLCM: Gray-Level Co-occurrence Matrix, GLRLM: Gray-Level Run-Length Matrix, GLSZM: Gray-Level Size Zone Matrix.

Transform-based features, also known as higher-order statistical features, are extracted using mathematical methods to enhance pattern identification and suppress noise in medical images [59]. These mathematical methods include wavelet transformation, Fourier transformation, Laplacian transformation, and Minkowski functionality. These transformed domains capture subtle variations in texture and intensity that may not be visible in the original image. The resulting features provide a more comprehensive quantification of tissue heterogeneity for clinical analysis [52], [59].

6.2.4 Feature selection

As approximately 500 different radiomic features are extracted, the fourth step of the radiomic workflow involves the selection of the most stable and prognostic features for further analysis [4], [9]. This can be achieved through various methods, such as filter or wrapper methods. Filter methods evaluate features based on their statistical relevance, independently of any model. Wrapper methods asses different subsets of features by training a model to determine which combinations have the best model performance [72]. This section explains different feature selection methods used in this master's thesis.

The **Spearman rank correlation test** is a filter-based method used to measure the correlation between two features. The application of this test is not limited to specific categories of data; it can be used with both continuous and discrete data, as well as with data that follows a normal distribution and data that does not follow a normal distribution [73], [74]. The underlying reason for this discrepancy is the utilization of rank variables rather than the original values of the data [75]. Equation (2) is used to calculate the Spearman rank correlation coefficient ρ [73], [76]:

$$\rho = 1 - \frac{6 \sum d_i^2}{n\,(n^2-1)} \tag{2}$$

Equation (2) represents the difference between the ranks of the paired variables as d_i, while n represents the total number of observations. Different values of ρ indicate different levels of correlation. The correlation value typically ranges from -1 to +1, with 0 indicating the absence of a correlation. The sign of the value indicates either a negative or positive correlation. A correlation value, positive or negative, between 0.10 and 0.29, between 0.30 and 0.49, and above 0.50 indicates little, medium, and strong correlations, respectively [73]. Features demonstrating a correlation above a defined threshold with the target variable are considered for elimination [75].

Pairwise elimination (pw), or pairwise feature selection, is a filter-based method that reduces the number of selected features by evaluating the correlation between feature pairs. The correlation between each feature is presented in a correlation matrix (Figure 15). The values in a correlation matrix can be the output of, for example, a Spearman correlation test, as previously explained. Systematically, one feature from a highly correlated pair, with a correlation value above a predefined threshold, is eliminated. The decision on which feature to eliminate can be based on various criteria, such as the feature's variance, its mean absolute correlation with other features, or its relevance to the outcome of interest [77]. A major advantage of pw is that the results are straightforward, and the visual representation of the results in the correlation matrix is easy to read. However, pw has difficulties with capturing non-linear relationships, which can lead to over-eliminating features, which will result in decreased model performance [78].

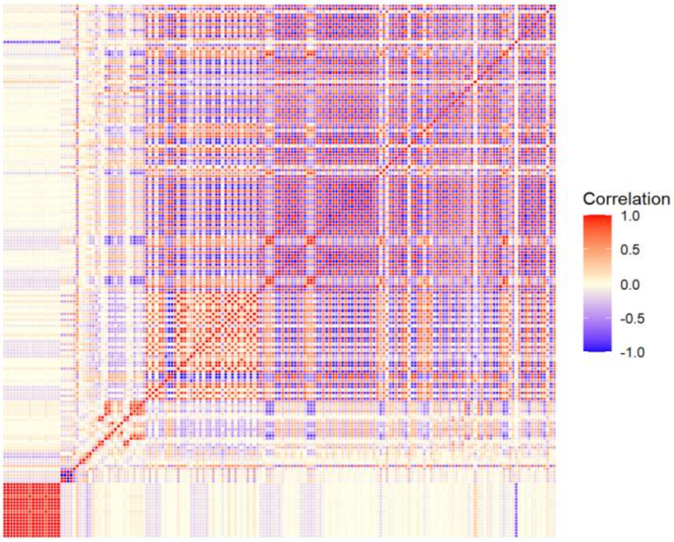


Figure 15: Example of a correlation matrix of radiomic features, with correlation values derived from a Spearman correlation test.

Univariate selection (uni) is a commonly used filter-based method for identifying features that are individually highly correlated with a target feature. This technique evaluates each feature independently using statistical tests such as the Chi-squared test or the Spearman correlation test [72], [79]. Univariate methods have a high computational efficiency and are known to be consistent and to have stable performance across multiple applications, including cancer classification [79]. However, univariate methods also have limitations. For example, they cannot capture interactions between variables or identify redundant features, potentially reducing model performance [80], [81].

Recursive feature elimination (rfe) is a wrapper-based method commonly used to reduce the number of radiomic features by successively removing less important ones, thereby improving the predictive accuracy of prognostic models. This process involves repeatedly building a model, identifying the least important features, and removing them one by one. At each step, the model is re-trained on the reduced feature set, which is repeated until a certain number of features remain [82]. The importance of each feature is based on how much it contributes to the model's predictions. Because rfe re-evaluates feature importance at every step, it also considers how features interact with one another rather than evaluating them individually [79]. Rfe leads to robust and generalizable radiomic models [83]. However, this method can be computationally intensive when using many features [78].

6.2.5 Performance testing

The last step in the radiomic workflow is performance testing. A variety of prognostic models can be used for radiomic analysis. For this study, two main types of prognostic models are used, namely logistic regression and rf with different feature selection methods.

Logistic regression is a commonly used statistical technique for predicting binary outcomes, such as the occurrence of a clinical event. It estimates the probability of an event, such as disease recurrence, based on one or more radiomic features [84], [85], [86].

In contrast to linear regression, which deals with predicting continuous variables, logistic regression models predict the probability of a binary event by applying the logit function, thereby transforming the binary outcome into a continuous scale of log odds. The logit function is given by (3), where p represents the probability of an event [85]:

$$logit\ (p_x) = \log\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \tag{3}$$

A coefficient β is assigned to each radiomic feature *X*, which acts as the input variable. The coefficient reflects how much a one-unit change in the feature affects the odds (log(px/1-px)) of the outcome [85], [86].

Logistic regression is easy to interpret and can adjust its results for disturbing variables, which is especially useful with large datasets. By including multiple features in a single model, it can isolate the unique contribution of each feature while checking for others. This makes it useful for building multivariable predictive models because it assesses not only individual feature relevance but also the correlation between features [84].

However, logistic regression also has its limitations. To avoid overfitting, the number of features included must be limited relative to the number of outcome events, often following the "10 events per feature" rule to avoid overfitting, which can cause problems for small datasets [85], [86].

Overall, logistic regression remains a fundamental and interpretable method for binary outcome modeling, offering both statistical insight and practical utility in predictive modeling.

Random forest (rf) is a machine learning technique effective for classification and regression tasks, such as predicting patient outcomes in radiomics. The core idea of rf is to build many decision trees for given data. Each tree is trained on a different subset of the data, and at each decision point in the tree, only a random subset of the available features is considered. This results in two randomization strategies: the data used to build the trees and the features used. The use of two strategies helps to reduce the correlation between trees, improving the model's ability to generalize and avoid overfitting [87], [88], [89].

Because rf combines the prediction of many trees, it is considered an ensemble learning technique [87]. Each tree makes its own prediction, and at the end, the overall decision is determined by either majority voting (for classification tasks) or by averaging (for regression tasks) [88]. By combining many trees, rf reduces the high variance that typically affects a single decision tree, making it less likely to memorize the training data and thus improving the model's accuracy [89].

One major advantage of rf is its ability to rank the importance of different features in making predictions. This is particularly useful in complex fields like radiomics, where many features are used [87]. However, a disadvantage of rf is that it is not immune to overfitting. It can sometimes show high performance on training data, which can be fitting noise or irrelevant patterns. However, when enough trees are used, above 500, the model tends to be more stable.

This makes it a reliable method, even in cases where the data may be noisy or incomplete. Studies have shown that increasing the number of trees does not degrade performance and can stabilize the model further [88], [90].

Rf is also relatively easy to use. Key parameters, like the number of trees and features considered at each decision point, have reasonable defaults that work well in most cases. This makes rf particularly useful in clinical settings, where datasets can be small [90]. In clinical and medical applications, rf is used for tasks like predicting patient survival outcomes. Its ability to handle high-dimensional data while maintaining high accuracy makes it a practical tool in healthcare research and decision-making [87], [90].

7 Model analysis

There are several methods to analyze the performance of a model. This thesis uses the receiver operating characteristic (ROC) curve as a visual representation and the area under the ROC curve (AUC) as a numerical representation of the results.

7.1 Confusion matrix

Confusion matrices, such as those shown in Figure 16, are a visual representation of the predicted outcome from the tested model compared to the true outcome of the tested data.

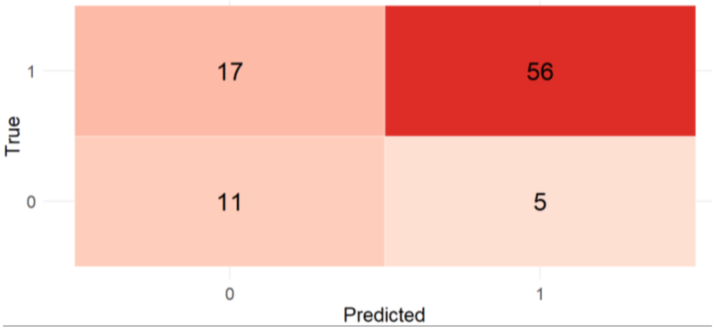


Figure 16: Confusion matrix of model S1 (TLG-only) using MV2 segmentation and an overall survival of 36 months as prognostic outcome, evaluated at a threshold of 0.5.

The confusion matrix shows how many true negatives (TN) are predicted in the bottom left cell. The true positives (TP) are shown in the upper right cell. Both cells show the number of correct predictions made by the model. However, the upper left cell, the false negatives (FN), and the lower right cell, the false positives (FP), show the false predictions made by the model. Confusion matrices always show the prediction for a certain threshold.

7.2 Receiver operating characteristics

A ROC curve can be created when all confusion matrices for thresholds between 0 and 1 are considered together. ROC analysis is used for evaluating the diagnostic performance of binary classification systems, particularly in medicine and machine learning. It is a visual way to assess how well a predictive model distinguishes between two options [91], [92], [93].

On the ROC curve, the horizontal axis represents the false positive rate (FPR), or 1-specificity, while the vertical axis represents the true positive rate (TPR), or sensitivity, illustrating how model performance varies across different thresholds [91], [93], [94]. The TPR, defined in (4), quantifies the fraction of actual positive cases that are correctly identified by the model.

$$TPR = \frac{true\ positive}{true\ positive + false\ negative} \tag{4}$$

The FPR, defined in (5), quantifies the proportion of actual negative cases that are incorrectly identified as positive by the test.

$$FPR = 1 - specificity = 1 - \frac{false\ positive}{false\ positive + true\ negative} \tag{5}$$

By calculating TPR and FPR for every possible cut-off value, a series of (FPR, TPR) coordinates is generated. These points are plotted to form the ROC curve [93]. The shape of the ROC curve (Figure 17) shows how well a test separates the positive and negative cases. If the curve goes toward the upper-left corner, this indicates a higher overall accuracy. A curve close to the diagonal line suggests random guessing [91], [94].

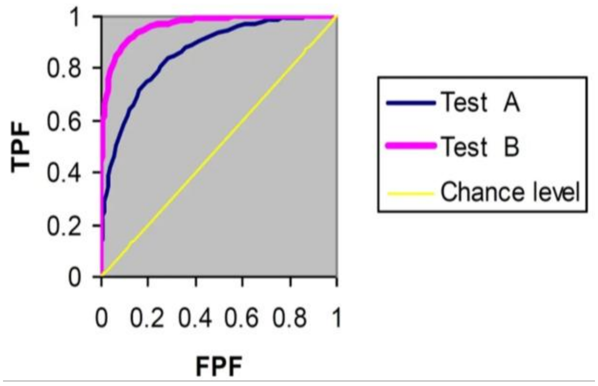


Figure 17: Graphical representation of a ROC curve. Test B (pink) separates positive from negative cases better than test A (purple). The yellow line indicates the diagonal resulting from purely random guessing [92]. TPF: true positive fraction, FPF: false positive fraction.

ROC analysis is independent of a specific threshold, unlike metrics such as accuracy or precision, which depend on a single threshold. The ROC curve shows how performance varies across all possible thresholds, which makes it a powerful tool for evaluating and comparing machine learning models [92].

7.3 Area under the curve

The ROC curve is a helpful visual representation of a model's prognostic performance. However, it is often useful to summarize this information with a single number, called the AUC. The AUC reflects the overall ability of the model to separate positive and negative cases [93], [94].

The empirical AUC can be calculated using the trapezoidal rule. It is mathematically equivalent to the Mann–Whitney U-statistic, which estimates the chance that a randomly chosen positive case will have a higher predicted occurrence than a randomly chosen negative case [95], [96]. Equation (6) shows the empirical AUC $\hat{\theta}$:

$$\hat{\theta} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\psi(X_i, Y_i) \tag{6}$$

where X_i and Y_j are the predicted scores for positive and negative events, and m and n are the number of positive and negative samples. $\psi(X_i, Y_j)$ is the Heaviside function, defined in (7) [96]:

$$\psi(X_i, Y_j) = \begin{cases} 1 & if\ X_i > Y_j \\ 0.5 & if\ X_i = Y_j \\ 0 & if\ X_i < Y_j \end{cases} \quad (7)$$

The Heaviside function is a step function used to compare the predicted scores. If the predicted score for a positive case is greater than the predicted score for a negative case ($X_i > Y_j$), then the function returns 1. If the scores are equal ($X_i = Y_j$), the function returns 0.5, and if the positive case has a lower predicted score than the negative case ($X_i < Y_j$), the function returns 0.

The Heaviside function allows the empirical AUC to account for the relative ranking of the predicted scores. AUC values range from 0 to 1. An AUC of 0.5 indicates random guessing and thus no distinguishing ability, whereas an AUC of 1.0 represents perfect distinguishment. Therefore, a higher AUC indicates a better overall performance [91], [94]. In clinical practice, AUC values are often categorized to make interpretation easier. As no realistic classifier is under 0.5 because of random guessing, the different categories are 0.50-0.70, 0.70-0.90, and above 0.90. These categories represent low, moderate, and high accuracy, respectively [93]. However, high AUC values do not inherently signify clinical efficacy. For example, a test with an AUC of 0.81 might still be unreliable if its 95% confidence interval goes from 0.65 to 0.95, since the lower end is close to the low accuracy category [93]. It is, therefore, important to study both the AUC and its confidence interval when judging the performance of a model.

In conclusion, the ROC curve and AUC values are powerful tools for evaluating prognostic models. The ROC curve illustrates the balance between TPR and FPR at various thresholds, while the AUC summarizes the overall prognostic performance.

7.4 Friedman test

The Friedman test is a nonparametric statistical method used to detect differences between three or more related groups [97]. It is particularly useful in repeated-measures designs, where the same subjects are evaluated under multiple conditions. This is useful when comparing the AUCs of different segmentation methods across multiple prognostic models, where each method is applied to the same dataset.

The test ranks the AUC values within each subject, where a subject refers to a single prognostic model. For each model, the segmentation methods are compared based on their AUC values. The segmentation method with the highest AUC value receives the highest rank, the second-highest receives the next highest rank, and so on. These ranks are then summed for each segmentation method across all models, providing a basis for statistical comparison [98]. The Friedman test statistic is calculated as χ^2 in (8) [97]:

$$\chi^2 = \frac{12}{nk(k+1)} \sum_{j=1}^{k} \{R_j - n(k+1)/2\}^2 \quad (8)$$

where n is the number of subjects (prognostic models), k is the number of groups being compared (segmentation methods), and R_j is the sum of the ranks for group j.

Under the null hypothesis that all segmentation methods perform equally, the Friedman test follows a chi-squared distribution with *(k−1)* degrees of freedom [97]. This allows the test to assess whether the observed rank differences are likely due to chance. If the resulting p-value is smaller than 0.05, the null hypothesis is rejected, indicating that at least one segmentation method performs differently from the others. A p-value greater than 0.05 suggests that any observed differences could be attributed to random variation [99].

Because the Friedman test does not assume normality and handles repeated measures effectively, it is particularly well-suited to medical imaging studies involving non-Gaussian distributions and correlated observations [98].

7.5 DeLong test

The DeLong test is a nonparametric statistical method used to determine whether AUCs are statistically significantly different from each other. This is useful when comparing models evaluated on the same dataset, such as clinical prediction models or machine learning classifiers [95], [96].

To compare the empirical AUCs of two models, DeLong's method treats them as related statistics and calculates their variance and covariance, allowing for statistical testing based on the normal distribution [96], [100]. Equation (9) calculates the test statistic as a z-score:

$$z = \frac{\hat{\theta}_A - \hat{\theta}_b}{\sqrt{Var(\hat{\theta}_A - \hat{\theta}_B)}} \tag{9}$$

Under the null hypothesis that the models have equal AUCs, this z-score follows a standard normal distribution. This means that it can be used to check if the difference in AUCs is likely due to chance. If the absolute value of z is greater than 1.96, the difference in AUCs is considered significant [95]. This indicates that it is unlikely to have happened by random variation alone.

Although the z-score indicates how far the observed difference in AUC deviates from the null hypothesis (measured in standard deviations), the p-value translates this into a probability, making the result more straightforward to interpret. The p-value indicates the probability of such a difference occurring by chance. It is calculated from the z-score using the standard normal distribution (z-distribution), which can be done using a statistical z-table or an online calculator. For a two-tailed test, a z-score of 1.96 corresponds to a p-value of 0.05. This means that there is a 5% chance of observing such a result due to random variation. Therefore, $p < 0.05$ is commonly used as the threshold for statistical significance [95], [96].

The method can also be extended to compare more than two models using a covariance or comparison matrix of AUCs and applying generalized U-statistics theory [96].

8 Materials and methods

The data used in this study are derived from the ProLUNG study funded by ‘Kom op tegen Kanker’ and performed at Ziekenhuis Oost-Limburg (ZOL) in Genk. All study participants were diagnosed with resectable stage I-IIIA NSCLC or early-stage NSCLC. All participants signed informed consent before inclusion.

A total of 123 patients were included in the ProLUNG study. However, 34 patients were excluded from the current research for various reasons, resulting in a total of 89 patients. All these patients were designated to the training cohort of the prognostic models. The different exclusion criteria are listed in Table 2, along with the number of excluded patients for each criterion.

Table 2: Exclusion criteria of the training cohort with the number of excluded patients per criterion. SCLC: small cell lung cancer.

Exclusion criteria	
No images available	1
Chemo before lobectomy	1
Carcinosarcoma	1
SCLC	3
No lobectomy	3
Carcinoid	4
Tumor is too small to segment	5
Only inflammation	16
Excluded	**34**

For the external validation of each prognostic model, a second, independent, patient cohort was assembled from the NCT02024113 trial performed at the Limburg PET-Center in Hasselt [101]. Only patients with resectable stage I-IIIA NSCLC were selected from this study, resulting in a second patient cohort of 32 patients, designated the validation cohort.

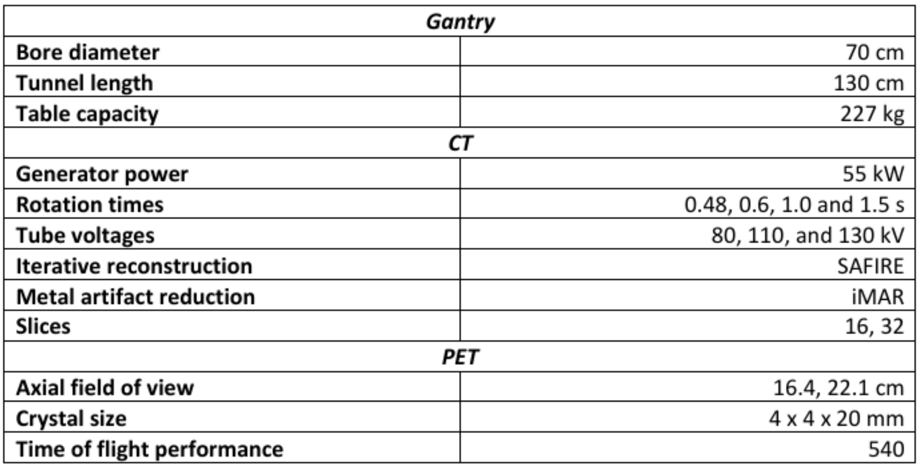
8.1 ^{18}F-FDG PET/CT scanning procedure

The ^{18}F-FDG PET/CT scanning procedure of both the training and validation cohort are explained in this part. All data, including the PET/CT scans from the ProLUNG study and the NCT02024113 trial, had previously undergone pseudonymization prior to its utilization in the current research.

8.1.1 Training cohort

A Biograph Horizon PET/CT scanner from Siemens with lutetium oxyorthosilicate (Lu2(SiO4)O) scintillation crystals was used to obtain PET/CT scans from the included patients of the training cohort. Technical specifications of the imaging device are listed in Table 3. PET scan attenuation correction was performed using the CT scans. All images were taken following the European Association of Nuclear Medicine's imaging procedure guidelines and saved in the Picture Archiving and Communications Systems.

Table 3: Technical specifications of the used Biograph Horizon PET/CT scanner from Siemens [102, p. 66]. CT: computed tomography, PET: positron emission tomography.



Gantry	
Bore diameter	70 cm
Tunnel length	130 cm
Table capacity	227 kg
CT	
Generator power	55 kW
Rotation times	0.48, 0.6, 1.0 and 1.5 s
Tube voltages	80, 110, and 130 kV
Iterative reconstruction	SAFIRE
Metal artifact reduction	iMAR
Slices	16, 32
PET	
Axial field of view	16.4, 22.1 cm
Crystal size	4 x 4 x 20 mm
Time of flight performance	540

To perform the PET scan of the training cohort, the radiopharmaceutical ^{18}F-FDG was injected into the patient. In order to minimize the radiation exposure to medical personnel, the Iris automated multidose injection system from Comecer administered 3 MBq/kg of the ^{18}F-FDG to the patient. Images were captured one hour after the injection was administered. Firstly, a CT scan (25 mA, 130 kV) was conducted, encompassing the region from the mid-thighs to the base of the skull. Secondly, a PET scan was conducted with a duration of 15-20 minutes, covering the same area. Depending on the patient's weight, a different emission time per bed position was used. Patients with a mass of less than 50 kg, between 50 and 80 kg, and more than 80 kg were scanned for one minute, one and a half minutes, and two minutes, respectively.

8.1.2 Validation cohort

A GEMINI TF Big Bore PET/CT scanner from Philips was used to obtain PET/CT scans from the patients included in the validation cohort. To perform the PET scan of the validation cohort, 3.75 MBq/kg of the radiopharmaceutical ^{18}F-FDG was injected into the patient. Images were captured one hour after the injection was administered. Firstly, a CT scan (80-175 mA, 120 kV) was conducted, encompassing the region from the mid-thighs to the base of the skull. Secondly, a PET scan was conducted with a duration of 15-20 minutes, covering the same area. Depending on the patient's body mass index (BMI), a different emission time, between 1 and 2 minutes per bed position, was used.

8.2 Data acquisition

After image acquisition, segmentation is the next step in the radiomic workflow, as explained in section 0 *Workflow of radiomics*. The acquired PET and CT images were loaded into the ACCURATE tool, developed by the research team of Prof. Dr. Ronald Boellaard (Amsterdam, UMC). Three different segmentation methods (SUV4, MV2, and AI-based tool LIONZ) were performed by this semi-automatic tool. An example of such segmentation is shown in Figure 18. Because of the large dataset, segmentations performed by LIONZ were run offline in batch mode to minimize user interference for every patient. Segmentations derived from all three methods could be manually corrected in the ACCURATE tool by adding, removing, or altering marked parts. If the segmentation of a patient was unclear, Prof. Dr. Liesbet Mesotten, head of the Nuclear Medicine Department at ZOL, was consulted. The final marked parts were saved as individual VOIs.

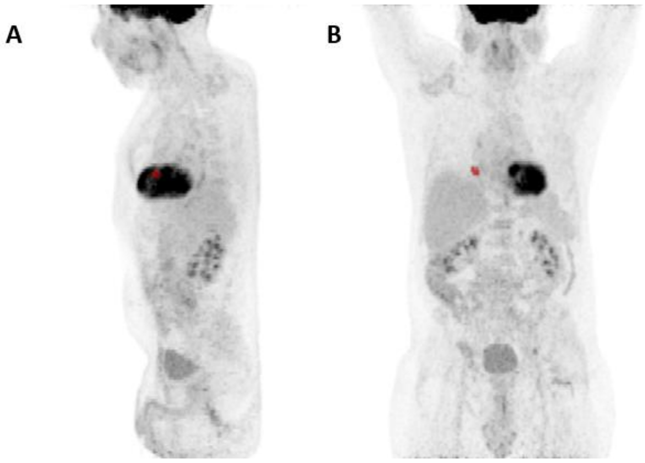


Figure 18: Example of a segmentation on a PET/CT image in the ACCURATE tool with (A) a sagittal view and (B) a frontal view. The tumor is segmented/indicated in red.

After segmenting and defining the VOIs, feature extraction was performed. This was accomplished by loading the VOI files into the RADIOMICS tool, which is also developed by the research team of Prof. Dr. Ronald Boellaard (Amsterdam, UMC). This tool extracted radiomic features from the VOI files, resulting in CSV files with six PET uptake metrics and 498 radiomic features. All parameters were saved in three separate Excel files, sorted per segmentation method. A more detailed explanation of how to perform the segmentation and feature extraction is added in Appendix A.

In order to perform prognostic analyses, additional patient information was combined with the radiomic data. The additional data could be split into four sections.

The first section included general patient information such as sex, age, and BMI of the patient. The second section provided information about the smoking history of the patient, where a value of '1' denotes a positive response, indicating that the subject has smoked. Conversely, the value of '0' denotes a negative response, indicating that the subject has never smoked. Furthermore, this section also includes the number of pack-years and the number of years a patient has stopped smoking. One pack-year is calculated by multiplying the number of packs (20 cigarettes) smoked daily by the number of years the individual has smoked.

In the third section, information about survival was gathered. The data in this section is used as an outcome for the prognostic models. Three types of events were taken into account: progression-free survival (PFS), overall survival (OS), and time to progression (TTP). PFS refers to the number of months a patient remained free from disease progression or death. For each patient, this duration was measured from the date of lobectomy to the date of progression or death. OS measures the time, in months, from the date of lobectomy to the date of death. TTP refers to the number of months a patient remained free from disease progression, starting from the date of lobectomy. For all three outcomes, patient data were labeled with ‘1’ if the event, disease progression, or death occurred and with ‘0’ if the event did not occur. In cases where an event did not occur, the end date was set as 1/04/2025, as this is the day the information was extracted.

The fourth section included information about tumor features such as the TNM stage, the overall stage, the type of NSCLC, the lobe position of the tumor, and the tumor’s diameter.

This process of data acquisition was performed for both the training and validation cohorts.

8.3 Feature reduction

Given that not all 498 radiomic features are relevant, feature reduction was executed in RStudio. First, features with constant values and/or minimal variance were eliminated, as these do not have any distinguishing power between patients. Second, the Spearman correlation test was used to eliminate features more than 75% correlated with volume or SUVmax, as these two features are well understood in the clinical world and are, therefore, worth keeping.

Three additional reduction methods were eventually performed, including pw, uni, and rfe. Each method was executed independently following the implementation of the Spearman correlation test. The top five remaining features of each separate method were evaluated in a logistic regression model. This procedure was repeated for each segmentation method.

8.4 Prognostic models

The R scripts that contain the feature reduction also contain the prognostic models. For every segmentation method, the R script is the same, except for the data used. Six different linear regression models and one rf model were developed by the research team of Prof. Dr. Ronald Boellaard (Amsterdam, UMC). These models were tested for four different prognostic outcomes. Table 4 provides an overview of all the different models and prognostic outcomes. In total, 84 combinations were tested because of the three segmentation methods, seven prognostic models, and four prognostic outcomes used.

Table 4: Overview of the different prognostic models and different prognostic outcomes used. TLG: total lesion glycolysis.

Model	s1 – TLG
	s2 – TLG and SUVmax
	s3 – TLG, SUVmax, and DmaxBulk
	pw – pairwise elimination
	uni – univariate selection
	rfe – recursive feature elimination
	rf – random forest
Prognostic outcome	OS24 – Overall survival after 24 months
	OS36 – Overall survival after 36 months
	TTP24 – Time to progression after 24 months
	TTP36 – Time to progression after 36 months

8.5 Model analysis

Seven models (s1, s2, s3, pw, uni, rfe, and rf) were trained using radiomics features obtained for all patients in the training cohort, using three different segmentation methods. The data was internally validated using a 5-fold 10-times cross-validation and externally validated using the independent validation cohort. As lesions from patients in the validation cohort were not segmented using SUV4, external validation was only possible for the MV2 and LIONZ segmentation methods.

The results from all 84 tested combinations, for cross-validation and external validation, were gathered in confusion matrices evaluated at a threshold of 0.5. The TPR and FPR values from the confusion matrices were used to make ROC curves. From the ROC curves, AUC values, their standard deviations, and a 95% confidence interval could be calculated for every combination.

The first objective was to assess whether the choice of segmentation method affects the performance of prognostic models. For this analysis, AUC values from the ROC curves were employed in a Friedman test. First, for each individual model, the segmentation methods were ranked based on their performance. Then, these rankings were compared across all the models to determine whether there are statistically significant differences in performance between the segmentation methods overall, with significance defined by a p-value of less than 0.05. Furthermore, a histogram was constructed to facilitate the visualization of the data. This histogram incorporates the AUC values from each segmentation method for every model.

The second objective was to determine the best-performing prognostic model. AUC values were compared using the DeLong test to assess whether statistically significant differences existed between the models. This analysis was carried out for each segmentation method separately. From the models where significant differences were observed, the AUC values and their standard deviations were further examined to determine the overall best-performing model, as well as the most best-performing model–segmentation method combination.

9 Results

In this chapter, the demographics of the patient cohorts are discussed, as well as the model analysis results. The model analysis is split into two parts, each discussing the results of one of the two main objectives of this thesis. All discussed results relate to the prognostic outcome OS36 as this is clinically the most relevant. More information about this specified cut-off is provided below. The results from all prognostic outcomes can be found in Appendices B-E.

9.1 Demographics

For training of the prognostic models, the training cohort of 89 patients derived from the ProLUNG study was used. In Table 5, an overview of the demographic features of this training cohort is shown. The table can be divided into four main parts: general patient information, smoking information, survival, and tumor information.

Table 5: Demographic information of the patients in the training cohort (N=89). Categorical variables are indicated with the amount (N), followed by the percentage in brackets. Continuous variables are indicated with the amount, followed by the standard deviation. BMI: body mass index, NSCLC: non-small cell lung cancer.

Total patients		89
Sex (N,(%))	Men	52 (58.4)
	Women	37 (41.6)
Age (years)	Median	70
	Average	69 ± 8
	Range	45-83
BMI (kg/m²)	Median	26.1
	Average	26.4 ± 5
	Range	16.6-50.4
Smoking status (N,(%))	Current smoker	45 (50.6)
	Ex-smoker	40 (44.9)
	Non-smoker	4 (4.5)
Packyears (years)	Median	37
	Average	38 ± 25
	Range	1-132
Years stopped smoking (years)	Median	20
	Average	21 ± 14
	Range	2-49
Survival outcome (N,(%))	Death	9 (10.1)
	Progression and death	13 (14.6)
	Progression	14 (15.7)
	Progression-free	53 (59.6)
Overall survival (months)	Median	53
	Average	51 ± 20
	Range	3-81

Time to progression (months)	Median	49
	Average	44 ± 23
	Range	3-81
Stage (N,(%))	IA	41 (46.1)
	IB	16 (18.0)
	IIA	2 (2.2)
	IIB	13 (14.6)
	IIIA	14 (15.7)
Type NSCLC (N,(%))	Adenocarcinoma	59 (66.3)
	Squamous cell carcinoma	26 (29.2)
	Large cell neuroendocrine carcinoma	4 (4.5)
Node inclusion (N,(%))	Yes	4 (4.5)
	No	85 (95.5)
Lobe position (N,(%))	Right upper	24 (27)
	Right middle	5 (6)
	Right lower	12 (13)
	Left upper	29 (33)
	Left lower	19 (21)
Tumor diameter (mm)	Median	25
	Average	28 ± 16
	Range	7-80

Figure 19 provides more details about the survival of the patients in the training cohort. It displays how survival events (1) and non-events (0) are distributed over time. An event is defined as the death of a patient, whereas a non-event indicates that the patient was still alive at the end of the observation period. The variation in follow-up time among non-events is due to differences in the patients' inclusion dates. There is a noticeable split around 36 months, supporting the use of OS36 as the most suitable prognostic outcome for further statistical analyses.

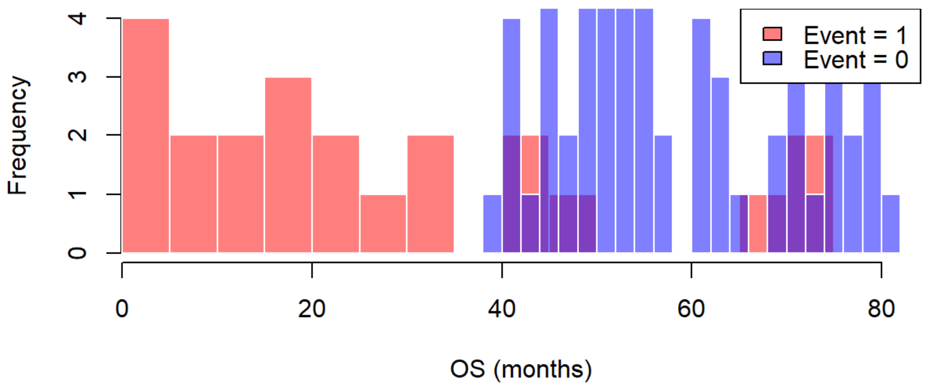


Figure 19: Distribution of overall survival (OS) events (1) and non-events (0) over time. An event is defined as the death of a patient, whereas a non-event indicates that the patient was still alive at the end of the observation period.

9.2 Influence of segmentation method on the performance of prognostic models

The first objective of this thesis is to determine if the choice of segmentation method has an impact on the performance of prognostic models.

9.2.1 Cross-validation results

ROC curves of all cross-validated prognostic models obtained using SUV4, MV2, and LIONZ segmentation are shown in Figure 20.

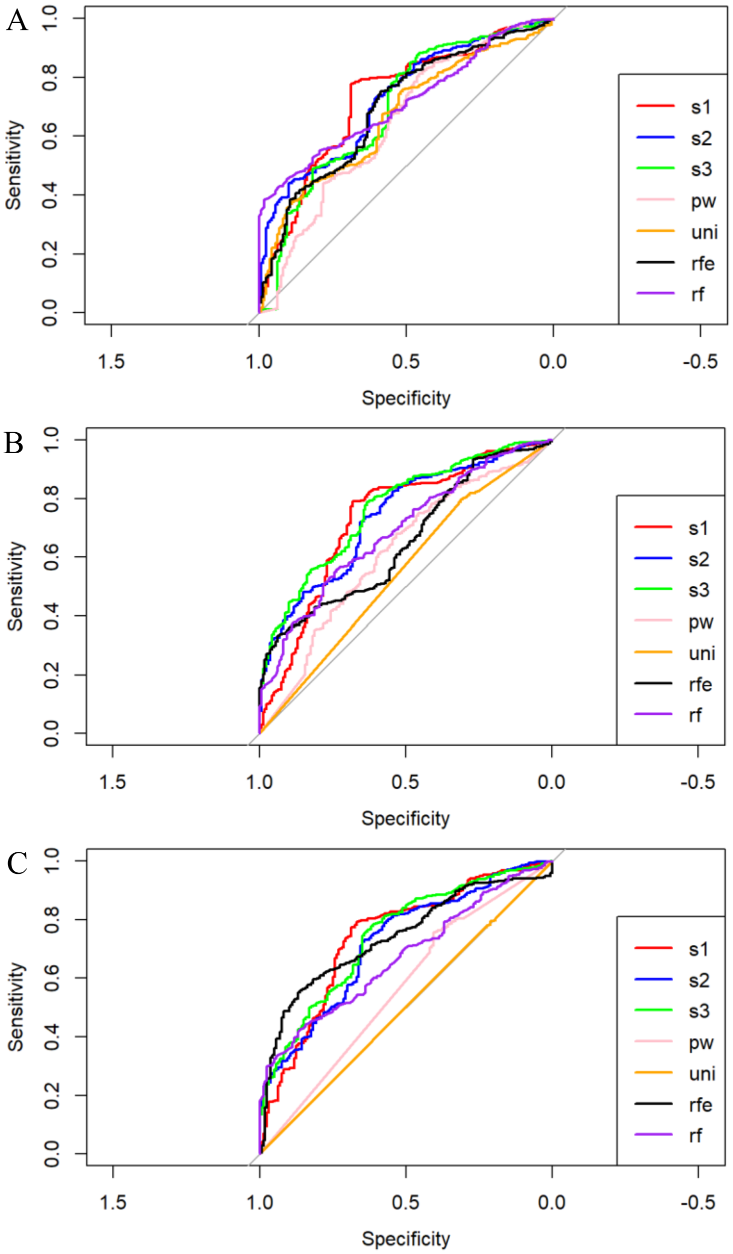


Figure 20: ROC curve of all cross-validated prognostic models with an overall survival of 36 months as prognostic outcome. (A) ROC curves of the prognostic models using SUV4 segmentation, (B) ROC curves of the prognostic models using MV2 segmentation, (C) ROC curves of the prognostic models using LIONZ segmentation.

Table 6 summarizes all average AUC values, along with their corresponding standard deviations and 95% confidence intervals, across the 5-fold, 10-times cross-validation for all models across each segmentation method.

Table 6: AUC values with standard deviation and 95% confidence interval of cross-validated prognostic models using SUV4, MV2, and LIONZ as segmentation methods. All models use an overall survival of 36 months as a prognostic outcome. AUC sd: area under the curve standard deviation, 95% CI: 95% confidence interval.

OS36	s1	s2	s3	pw	uni	rfe	rf
	SUV4 (mean AUC: 0.72 ± 0.14)						
AUC	0.75	0.74	0.72	0.68	0.68	0.72	0.72
AUC sd	0.15	0.10	0.15	0.17	0.17	0.15	0.12
95% CI	[0.713, 0.795]	[0.714, 0.771]	[0.677, 0.760]	[0.636, 0.732]	[0.634, 0.737]	[0.681, 0.764]	[0.683, 0.749]
	MV2 (mean AUC: 0.69 ± 0.15)						
AUC	0.76	0.75	0.77	0.66	0.54	0.67	0.70
AUC sd	0.13	0.14	0.16	0.18	0.12	0.16	0.13
95% CI	[0.722, 0.796]	[0.706, 0.785]	[0.729, 0.815]	[0.612, 0.710]	[0.511, 0.577]	[0.622, 0.710]	[0.662, 0.734]
	LIONZ (mean AUC: 0.68 ± 0.14)						
AUC	0.76	0.74	0.76	0.57	0.50	0.76	0.69
AUC sd	0.14	0.16	0.13	0.15	0.14	0.14	0.13
95% CI	[0.720, 0.795]	[0.690, 0.781]	[0.729, 0.798]	[0.529, 0.613]	[0.464, 0.542]	[0.722, 0.799]	[0.653, 0.723]

In a ROC plot, a curve approaching the top-left corner indicates near-perfect performance, whereas a curve on the diagonal line indicates random guessing.

Figure 20A does not show a lot of difference in model performance between the tested models for SUV4 segmentation. When inspecting the ROC curves more closely, a steep initial rise is detected for models rf and s2. Following this initial rise, the curve of the rf model bends towards the diagonal, while the curve of the s2 model remains further away from the diagonal, indicating a stronger overall performance. Model s1 does not have this steep initial rise, but remains furthest from the diagonal in other regions of the curve, which can also indicate strong performance. Around the threshold value of 1.0, the pw model performs worse than random guessing. As it approaches the threshold value of 0.5, the pw model seems to recover, but still seems to perform the worst of all the models. This seemingly low model performance, is confirmed by the AUC values in Table 6. These values also indicate that models s1 and s2 are the best performers for SUV4 segmentation. When comparing these two models, model s1 has a higher average AUC value, but also has a higher standard deviation compared to model s2.

Figure 20B clearly shows that the uni model has the weakest performance for MV2 segmentation, as its curve lies close to the diagonal. All other models demonstrate similar performance at the beginning and end of their respective curves. However, in the middle region, the simple models (s1, s2, and s3) can be clearly distinguished from the more complex models (pw, rfe, and rf), with the simple models showing better performance. These visual findings are confirmed by the AUC values in Table 6, where the average AUC values for the simple models are above 0.75, whereas the average AUC values of the other models are below 0.70. The uni model shows especially weak performance, with an average AUC value of 0.54, indicating nearly random guessing.

When using LIONZ segmentation (Figure 20C), the uni model once again has the weakest performance, with a curve that lies on top of the diagonal. The pw and rf models also seem to perform poorly, with curves close to the diagonal. The remaining models (s1, s2, s3, and rfe) appear to have a similar model performance, with model rfe performing better at thresholds closer to 1.0 and the other models performing better at thresholds closer to 0.0. These visual findings are confirmed by the AUC values in Table 6. The s1, s2, s3, and rfe models have AUC values above 0.74, whereas the other models have AUC values below 0.70. The uni model's AUC value of 0.50 indicates random guessing, as previously thought.

The quantitative AUC values in Table 6 were compared both visually (using the histogram in Figure 21) and statistically (by performing a Friedman test).

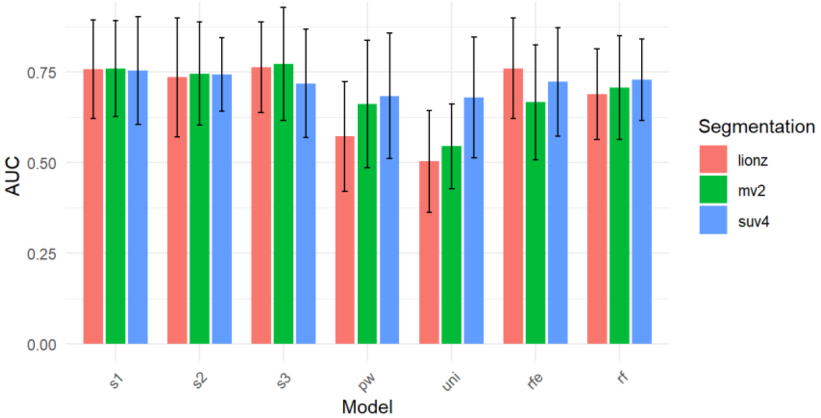


Figure 21: Histogram comparing the AUC values of all cross-validated models using LIONZ, MV2, and SUV4 segmentations with an overall survival of 36 months as prognostic outcome. Orange bars represent the AUC value using LIONZ segmentations, green bars represent the AUC value using MV2 segmentations, and blue bars represent the AUC value using SUV4 segmentations.

Visually, the histogram shows that the three segmentation methods have different performances across the evaluated models. MV2 is the most consistent performer, often producing the highest or second-highest AUC values. It outperforms the other segmentation methods in the simple models. In the other models, MV2 remains competitive, but underperforms compared to SUV4. SUV4 seems the most stable method, demonstrating consistently high performance across the models, indicating a strong performance regardless of model complexity. LIONZ generally underperforms relative to MV2 and SUV4, especially in the pw and uni models. Nevertheless, the performance of LIONZ is comparable to the other segmentation methods in the simple models.

In addition, the Friedman test ranks the segmentation methods, for each individual prognostic model, based on their performance. Then, these rankings are compared across all the models to determine whether there are statistically significant differences in performance between the segmentation methods overall. The test showed a p-value of 0.6065, indicating no significant difference between the three segmentation methods.

In conclusion, the visual findings together with the results of the Friedman test, indicate that there is no significant difference in the performance of the prognostic models for early-stage NSCLC when using different segmentation methods.

9.2.2 External validation results

To confirm the results obtained by cross-validation, the same trained models were applied to external validation testing using the validation cohort. Figure 22 shows the ROC curves of all externally validated prognostic models obtained using MV2 and LIONZ segmentation. As mentioned before, the external dataset did not contain any information about SUV4 segmentations. Therefore, SUV4 segmentation is not discussed in this part of the study.

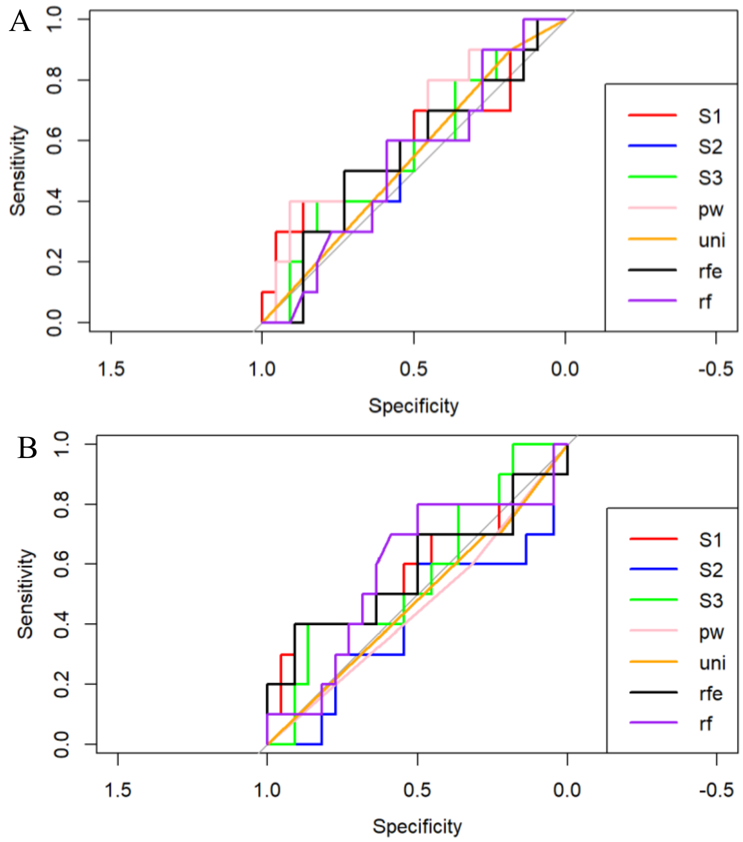


Figure 22: ROC curve of all externally validated prognostic models with an overall survival of 36 months as prognostic outcome. (A) ROC curves of all externally validated models using MV2 segmentation, (B) ROC curves of all externally validated models using LIONZ segmentation.

Table 7 summarizes all averaged AUC values for the externally validated models across each segmentation method.

Table 7: The AUC values of external validation (AUC ext) of every model for every segmentation method with an overall survival of 36 months as prognostic outcome.

OS36	s1	s2	s3	pw	uni	rfe	rf
	MV2 (mean AUC: 0.57)						
AUC ext	0.59	0.57	0.57	0.64	0.54	0.55	0.53
	LIONZ (mean AUC: 0.53)						
AUC ext	0.60	0.42	0.57	0.46	0.48	0.58	0.58

Figure 22A (MV2 segmentation) shows that all models have ROC curves that lie very close to the diagonal reference line, indicating limited predictive power for all models. The pw and uni models show slightly better performance, with their curves floating just above the diagonal. The s2, s3, and rfe models display highly similar patterns, with each model crossing the diagonal line multiple times, pointing to inconsistent prognostic performance. The rf model generally performs the worst, as it crosses the diagonal most often. The AUC values of all models are close to each other and differ by only 0.11 between the extremes, as shown in Table 7. The simple models tend to have higher AUC values than the more complex models, except for the pw model, which has the highest AUC value. Overall, the models demonstrate minimal prognostic performance.

Figure 22B (LIONZ segmentation) shows that the s2, pw, and uni models perform the worst. These models' curves practically always lie below the diagonal reference line, which represents random guessing. Model s1 shows the next weakest performance, as its curve is close to the diagonal line at all thresholds. The remaining three models (s3, rfe, and rf) show similar performance. The three curves alternate as the best-performing model and also have a region where they cross the diagonal. Looking at Table 7, model s1 has the highest AUC value, followed closely by model s3, rfe, and rf. As the ROC curves suggest, models s2, pw, and uni demonstrate the worst model performance. Overall, none of the models demonstrate consistently strong performance.

The AUC values in Table 7 were compared visually using the histogram in Figure 23.

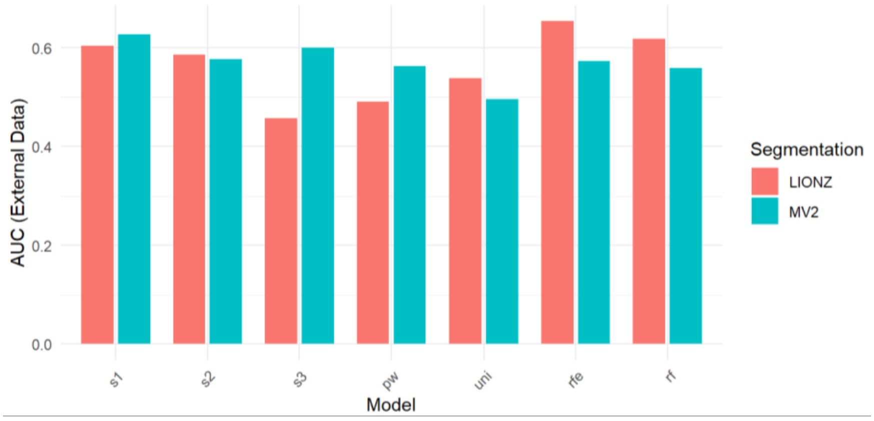


Figure 23: Histogram comparing the AUC values of all externally validated models obtained using LIONZ and MV2 segmentations with an overall survival of 36 months as prognostic outcome. Orange bars represent models using LIONZ segmentations, and blue bars represent models using MV2 segmentation.

The histogram in Figure 23 shows that the bars of MV2- and LIONZ-based prognostic models are nearly the same height for the s1 and s2 models, indicating no significant difference in model performance. Other models, such as s3, pw, and uni have AUC values around 0.5, indicating random guessing. The two most complex models, rfe and rf, show a better performance for LIONZ segmentation than for MV2 segmentation.

In conclusion, there is no consistency in how a segmentation method performs across the different prognostic models with external validation. The results of external validation do not add value to the conclusion made from the cross-validation.

9.3 Best-performing model

A second objective of this thesis is to determine the best-performing prognostic model overall and the best model-segmentation combination. To determine this, the AUC values from Table 6 and 7 were used to perform a DeLong test for every model across all segmentation methods.

9.3.1 DeLong test on cross-validation results

First, a DeLong test was performed on the AUC values from the cross-validation. This determined if there were statistically significant differences between the different prognostic models. The resulting p-values are displayed in a comparison matrix for all prognostic models within each segmentation method. Cells highlighted in green indicate a significant difference between two compared models. If a difference was found, the number of the best-performing model is shown below the p-value. Figure 24 shows the comparison matrix from the DeLong test performed on the seven prognostic models using SUV4, MV2, and LIONZ segmentation.

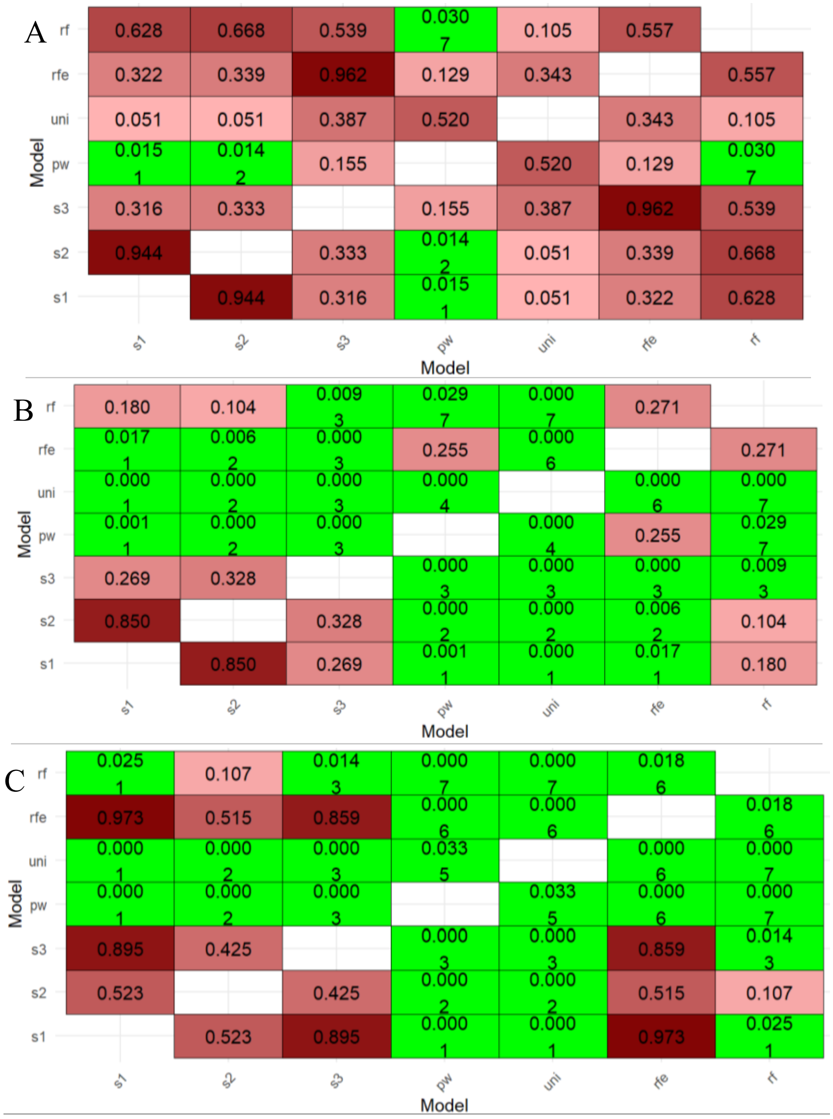


Figure 24: Comparison matrix of the DeLong test performed on the seven cross-validated prognostic models. (A) Comparison matrix for models using SUV4 segmentation, (B) Comparison matrix for models using MV2 segmentation, (C) Comparison matrix for models using LIONZ segmentation.

Figure 24A shows that a few significant differences were observed in the cross-validated prognostic models using SUV4 as segmentation method. Only comparisons between s1-pw, s2-pw, and rf-pw were significantly different. In all three cases, the pw model performed the worst. Consequently, the best-performing models using SUV4 segmentation are s1, s2, and rf.

Figure 24B shows many significant differences between the AUC values of the cross-validated prognostic models using MV2 segmentation. All models were significantly different, with exception of the s1-s2, s1-s3, s1-rf, s2-s3, s2-rf, and pw-rfe models. Overall, the uni model was the worst-performing model. Comparing all significantly different models, the best-performing models using MV2 segmentation are s1, s2, and s3.

Figure 24C illustrates many significant differences between the AUC values of the cross-validated prognostic models using LIONZ segmentation. All models were significantly different, with the exception of the s1-s2, s1-s3, s1-rfe, s2-s3, s2-rfe, s2-rf, and s3-rfe models. Overall, uni and pw were the worst-performing models. Of the significantly different models, s1, s3, and rfe are the best-performing models for LIONZ.

After identifying the best-performing models for each segmentation method, the next step is to determine the best overall model and the model-segmentation combination. First, the average AUC value was calculated for each model across all segmentation methods, regardless of significant differences. Calculating the average AUC value helps to summarize a model's overall performance and shows which model performs better, regardless of the segmentation method. Model s1 had the highest average AUC (0.76 ± 0.14), followed by model s3 with an average AUC of 0.75 ± 0.15.

Next, the AUC values for each significantly different combination were examined by determining the highest lower bound of the 95% confidence interval. This was calculated using (10) and (11), with *n* equal to 50, as a 5-fold 10-times cross-validation was performed.

$$lower\ bound\ 95\%\ CI = AUC - 1.96 * SE \tag{10}$$

$$SE = \frac{SD}{\sqrt{n}} \tag{11}$$

The model-segmentation combination with the highest lower bound was model s3 with MV2 segmentation, with a lower bound of 0.729 and an AUC of 0.77 ± 0.16, as shown in Table 6.

9.3.2 Confusion matrices for cross-validation results

Confusion matrices can be used to confirm or deny the results coming from the DeLong test. Figure 25 shows the confusion matrices with a threshold of 0.5 for the significantly different model-segmentation combinations.

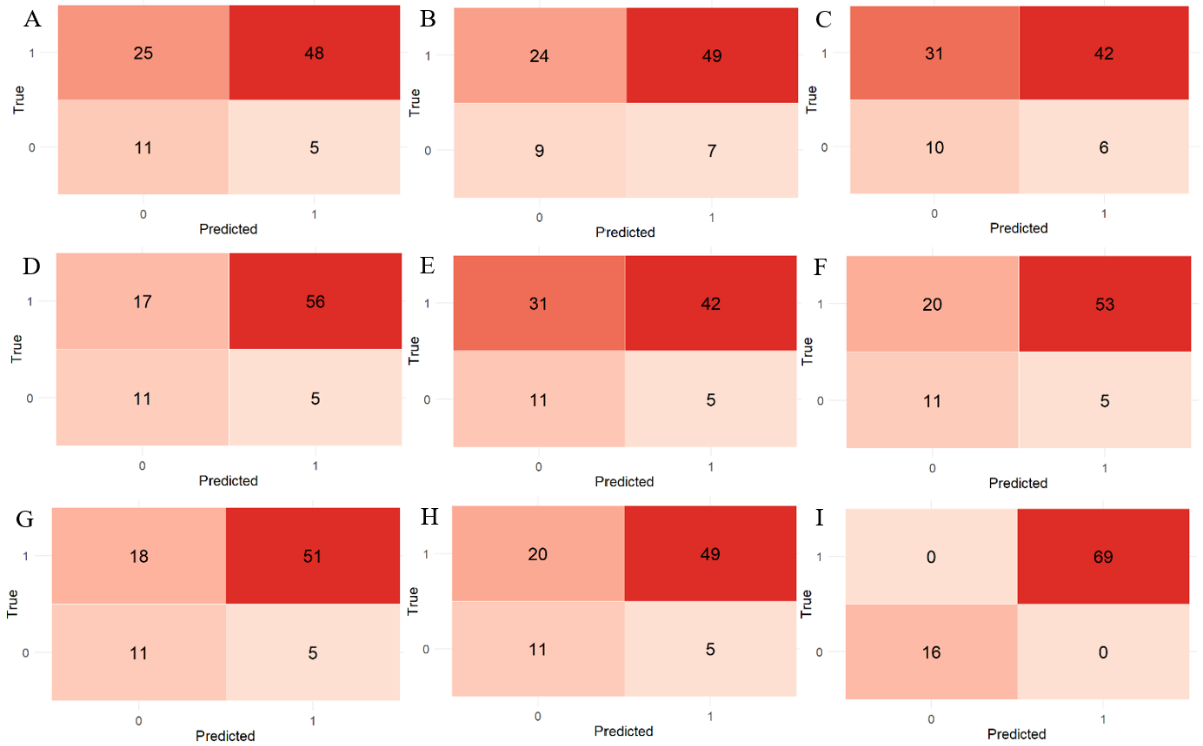


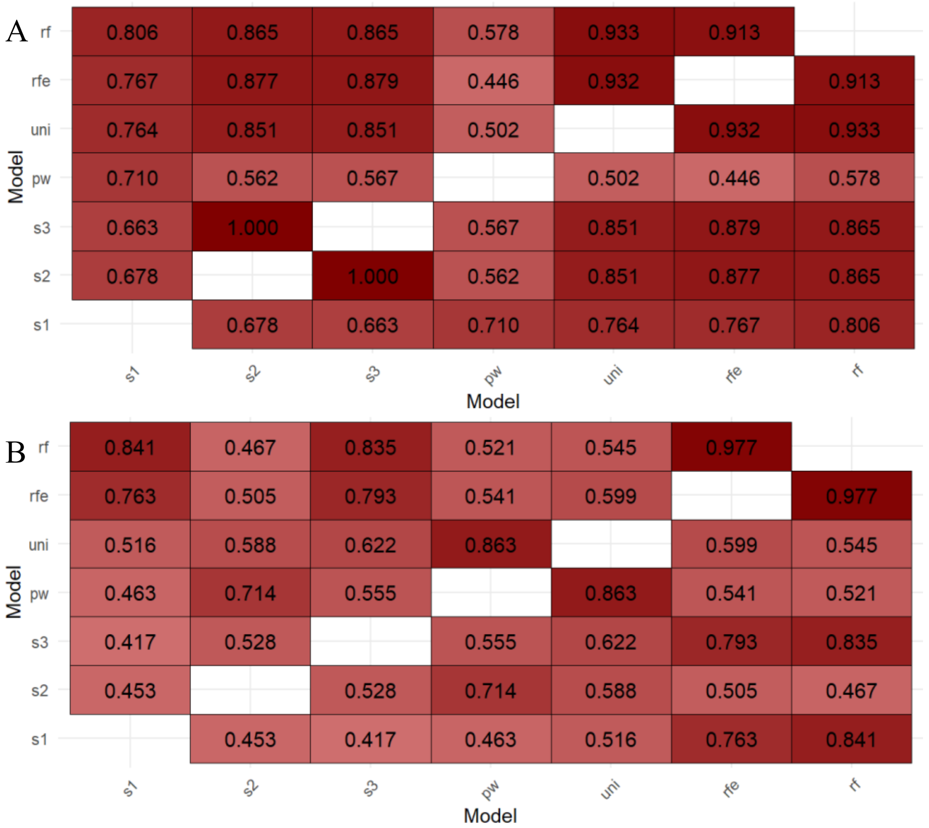
Figure 25: Confusion matrices with a threshold of 0.5 for all cross-validated model-segmentation combinations that were proven statistically significantly different by the DeLong test. (A) Model s1 with SUV4 segmentation, (B) Model s3 with SUV4 segmentation, (C) Model rfe with SUV4 segmentation, (D) Model s1 with MV2 segmentation, (E) Model s2 with MV2 segmentation, (F) Model s3 with MV2 segmentation, (G) Model s1 with LIONZ segmentation, (H) Model s2 with LIONZ segmentation, (I) Model rf with LIONZ segmentation.

As mentioned above, the results of the 95% confidence interval indicate that the best-performing model-segmentation combination is model s3 with MV2 segmentation. Figure 25F shows the corresponding confusion matrix for this combination at a threshold of 0.5. To have a strong prognostic model for cancer, an accurate identification of TP and TN is required. Additionally, achieving a low FN rate is important, as this can result in undertreatment of patients. When comparing the s3-MV2 combination to combinations using SUV4 segmentation (Figure 25A, 25B, and 25C), the s3-MV2 combination is able to identify more TP and TF while simultaneously identifying fewer FN, pointing towards a better performance of the s3-MV2 combination. The same applies to combinations using the s2 model (Figures 25E and 25H) and to the s1-LIONZ combination (Figure 25G). However, when comparing the s3-MV2 combination with the s1-MV2 combination in Figure 25D, the latter shows a better performance, with higher TP and lower FN values. Figure 25I shows that the rf-LIONZ combination exhibits extreme behavior, as it predicted every case correctly. This indicates a bias, as perfect prediction is nearly impossible and can therefore not be trusted as a correct indicator of the overall performance of this combination.

The results from the confusion matrix do not indicate the s3-MV2 combination as the best, but rather the third-best-performing combination. This still indicates a strong model-segmentation performance. Furthermore, it must be kept in mind that confusion matrices only provide results for a specific threshold and do not offer an overall performance evaluation. This may explain why these results differ from those of the 95% confidence interval seen previously.

9.3.3 External validation results

Additionally, a DeLong test was also performed on the AUC values derived from the external validation. Figure 26 shows the comparison matrices from the DeLong test performed on the externally validated prognostic models using MV2 and LIONZ segmentation.



A

Model	s1	s2	s3	pw	uni	rfe	rf
rf	0.806	0.865	0.865	0.578	0.933	0.913	
rfe	0.767	0.877	0.879	0.446	0.932		0.913
uni	0.764	0.851	0.851	0.502		0.932	0.933
pw	0.710	0.562	0.567		0.502	0.446	0.578
s3	0.663	1.000		0.567	0.851	0.879	0.865
s2	0.678		1.000	0.562	0.851	0.877	0.865
s1		0.678	0.663	0.710	0.764	0.767	0.806

B

Model	s1	s2	s3	pw	uni	rfe	rf
rf	0.841	0.467	0.835	0.521	0.545	0.977	
rfe	0.763	0.505	0.793	0.541	0.599		0.977
uni	0.516	0.588	0.622	0.863		0.599	0.545
pw	0.463	0.714	0.555		0.863	0.541	0.521
s3	0.417	0.528		0.555	0.622	0.793	0.835
s2	0.453		0.528	0.714	0.588	0.505	0.467
s1		0.453	0.417	0.463	0.516	0.763	0.841

Figure 26: Comparison matrix of the DeLong test performed on the seven prognostic models using external validation data. (A) Comparison matrix for the externally validated models using MV2 segmentation, (B) Comparison matrix for the externally validated models using LIONZ segmentation.

As shown in Figure 26A, no significant differences were observed between the AUC values of the externally validated prognostic models using MV2 segmentation. Figure 26B also shows that no significant differences were observed for the externally validated prognostic models using LIONZ segmentation.

Based on these results, it can be concluded that testing with external validation did not provide any additional information about the best-performing prognostic model.

10 Discussion

This research evaluated seven prognostic models, which were divided into three categories. The first category includes simple logistic regression models that use one to three conventional PET metrics such as TLG, SUVmax, and DmaxBulk (s1, s2, and s3). The second category includes intermediate logistic regression models in which radiomic features are selected using three different feature selection methods (pw, uni, and rfe). The third category includes a complex machine learning model (rf). The performance of all models was tested using three different segmentation methods (SUV4, MV2, and LIONZ) and four different prognostic outcomes (OS24, OS36, TTP24, and TTP36). This setup led to two main findings that could improve the radiomic workflow, leading to better prognostic models for early-stage NSCLC:

- there is no statistically significant difference in prognostic model performance for early-stage NSCLC when using different segmentation methods;
- the TLG-only model (s1) prevailed as best-performing model overall, and the model using TLG, SUVmax, and DmaxBulk (s3) with an MV2 segmentation was the best-performing model-segmentation combination. These results suggest that simple prognostic models outperform more complex prognostic models.

The first major finding relates to segmentation, the second step of the radiomic workflow. A Friedman test conducted on the cross-validated AUC values from the training cohort yielded a p-value of 0.6065. Since this value exceeds the conventional significance level of 0.05, it indicates that there is no significant difference in model performance across the different segmentation methods evaluated – SUV4, MV2, and LIONZ. This suggests that, from a purely performance-based perspective, the choice of segmentation method does not critically influence the outcome of radiomic models.

Since no significant differences were observed among the different segmentation methods, the choice of method for clinical use must depend on factors other than model performance. Several factors besides model performance can determine the choice of segmentation method. For example, consider the reliability of the segmentation method. Semi-automated segmentation methods are generally more robust and reliable than manual methods [103]. All of the segmentation methods used in this study are semi-automated. However, it has also been shown that AI-based segmentation methods have a higher repeatability than threshold-based methods [104]. Consequently, this factor favors using LIONZ over SUV4 and MV2. Another factor is the clinical understanding of the used segmentation method. Threshold-based segmentation methods are the most frequently used for lung cancer segmentation [105]. Among these methods, SUV methods, such as SUV4, are widely used due to their comprehensive evaluation of cellular metabolism, which is biologically significant and useful to physicians. In contrast, AI-based methods such as LIONZ often operate as “black boxes”, offering less transparency in clinical decision-making. A third factor is human intervention. The extent of human intervention required also plays a significant role in determining the practicality of a segmentation method. Reducing human intervention improves feature reliability and facilitates clinical implementation [103]. In this study, MV2 required the least human intervention, with adjustments needed in only 5.6% (one minor and four major adjustments) of cases, followed by LIONZ at 7.9% (three minor and four major adjustments), and SUV4 at 21.3% (one minor and 18 major adjustments). These findings align with previous research in Classical Hodgkin Lymphoma, which concluded that the most effective segmentation method was the one requiring the least human intervention [67]. Taken together, these factors suggest that while no single segmentation method is definitively superior, each presents a unique balance of strengths and limitations.

One potential limitation regarding this first finding is that all segmentations were performed by a single observer, which could introduce subjectivity [58]. However, prior research shows that semi-automatic tools allow even untrained observers to produce contours comparable to those created manually by trained physicians [103]. Moreover, consistency in using a single segmentation method is emphasized as more critical than inter-observer variability, framing this issue more as a methodological note than a limitation.

The second major finding of this study relates to the performance of different prognostic models. The best-performing model was determined by identifying the model with the highest average AUC value. Model s1, which only includes TLG, showed the best overall performance, closely followed by model s3, which combines TLG, SUVmax, and DmaxBulk. Additionally, DeLong tests confirmed significant differences between models within each segmentation method. When the significant model-segmentation combinations were ranked based on the highest lower boundary of a 95% confidence interval, model s3 with MV2 segmentation showed the highest performance. Interestingly, this model also ranked as the second-best-performing model overall, suggesting that combining multiple PET-derived features may improve prognostic accuracy.

As mentioned above, the results indicate that simple models using conventional PET metrics, such as TLG, SUVmax, and DmaxBulk, perform better than more complex models. Previous studies have shown that TLG is a strong prognostic parameter for overall survival in patients with advanced-stage NSCLC and lymphoma [106], [107]. However, the evidence for early-stage NSCLC remains limited and somewhat inconsistent. The consistent inclusion of TLG in the best-performing models in this study cautiously supports its prognostic value in early-stage NSCLC. Although DmaxBulk contributed to the performance of model s3, its influence in this study is limited because early-stage NSCLC rarely contains metastasis. Only seven out of 89 patients in this study had a non-zero DmaxBulk value, yet its inclusion still improved model performance. This was shown by the strong performance of model s3 and suggests that even minimal spatial information can enhance prognostic accuracy when combined with metabolic features, highlighting the potential value of spatial features in NSCLC [71]. Although both models, s2 and s3, include SUVmax, previous research has shown that its significance as a prognostic parameter for OS varies based on staging and other factors [106], [108], [109]. This may explain why model s2, which combines TLG and SUVmax, performed worse than model s3, which also included DmaxBulk. These results suggest that SUVmax does not meaningfully improve prognostic accuracy, whereas spatial features like DmaxBulk and metabolic features like TLG can strengthen model performance.

Overall, the simpler models (s1, s2, and s3) performed better than the intermediate (pw, uni, and rfe) and complex (rf) models. One possible explanation is that this study focused on early-stage NSCLC, which mostly includes rather small lesions. Conventional PET metrics are closely correlated with features extracted from the small lesions; thus, integrating extra radiomic features into the models does not provide additional value [110]. Another possible explanation is overfitting. Although complex models were restricted to their top five features, the small dataset (n=89) increases the risk of overfitting, potentially causing the models to learn noise instead of meaningful patterns. This leads to models that perform well on the training data but fail to generalize to unseen data [52]. This concern is further supported by the results of unseen, external validation data. In both segmentation and model performance analyses, no statistically significant results were obtained in the external validation cohort. Two primary factors may explain this. First, the validation cohort was limited to only 32 patients. A commonly used guideline recommends including at least ten patients per feature used in a model [85]. This threshold was achieved for the simple methods but not for the more complex models, which use their top five features. Small validation cohorts reduce statistical power and increase variability, resulting in higher p-values and wider confidence intervals. This is a frequent problem in radiomics research because the field is still relatively new, and available datasets remain limited in size [111]. A second possible reason for the failing external validation is the use of a different PET scan in the validation dataset compared to the training dataset. Studies have shown that using different PET scanners, especially those from different manufacturers, can produce varying SUV measurements [112]. This variation increases when scanning small lesions because the partial volume effect leads to an underestimation of SUV due to limited image resolution [113]. This study exhibits both scanner variability and the presence of small lesions, suggesting that SUV measurements from the training cohort may differ from those in the validation cohort.

Looking ahead, there are several interesting directions for future research. First, the findings from this study can be externally validated. This can be achieved by using a larger external dataset to confirm the reliability of the finding and overcome current limitations, such as overfitting, as well as by segmenting the patients in the validation cohort using SUV4. This would allow for a comparison of all three segmentation methods. In addition, the training cohort can be expanded by adding more patients and including all visible lymph nodes on the PET scan, including those that are not clinically positive. Another promising step would be to combine features from the current prognostic models, such as the TLG-only model, with metabolomic data. This could lead to more accurate models and offer new insights into the connection between lung cancer tissue and metabolism. These insights could eventually help to improve early detection and reduce mortality. It would also be valuable to compare the current models with deep learning approaches, as these have demonstrated significant potential in related fields. This comparison could reveal ways to improve performance further.

In the long term, this study could support the development of more accurate prognostic models for early-stage NSCLC. Using PET/CT imaging to predict disease progression and survival could help to determine further treatment and provide patients with clearer information about their prognosis, potentially improving their peace of mind and overall quality of life.

11 Conclusion

During this research, the PET/CT scans of 121 patients were analyzed and divided into a training cohort of 89 patients and an independent validation cohort of 32 patients. Three tumor segmentation methods were used: SUV4, MV2, and an AI-based method, LIONZ. From all segmented images, radiomic parameters were extracted and evaluated across seven different prognostic models. Three models were simple logistic regression models using preselected conventional PET metrics such as TLG, SUVmax, and DmaxBulk; three models were intermediate logistic regression models using various feature selection methods; and the final model was a complex random forest model. All segmentation methods and models were tested across four different prognostic outcomes, with OS36 selected as the primary outcome. The training cohort was used to train and cross-validate the prognostic models, whereas the validation cohort was used for external validation.

The objectives of this thesis are to investigate whether the choice of tumor segmentation method has an impact on the performance of different prognostic models for early-stage NSCLC, and additionally to determine which model is preferred to create the most accurate prognostic model for early-stage NSCLC.

The results show that the choice of tumor segmentation method does not significantly affect the performance of prognostic models for early-stage NSCLC. This conclusion is supported by the Friedman test on cross-validated AUC values, which found no statistically significant differences between segmentation methods across the different models. While external validation was performed, the cohort size was too small to draw statistically meaningful conclusions. Regarding the second objective, model comparison using the DeLong test demonstrated that simple prognostic models outperformed more complex models, such as logistic regression after feature elimination or random forest. The TLG-only model (s1) achieved the highest average AUC, while the best model-segmentation combination was the s3 model with MV2 segmentation, which includes TLG, SUVmax, and DmaxBulk. These findings suggest that simple and easy-to-interpret models can be effective in clinical practice as long as the imaging data is segmented consistently. Overall, this study supports the use of straightforward radiomic models for predicting prognostic outcomes in early-stage NSCLC.

References

[1] World Health Organization, "Global cancer burden growing, amidst mounting need for services," 1 February 2024. [Online]. Available: https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services. [Accessed 13 March 2025].

[2] National cancer institute, "Cancer Staging," 14 October 2022. [Online]. Available: https://www.cancer.gov/about-cancer/diagnosis-staging/staging. [Accessed 13 March 2025].

[3] R. J. Gillies, P. E. Kinahan and H. Hricak, "Radiomics: Images Are More than Pictures, They Are Data," *Radiology,* vol. 278, no. 2, 2015.

[4] G. Lee, H. Park, S. H. Bak and H. Y. Lee, "Radiomics in Lung Cancer from Basic to Advanced: Current Status and Future Directions," *Korean journal of radiology,* vol. 21, no. 9, pp. 159-171, 2020.

[5] P. Bisoyi, "A brief tour guide to cancer disease," in *Understanding cancer*, New Delhi, Academic Press, 2022, pp. 1-20.

[6] T. A. Brown, "Regulation of genome activity," in *Genomes*, Garland Science, 2002.

[7] Stichting tegen kanker, "Kanker," [Online]. Available: https://kanker.be/kanker/kanker/. [Accessed 3 May 2025].

[8] National cancer institute, "What is cancer?," 11 October 2021. [Online]. Available: https://www.cancer.gov/about-cancer/understanding/what-is-cancer. [Accessed 3 May 2025].

[9] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell,* vol. 100, no. 1, pp. 57-70, 2000.

[10] D. Hanahan, "Hallmarks of cancer: new dimensions," *Cancer discovery,* vol. 12, no. 1, pp. 31-46, 12 January 2022.

[11] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *Cell,* vol. 144, no. 5, pp. 646-674, 4 March 2011.

[12] E. J. Morris and N. J. Dyson, "Retrinoblastoma protein partners," *Advances in cancer research,* vol. 82, pp. 1-54, 2001.

[13] H. Wang, M. Guo, H. Wei and Y. Chen, "Targetting p53 pathways: mechanisms, structures and advances in therapy," *Signal transduction and targeted therapy,* vol. 8, no. 92, 2023.

[14] Y. Xu, J. Cui and D. Puett, "Basic cancer biology," in *Cancer bioinformatics*, London, Springer, 2014, pp. 1-39.

[15] C. Sumner, "Hallmarks of cancer: activation invasion and metastasis," Cell signaling technology, [Online]. Available: https://blog.cellsignal.com/hallmarks-of-cancer-activation-invasion-and-metastasis#:~:text=Cancer%20cells%20invade%20local%20tissue,cells%20expand%20into%20nearby%20environments. [Accessed 3 May 2025].

[16] G. C. Blobe, W. P. Schiemann and H. F. Lodisch, "Role of transforming growth factor beta in human disease," *The new England journal of medicine,* vol. 342, no. 18, pp. 1350-1358, 4 May 2000.

[17] M. G. Vander Heiden, L. C. Cantley and C. B. Thompson, "Understanding the Warburg effect: the metabolic requirements of cell proliferation," *Science,* vol. 324, no. 5930, pp. 1029-1033, 22 May 2009.

[18] C. Summer, "Hallmarks of cancer: tumor-promoting inflammation," Cell signaling technology, 2025. [Online]. Available: https://blog.cellsignal.com/hallmarks-of-cancer-tumor-promoting-inflammation. [Accessed 14 May 2025].

[19] C. Summer, "Hallmarks of cancer: genome instability and mutation," Cell signaling technology, 2025. [Online]. Available: https://blog.cellsignal.com/hallmarks-of-cancer-genome-instability-and-mutation. [Accessed 14 May 2025].

[20] P. Bisoyi, "Malignant tumors - as cancer," in *Understanding cancer: from basics to therapeutics*, New Delhi, Academic Press, 2022.

[21] American cancer society, "Cancer staging," 10 September 2024. [Online]. Available: https://www.cancer.org/cancer/diagnosis-staging/staging.html. [Accessed 3 May 2025].

[22] A. Mahmood and R. Srivastava, "Etiology of cancer," in *Understanding cancer: from basics to therapeutics*, New Delhi, Academic Press, 2022, pp. 37-62.

[23] NHS, "Lung cancer," 1 November 2022. [Online]. Available: https://www.nhs.uk/conditions/lung-cancer/. [Accessed 3 May 2025].

[24] American cancer society, "Lung cancer survival rates," 29 January 2024. [Online]. Available: https://www.cancer.org/cancer/types/lung-cancer/detection-diagnosis-staging/survival-rates.html. [Accessed 3 May 2025].

[25] F. O. Stephens and K. R. Aigner, "Lung cancer (bronchogenic carcinoma)," in *Basics of oncology*, Springer, 2016, pp. 141-146.

[26] W. W. Tan, "Non-Small Cell Lung Cancer (NSCLC) - Background," Medscape, 21 December 2024. [Online]. Available: https://emedicine.medscape.com/article/279960-overview#a2. [Accessed 9 March 2025].

[27] M. C. Stöppler, "Non-Small-Cell Lung Cancer (NSCLC)," eMedicineHealth, 28 January 2025. [Online]. Available: https://www.emedicinehealth.com/non-small-cell_lung_cancer/article_em.htm. [Accessed 9 March 2025].

[28] National cancer institute, "Non-small cell lung cancer treatment - Patient version," 28 March 2025. [Online]. Available: https://www.cancer.gov/types/lung/patient/non-small-cell-lung-treatment-pdq. [Accessed 3 May 2025].

[29] Australian radiation protection and nuclear safety agency, "X-rays," [Online]. Available: https://www.arpansa.gov.au/understanding-radiation/what-is-radiation/ionising-radiation/x-ray. [Accessed 3 May 2025].

[30] M. Berger, Q. Yang and A. Maier, "X-ray Imaging," in *Medical imaging systems: an introductory guide*, vol. 11111, A. Maier, S. Steidl, V. Christlein and J. Hornegger, Eds., Springer, 2018, pp. 119-144.

[31] G. Lloyd-Jones, "Basics of X-ray physics - X-ray production," Radiology masterclass, February 2016. [Online]. Available: https://www.radiologymasterclass.co.uk/tutorials/physics/x-ray_physics_production. [Accessed 3 May 2025].

[32] T. Suzuki, "Irradiation versus radiation," Pulstec, 28 March 2024. [Online]. Available: https://www.pulstec.net/what-is-x-ray-irradiation/. [Accessed 3 May 2025].

[33] C. T. Badea, "Principles of micro X-ray computed tomography," in *Molecular imaging: principles and practice*, Durham, Academic Press, 2021, pp. 47-64.

[34] D. Pavlenko, *De rol van 18F-FDG PET/CT in de uitwerking van aseptische inflammatie*, Gent, 2019.

[35] G. Lloyd-Jones, "Basics of X-ray physics - Tissue densities," Radiology masterclass, March 2016. [Online]. Available: https://www.radiologymasterclass.co.uk/tutorials/physics/x-ray_physics_densities#top_1st_img. [Accessed 3 May 2025].

[36] G. S. Kushwaha, N. S. Bhavesh, N. Misra and M. Suar, "Biomedical techniques in cellular and molecular diagnostics: journey so far and the way forward," in *Biomedical imaging instrumentation*, Academic press, 2021.

[37] B. Reniers, *X-Ray, Computed Tomography and cone beam CT* [cursus], Diepenbeek: Gezamenlijke opleiding Industriële Ingenieurswetenschappen UHasselt & KU Leuven, 2024, p. 10.

[38] R. R. Gharieb, "X-ray and computed tomography scan imaging: instrumentation and medical applications," in *Computed-tomography (CT) scan*, IntechOpen, 2022.

[39] M. Momcilovic and D. B. Shackelford, "Imaging cancer metabolism," *Biomolecules & therapeutics,* vol. 26, no. 1, pp. 81-92, 7 Dec 2017.

[40] A. van der Plas, "X-ray/CT technique," 10 Juli 2023. [Online]. Available: https://www.radiology.expert/en/modules/xrayct-technique/types-of-ct-techniques/. [Accessed 3 May 2025].

[41] R. A. Racicot, "Fossil secrets revealed: X-ray CT scanning and applications in paleontology," *The paleontological society papers,* vol. 22, pp. 21-38, 2017.

[42] P. Cheebsumon, et al., "Assessment of tumour size in PET/CT lung cancer studies: PET- and CT-based methods compared to pathology," *EJNMMI Research,* vol. 2, no. 56, 2012.

[43] M. Bai, *Synthesis and application of targeted molecular imaging agents for enhanced disease imaging and therapy*, 2007.

[44] A. Zhu, D. Lee and H. Shim, "Metabolic PET imaging in cancer detection and therapy response," *Seminars in oncology,* vol. 38, no. 1, pp. 55-69, 2011.

[45] D. Carrick, J. Dickson and A. Bradley, "Basic principles of PET/CT imaging," in *PET/CT imaging: basics and practice*, Springer, 2022, pp. 1-12.

[46] B. Kemp, in *PET-CT and PET-MRI: a practical guide*, Berlin, Springer, 2012, pp. 3-18.

[47] B. Reniers, *Positron Emission Tomography/Single Photon Emission Computed Tomography* [cursus], Diepenbeek: Gezamenlijke opleiding Industriële Ingenieurswetenschappen UHasselt & KU Leuven, 2024.

[48] N. D. Volkow, N. A. Mullani and B. Bendriem, "Positron emission tomography instrumentation: an overview," *American journal of physiologic imaging,* vol. 3, no. 3, pp. 142-153, 1988.

[49] S. Shaikh, "PET-CT imaging and applications," in *Computed-tomography (CT) scan*, IntechOpen, 2022.

[50] S. Kukava and M. Baramia, "Place and role of PET/CT in the diagnosis and staging of lung cancer," in *Advances in radiation oncology in lung cancer*, Springer, 2023, pp. 85-112.

[51] Y. Zhang, A. Oikonomou, A. Wong, M. A. Haider and F. Khalvati, "Radiomics-based prognosis analysis for non-small cell lung cancer," *Scientific reports,* vol. 7, 2017.

[52] M. E. Mayerhoefer, et al., "Introduction to radiomics," *The journal of nuclear medicine,* vol. 61, no. 4, pp. 488-495, 2020.

[53] J. E. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi and B. Baessler, "Radiomics in medical imaging - "how-to" guide and critical reflection," *Insights into imaging,* vol. 11, no. 91, 2020.

[54] C. McCague, et al., "Introduction to radiomics for a clinical audience," *Clinical radiology,* vol. 78, pp. 83-98, 2023.

[55] S. Hawkins, et al., "Predicting malignant nodules from screening CT scans," *Journal of thoracic oncology,* vol. 11, no. 12, 2016.

[56] W. Wu, H. Hu, J. Gong, X. Li, G. Huang and S. Nie, "Malignant-benign classification of pulmonary nodules based on random forest aided by clustering analysis," *Physics in medicine & biology,* vol. 64, no. 3, 2019.

[57] F. Bianconi, I. Palumbo, A. Spanu, S. Nuvoli, M. L. Fravolini and B. Palumbo, "PET/CT radiomics in lung cancer: an overview," *Applied biosciences and bioengineering,* vol. 10, no. 5, 2020.

[58] A. K. Anagnostopoulos, et al., "Radiomics/radiogenomics in lung cancer: basic principles and initial clinical results," *Cancers,* vol. 14, no. 7, 2022.

[59] S. K. Saini, N. Thakur and M. Juneja, "Radiomics based diagnosis with medical imaging: a comprehensive study," *Wireless personal communications,* vol. 130, pp. 481-514, 2023.

[60] B. Foster, U. Bagci, A. Mansoor, Z. Xu and D. J. Mollura, "A review on segmentation of positron emission tomography images," *Computers in biology and medicine,* vol. 50, pp. 76-96, 1 July 2014.

[61] O. Drieskens, "PET/CT," Genk, 2018.

[62] F. C. F. Dionisio, L. S. Oliveira, M. d. A. Hernandes, E. E. Engels, P. M. de Azevedo-Marques and M. H. Nogueira-Barbos, "Manual versus semiautomatic segmentation of soft-tissue sarcomas on magnetic resonance imaging: evaluation of similarity and comparison of segmentation times," *Radiologia brasileira,* vol. 54, no. 3, pp. 155-164, 2021.

[63] M. Hatt, J. A. Lee, C. R. Schmidtlein, I. El Naqa, C. Caldwell and E. De Bernardi, "Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211," *Medical Physics,* vol. 44, no. 6, pp. e1-e42, 2017.

[64] A. Baazaoui, W. Barhoumi, E. Zagrouba and R. Mabrouk, "A survey of PET image segmentation: applications in oncology, cardiology and neurology," *Current medical imaging,* vol. 12, no. 1, pp. 13-27, 2016.

[65] K. Mah and C. B. Caldwell, "Biological target volume," in *PET-CT in radiotherapy treatment planning*, Saunders, 2008, pp. 52-89.

[66] E. Pfaehler, et al., "PET segmentation of bulky tumors: strategies and workflows to improve inter-observer variability," *Plos One,* vol. 15, no. 3, 2020.

[67] J. Driessen, et al., "The impact of semi-automatic segmentation methods on metabolic tumor volume, intensity and dissemination radiomics in 18F-FDG PET scans of patients with classical Hodgkin lymphoma," *Journal of nuclear medicine,* vol. 63, no. 9, pp. 1424-1430, 6 January 2022.

[68] M. Droguet, et al., "Automated segmentation of lesions in [18F]FDG PET/CT images of lung cancer patients: external evaluation of an AI-driven lesion segmentation tool (LION)," *Journal of nuclear medicine,* vol. 65, no. 2, 2024.

[69] A. Zwanenburg, S. Leger, M. Vallières and S. Löck, *The image biomarker standardisation initiative*, 2019.

[70] P. E. Kinahan and J. W. Fletcher, "PET/CT Standardized Uptake Values (SUVs) in Clinical Practice and Assessing Response to Therapy," *Semin Ultrasound CT MR,* vol. 31, no. 6, pp. 496-505, 2010.

[71] S. Pellegrino, et al., "Prognostic Value of Tumor Dissemination (Dmax) Derived from Basal 18F-FDG Positron Emission Tomography/Computed Tomography in Patients with Advanced Non-Small-Cell Lung Cancer," *Biomedicines,* vol. 13, no. 2, p. 477, 2025.

[72] D. Kiptoon, "Feature selection in machine learning," 18 August 2023. [Online]. Available: https://medium.com/@jdkiptoon/feature-selection-in-machine-learning-20417d052b80. [Accessed 13 May 2025].

[73] Intellectus consulting, "Correlation (Pearson, Kendall, Spearman)," [Online]. Available: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/correlation-pearson-kendall-spearman/. [Accessed 13 May 2025].

[74] S. Tu, C. Li and B. E. Shepherd, "Between- and within-cluster spearman rank correlations," *Statistics in medicine,* vol. 44, no. 3-4, p. e10326, 24 January 2024.

[75] A. Liu, S. Raza and A. McGuire, "Feature selection methods," Medium, 31 October 2023. [Online]. Available: https://2os.medium.com/feature-selection-methods-25ebe0e08896. [Accessed 13 May 2025].

[76] J. Cohen, P. Cohen, S. G. West and L. S. Aiken, "Rank correlation," in *Applied multiple regression/correlation analysis for the behavioral sciences*, 3th ed., New Jersey, Lawrence Erlbaum Associates, 2003, pp. 31-32.

[77] T. Bex, "How to use pairwise correlation for robust feature selection," Medium, 13 April 2021. [Online]. Available: https://medium.com/data-science/how-to-use-pairwise-correlation-for-robust-feature-selection-20a60ef7d10. [Accessed 13 May 2025].

[78] E. Johnson, "Effective feature selection methods in machine learning," Synapse waves, [Online]. Available: https://synapsewaves.com/articles/feature-selection-techniques-machine-learning/?utm_source=chatgpt.com. [Accessed 19 May 2025].

[79] C. Lai, M. J. T. Reinders, L. J. van 't Veer and L. F. Wessels, "A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets," *BMC Bioinformatics,* vol. 7, no. 235, 2 May 2006.

[80] D. Dissanayake and R. Navarathna, "Feature selection with univariate filtering," Medium, 17 January 2023. [Online]. Available: https://octave-jkh.medium.com/feature-selection-with-univariate-filtering-41c6061579e5. [Accessed 13 May 2025].

[81] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in bioinformatics,* vol. 2, 27 June 2022.

[82] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning,* vol. 46, pp. 389-422, January 2002.

[83] M. Johannes, et al., "Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients," *Bioinformatics,* vol. 26, no. 17, pp. 2136-2144, September 2010.

[84] H. M. Castro and J. C. Ferreira, "Linear and logistic regression models: when to use and how to interpret them?," *Jornal brasileiro de pneumologia,* vol. 48, no. 6, p. e20220439, 2022.

[85] E. C. Zabor, C. A. Reddy, R. D. Tendulkar and S. Patil, "Logistic regression in clinical studies," *International journal of radiation oncology - biology - physics,* vol. 122, no. 2, pp. 271-277, 2022.

[86] J. Tolles and W. J. Meurer, "Logistic regression: relating patient characteristics to outcomes," *Jama,* vol. 316, no. 5, pp. 533-534, 2016.

[87] C. F. Uribe, et al., "Machine learning in nuclear medicine: part 1 - introduction," *The journal of nuclear medicine,* vol. 60, no. 4, pp. 451-458, 2019.

[88] L. Barrenada, P. Dhiman, D. Timmerman, A.-L. Boulesteix and B. Van Calster, "Understanding overfitting in random forest for probability estimation: a visualization and simulation study," *Diagnostic and prognostic research,* vol. 8, no. 14, 2024.

[89] L. Breiman, "Random forests," *Machine learning,* vol. 45, pp. 5-32, 2001.

[90] S. Ahlawat, "Random forest," in *Statistical quantitative methods in finance*, 2025, pp. 219-239.

[91] N. A. Obuchowski and J. A. Bullen, "Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine," *Physics in medicine and biology,* vol. 63, 2018.

[92] S. Dash, "Understanding the ROC and AUC intuitively," Medium, 19 October 2022. [Online]. Available: https://medium.com/@shaileydash/understanding-the-roc-and-auc-intuitively-31ca96445c02. [Accessed 13 May 2025].

[93] S. K. çorbacioglu and G. Aksel, "Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value," *Turkish journal of emergency medicine,* vol. 23, no. 4, pp. 195-198, 2023.

[94] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters,* vol. 27, pp. 861-874, 2006.

[95] R. Draelos, "Comparing AUCs of machine learning models with DeLong's test," Glass box, 4 February 2020. [Online]. Available: https://glassboxmedicine.com/2020/02/04/comparing-aucs-of-machine-learning-models-with-delongs-test/. [Accessed 21 May 2025].

[96] E. R. DeLong, D. M. DeLong and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics,* vol. 44, no. 3, pp. 837-845, 1988.

[97] R. Eisinga, T. Heskes, B. Pelzer and M. Te Grotenhuis, "Exact p-values for pairwise comparsion of Friedman rank sums, with application to comparing classifiers," *BMC Bioinformatics,* vol. 18, no. 68, 2017.

[98] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association,* vol. 32, no. 200, pp. 675-701, 1937.

[99] E. Alsyed, R. Smith, L. Bartley, C. Marshall and E. Spezi, "A heterogeneous phantom study for investigating the stability of PET images radiomic features with varying reconstruction settings," *Frontiers in Nuclear Medicine,* vol. 3, 2023.

[100] X. Sun and W. Xu, "Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves," *Signal processing letters,* vol. 21, no. 11, 2014.

[101] K. Vanhove, et al., "Correlations between the metabolic profile and 18F-FDG-Positron Emission Tomography-Computed Tomography parameters reveal the complexity of the metabolic reprogramming within lung cancer patients," *Scientific Reports,* vol. 9, 2019.

[102] L. Deckers and R. Truyens, *The potential of radiomics with PET/CT: study of correlations with metabolic profile and its discriminative power* [master's thesis], Diepenbeek: Gezamenlijke opleiding Industriële Ingenieurswetenschappen UHasselt & KU Leuven, 2021.

[103] C. A. Owens, et al., "Lung tumor segmentation methods: Impact on the uncertainty of radiomics features for non-small cell lung cancer," *PloS ONE,* vol. 13, no. 10, 2018.

[104] E. Pfaehler, et al., "Repeatability of two semi-automatic artificial intelligence approaches for tumor segmentation in PET," *EJNMMI Research,* vol. 11, no. 4, 2021.

[105] Y. Zhang, W. Huang, H. Jiao and L. Kang, "PET radiomics in lung cancer: advances and translational challenges," *EJNMMI Physics,* vol. 11, no. 81, 2024.

[106] F. Yildirim, A. S. Yurdakul, S. Özkaya, Ü. Ö. Akdemir and C. Öztürk, "Total lesion glycolysis by 18F-FDG PET/CT is independent prognostic factor in patients with advanced non-small cell lung cancer," *The clinical respiratory journal,* vol. 11, no. 5, pp. 602-611, 2017.

[107] L. Ceriani, et al., "SAKK38/07 study: integration of baseline metabolic heterogeneity and metabolic tumor volume in DLBCL prognostic model," *Blood advances,* vol. 4, no. 6, pp. 1082-1092, 2020.

[108] M. A. Arshad, et al., "Discovery of pre-therapy 2-deoxy-2-18F-fluoro-D-glucose positron emission tomography-based radiomics classifiers of survival outcome in non-small-cell lung cancer patients," *European journal of nuclear medicine and molecular imaging,* vol. 46, no. 2, pp. 455-466, 2019.

[109] F. Na, J. Wang, C. Li, L. Deng, J. Xue and Y. Lu, "Primary tumor standardized uptake value measured on F18-Fluorodeoxyglucose positron emission tomography is of prediction value for survival and local control in non-small-cell lung cancer receiving radiotherapy: meta-analysis," *Journal of Thoracic oncology,* vol. 9, no. 6, pp. 834-842, 2014.

[110] E. Pfaehler, et al., "Plausibility and redundancy analysis to select FDG-PET textural features in non-small cell lung cancer," *Medical physics,* vol. 48, no. 3, pp. 1226-1238, 2021.

[111] T. Konert, J. B. van de Kamer, J.-J. Sonke and W. V. Vogel, "The developing role of FDG PET imaging for prognostication and radiotherapy target volume delineation in non-small cell lung cancer," *Journal of Thoracic Disease,* vol. 10, pp. S2508-S2521, 2018.

[112] K.-W. Park, R. Ashlock, W. B. Chang, J. Lahner, B. Line and C. Kim, "High variation in standardized uptake values among PET systems from different manufacturers," *Journal of nuclear medicine,* vol. 48, p. 185, 2007.

[113] M. C. Adams, T. G. Turkington, J. M. Wilson and T. Z. Wong, "A Systematic Review of the Factors Affecting Accuracy of SUV Measurements," *American journal of roentgenology,* vol. 195, no. 2, 2012.

[114] OpenAI, "ChatGPT," November 30 2022. [Online]. Available: https://chatgpt.com/.

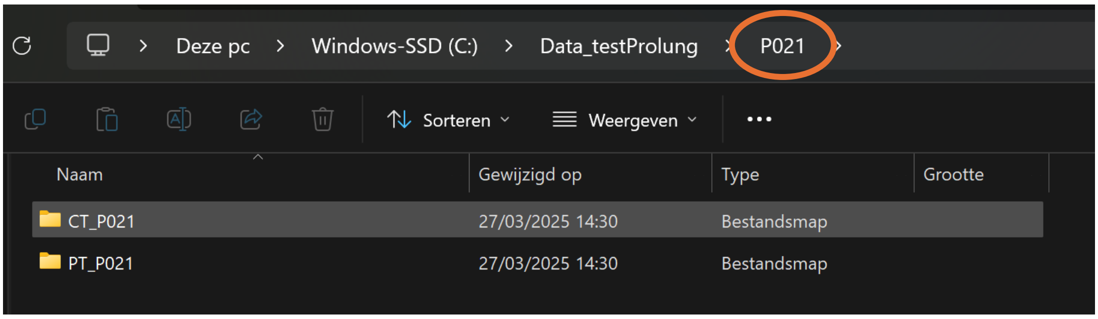
[115] OpenAI, "ChatGPT," June 1 2025. [Online]. Available: https://chatgpt.com/.

List of appendices

Appendix A: An instruction manual on how to extract radiomic data for three segmentation methods using the ACCURATE and RADIOMICS tools (Prof. Dr. Boellaard, Amsterdam UMC)

Note: This instruction manual is written to analyze one patient. All steps need to be repeated when analyzing multiple patients.
Note 2: Before following this instruction manual, the following programs should be installed: ACCURATE, RADIOMICS, LIONZ, and Python 3.11.

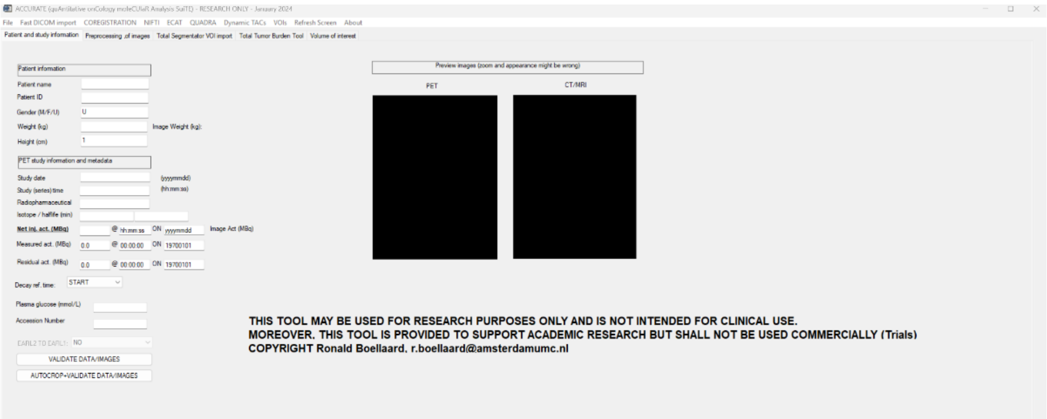
Step 1: First of all, make a map containing the CT and PET scans of the patient. This map is preferably called PXXX, with XXX indicating the pseudonymization number of each patient. For example, the patient map of patient 21 will be called P021.



Naam	Gewijzigd op	Type	Grootte
CT_P021	27/03/2025 14:30	Bestandsmap	
PT_P021	27/03/2025 14:30	Bestandsmap	

Note: Make sure that the path to the patient map does not contain any spaces. Preferably, the patient map is put in a map that is located directly in de C:\ disk of the computer, where the ACCURATE and RADIOMICS tool are located as well. An example of a correct path is shown in the image above.

Step 2: Open the ACCURATE tool (developed by the research team of Prof. Dr. Ronald Boellaard at Amsterdam UMC). The ACCURATE tool should be located on the C:\ disk of the computer and can be opened by launching 'accurate4petct_v10072024'. Then press 'Click to continue' and the application should open. When opened, the application should look like the image below.


ACCURATE (quAntitative onCology moleCUlaR Analysis SuiTE) - RESEARCH ONLY - January 2024
File
Fast DICOM import
COREGISTRATION
NIFTI
ECAT
QUADRA
Dynamic TACs
VOIs
Refresh Screen
About
Patient and study information
Preprocessing .of images
Total Segmentator VOI import
Total Tumor Burden Tool
Volume of interest
Patient information
PET study information and metadata
Preview images (zoom and appearance might be wrong)
PET
CT/MRI
VALIDATE DATA/IMAGES
AUTOCROP+VALIDATE DATA/IMAGES

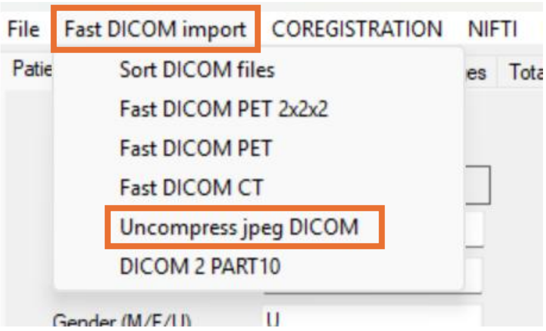
Patient name
Patient ID
Gender (M/F/U) U
Weight (kg) Image Weight (kg):
Height (cm) 1

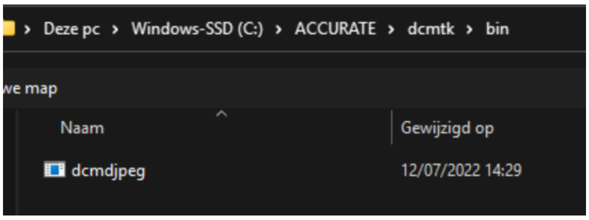
Study date (yyyymmdd)
Study (series) time (hh:mm:ss)
Radiopharmaceutical
Isotope / halflife (min)
Net inj. act. (MBq) @ hh:mm:ss ON yyyymmdd Image Act (MBq)
Measured act. (MBq) 0.0 @ 00:00:00 ON 19700101
Residual act. (MBq) 0.0 @ 00:00:00 ON 19700101
Decay ref. time: START
Plasma glucose (mmol/L)
Accession Number
EARL2 TO EARL1: NO

THIS TOOL MAY BE USED FOR RESEARCH PURPOSES ONLY AND IS NOT INTENDED FOR CLINICAL USE.
MOREOVER. THIS TOOL IS PROVIDED TO SUPPORT ACADEMIC RESEARCH BUT SHALL NOT BE USED COMMERCIALLY (Trials)
COPYRIGHT Ronald Boellaard. r.boellaard@amsterdamumc.nl

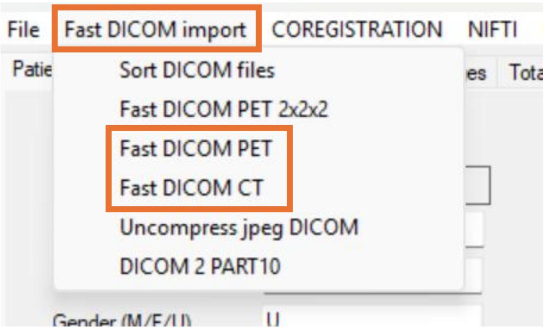
Note: During the use of the application, there is a possibility that it will crash when an error occurs or if you miss-clicked. If this happens, just reopen the application.

Step 3: To be sure that the PET/CT image will be correctly imported into the tool, it is important that uncompressed slices are uploaded. The ACCURATE tool can uncompress files by pressing ‘Fast DICOM import’, followed by ‘uncompress jpeg DICOM’. This is shown in the left figure below.
Next, a window will open. You need to follow the path ‘ACCURATE\dcmtk\bin\dcmdjpeg’. This is shown in the right figure below. Then, a white window will open where you need to indicate the map where the CT or PET slices are located. This step should be done individually for every map that contains either CT or PET slices.





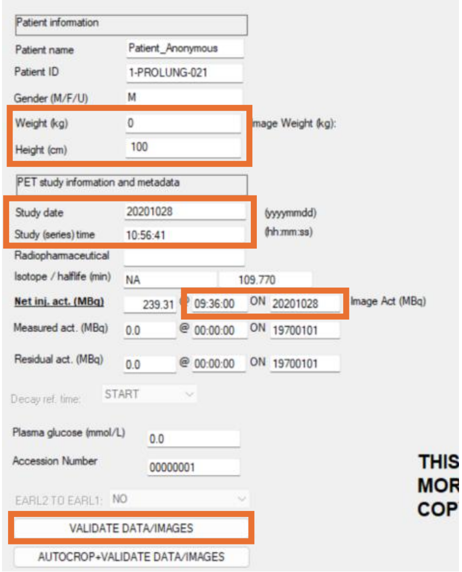
Step 4: When both the CT and PET scan are uncompressed, a project can be made. A project is the overlaying of the CT and PET scan to form a PET/CT. This can be done via the tab ‘Fast DICOM import’ followed by pressing firstly ‘Fast DICOM PET’ and next ‘Fast DICOM CT’. The different tabs are shown in the figure below.



Note: When uploading the CT, two warnings will be displayed after one another. By pressing ‘ok’ these will disappear without giving an error.

Note 2: It can be possible that when uploading the CT scan, it will give a weird image (e.g.: blurry images containing lines). In this case, it is recommended to redownload and re-uncompress the CT scan. If this doesn’t resolve the problem, check whether the CT is a full body CT scan.

Step 5: When the PET/CT is uploaded, essential parameters need to be filled in. The weight and size of the patient should be included, as well as the study date, study time, and moment of injection. The necessary boxes are shown in the picture below. When the boxes are filled in, the button ‘VALIDATE DATA/IMAGES’ should be pressed to validate everything.



Patient information

Patient name: Patient_Anonymous

Patient ID: 1-PROLUNG-021

Gender (M/F/U): M

Weight (kg): 0

Height (cm): 100

mage Weight (kg):

PET study information and metadata

Study date: 20201028 (yyyymmdd)

Study (series) time: 10:56:41 (hh:mm:ss)

Radiopharmaceutical:

Isotope / halflife (min): NA 109.770

Net inj. act. (MBq): 239.31 @ 09:36:00 ON 20201028 Image Act (MBq)

Measured act. (MBq): 0.0 @ 00:00:00 ON 19700101

Residual act. (MBq): 0.0 @ 00:00:00 ON 19700101

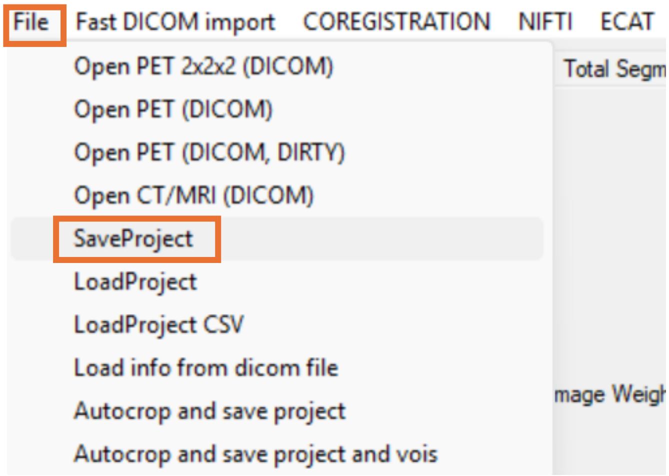
Decay ref. time: START

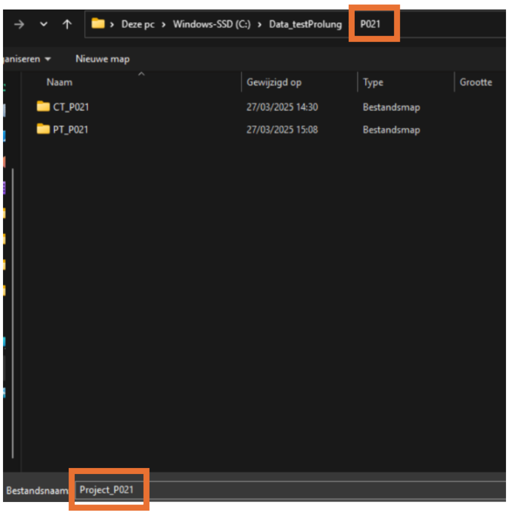
Plasma glucose (mmol/L): 0.0

Accession Number: 00000001

EARL2 TO EARL1: NO

Step 6: When the project is validated, it can be saved. Via the tab 'File', you can press 'SaveProject', which will open a window to save the project. This is shown in the left figure. The project should be saved in the corresponding patient map. To remain consistent, the project needs to be named 'Project_PXXX'. Thus, the project of patient 21 will be saved in the map P021 and will be named 'Project_P021'. This is shown in the right figure.



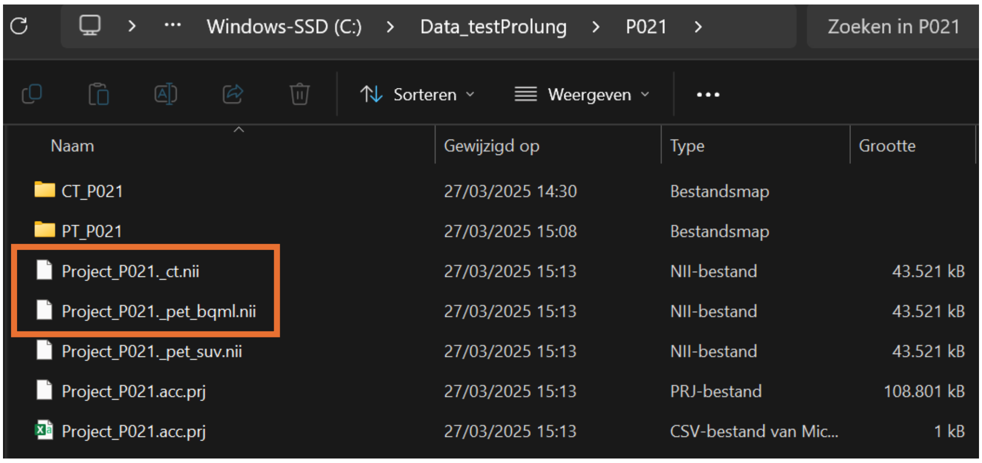


Deze pc › Windows-SSD (C:) › Data_testProlung › P021

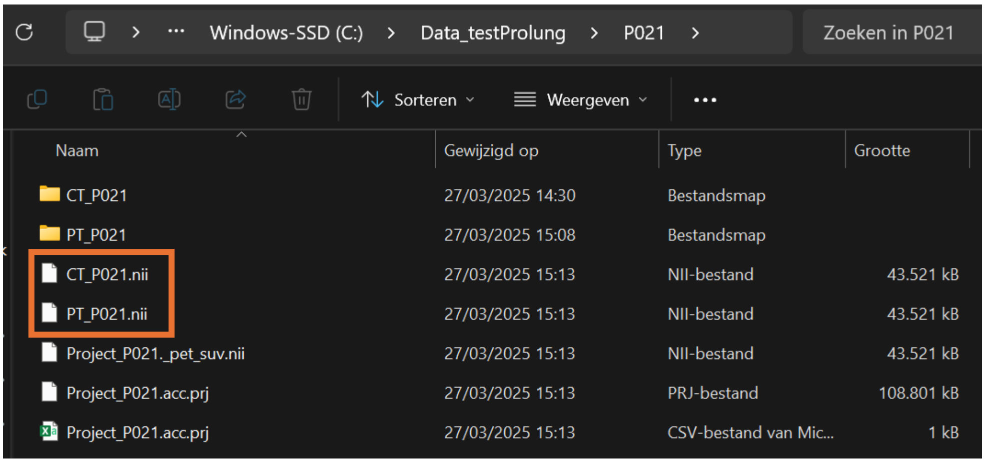
Naam	Gewijzigd op	Type	Grootte
CT_P021	27/03/2025 14:30	Bestandsmap	
PT_P021	27/03/2025 15:08	Bestandsmap	

Bestandsnaam: Project_P021

Step 7: For segmentation with AI, the project files should have a specific name. ACCURATE should be closed, and two files in the patient map should be renamed. The file 'Project_PXXX_ct.nii' should be renamed to 'CT_PXXX', and the file 'Project_PXXX_pet_bqml.nii' should be renamed to 'PT_PXXX'. Both unchanged and changed files are shown in the figures below.



Naam	Gewijzigd op	Type	Grootte
CT_P021	27/03/2025 14:30	Bestandsmap	
PT_P021	27/03/2025 15:08	Bestandsmap	
Project_P021._ct.nii	27/03/2025 15:13	NII-bestand	43.521 kB
Project_P021._pet_bqml.nii	27/03/2025 15:13	NII-bestand	43.521 kB
Project_P021._pet_suv.nii	27/03/2025 15:13	NII-bestand	43.521 kB
Project_P021.acc.prj	27/03/2025 15:13	PRJ-bestand	108.801 kB
Project_P021.acc.prj	27/03/2025 15:13	CSV-bestand van Mic...	1 kB

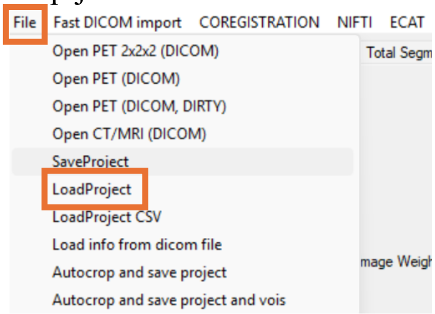


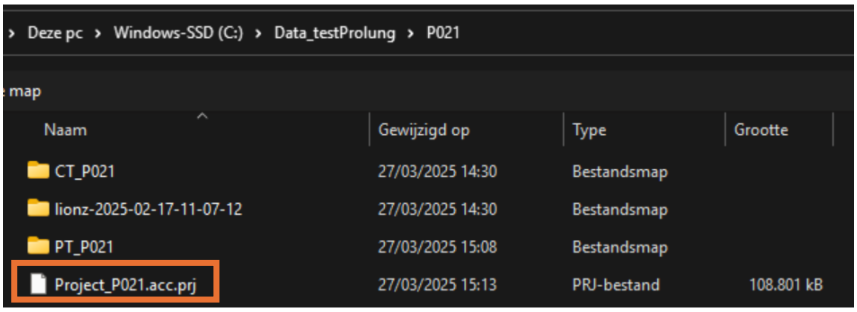
Naam	Gewijzigd op	Type	Grootte
CT_P021	27/03/2025 14:30	Bestandsmap	
PT_P021	27/03/2025 15:08	Bestandsmap	
CT_P021.nii	27/03/2025 15:13	NII-bestand	43.521 kB
PT_P021.nii	27/03/2025 15:13	NII-bestand	43.521 kB
Project_P021._pet_suv.nii	27/03/2025 15:13	NII-bestand	43.521 kB
Project_P021.acc.prj	27/03/2025 15:13	PRJ-bestand	108.801 kB
Project_P021.acc.prj	27/03/2025 15:13	CSV-bestand van Mic...	1 kB

Step 8: When the AI segmentation tool LIONZ is installed, it can be run via the command prompt of the computer. The commands that should be run are:
cd C:\Python\Python311\ LIONZ-env\Scripts
activate
LIONZ -d C:\Data_testProlung -m fdg

Note: Depending on where Python is installed and where the patient files are located, the directories, that are underlined, can differ.
Note 2: While LIONZ is running, steps 9-12B can be performed.

Step 9: The patient project must be loaded again in ACCURATE. By clicking 'File' in the toolbar and selecting 'LoadProject', you can choose the correct patient project in the format of a '.prj' file.

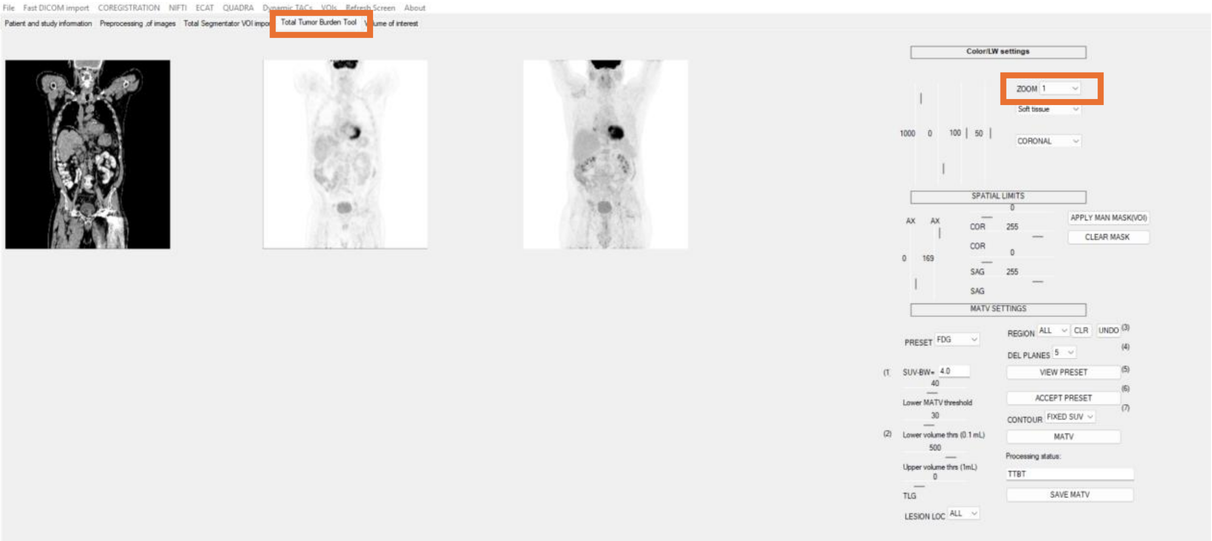




Naam	Gewijzigd op	Type	Grootte
CT_P021	27/03/2025 14:30	Bestandsmap	
lionz-2025-02-17-11-07-12	27/03/2025 14:30	Bestandsmap	
PT_P021	27/03/2025 15:08	Bestandsmap	
Project_P021.acc.prj	27/03/2025 15:13	PRJ-bestand	108.801 kB

Now that the project is loaded, it needs to be validated again by pressing the button 'VALIDATE DATA/IMAGES'.

Step 10: Open the tab ‘Total Tumor Burden Tool’. One CT, one PET, and one PET/CT image should appear, as shown in the figure below. To make the images larger, the ‘ZOOM’ function in the upper right corner can be used.



Step 11A: To do the presegmentation for SUV4, there are several steps, which are indicated with corresponding numbers on the figure below:

i) Set the preset to FDG, now red spots will light up on the images. These spots indicate parts of the body with high glucose consumption levels.

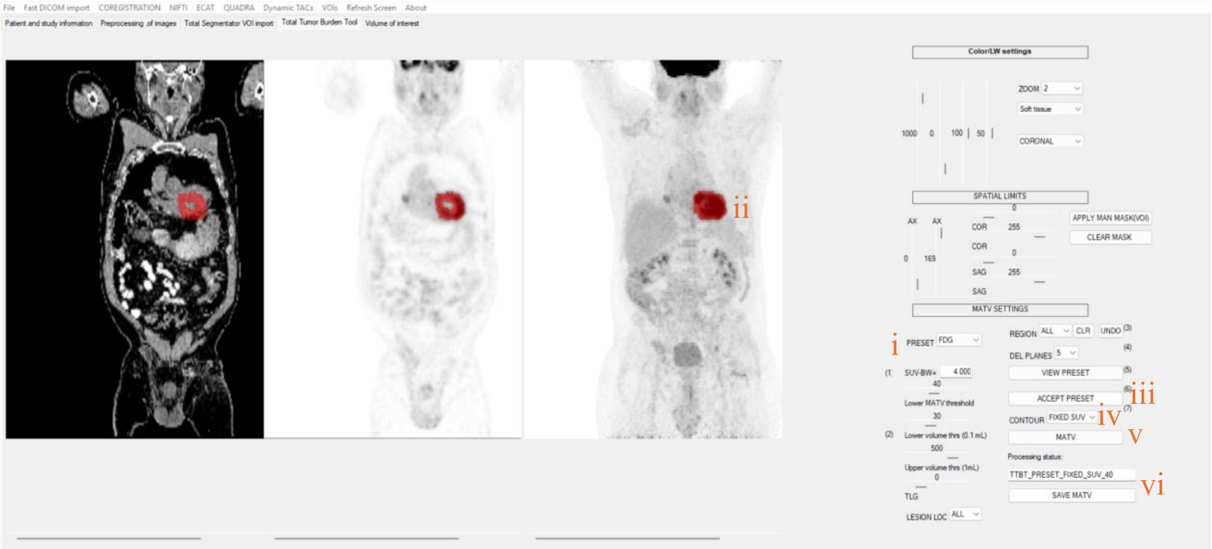
ii) Not all high-glucose-consuming parts are tumors. These parts, such as the head, heart, bladder, or intestines need to be unselected by clicking on them with the right mouse button.

iii) Press the button ‘ACCEPT PRESET’.

iv) Choose as ‘CONTOUR’ the option ‘Fixed SUV’

v) Press the button ‘MATV’

vi) Save the MATV by pressing the button ‘SAVE MATV’. It is recommended to keep the name that ACCURATE suggests. This will likely be ‘TTBT_PRESET_FIXED_SUV_40’



Step 11B: To do the final segmentation for SUV4, there are several steps, which are indicated with corresponding numbers on the figure below:

i) Go to the tab ‘Volume of interest’

ii) Choose a view that suits for you. There are four options: AX (axial), COR (coronal), SAG (sagittal), or MIP (Maximum Intensity Projection).

iii) Press the button ‘SUV40’

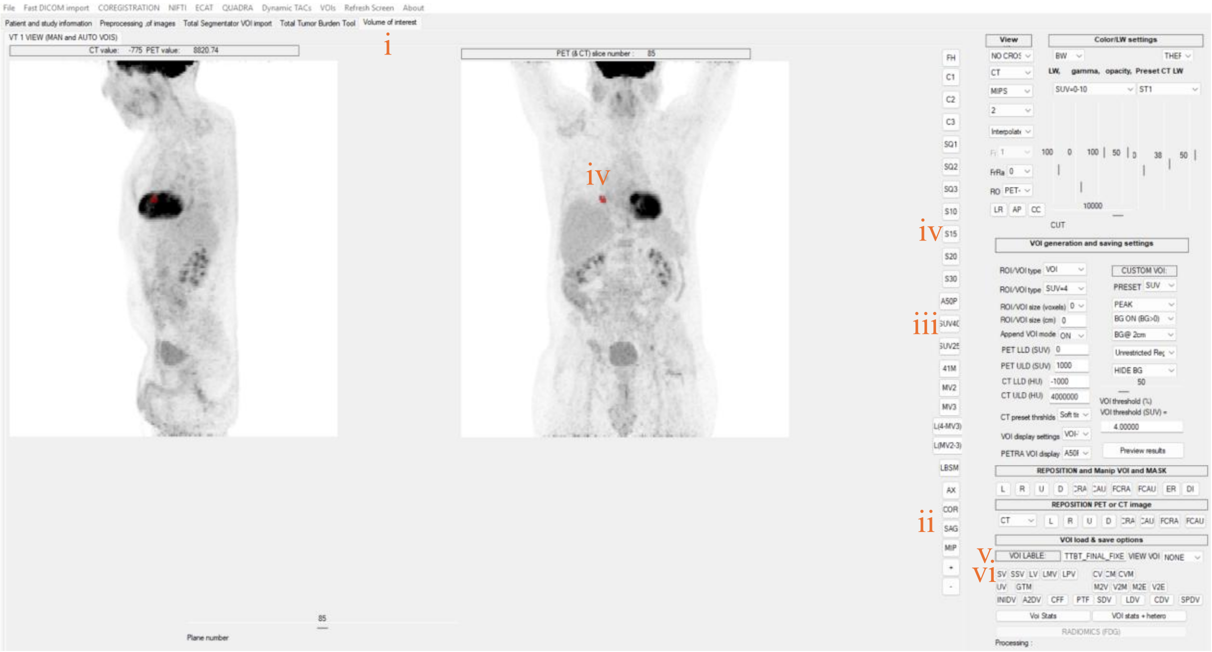
iv) Click, with the right mouse button, on the image where the tumor is located so that it will be marked red on the image.

<u>Note</u>: It is possible that you need to use different views or that you need to zoom.

<u>Note 2</u>: If the tumor will not turn red, choose the button ‘S15’. This will put a sphere where you click on the image.

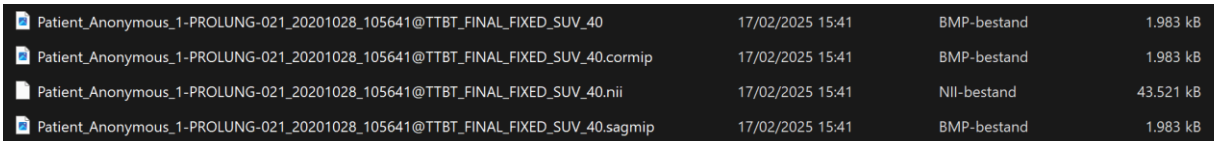
v) Change in the ‘VOI LABLE’, the word ‘PRESET’ to ‘FINAL’

vi) Press the button ‘SV’



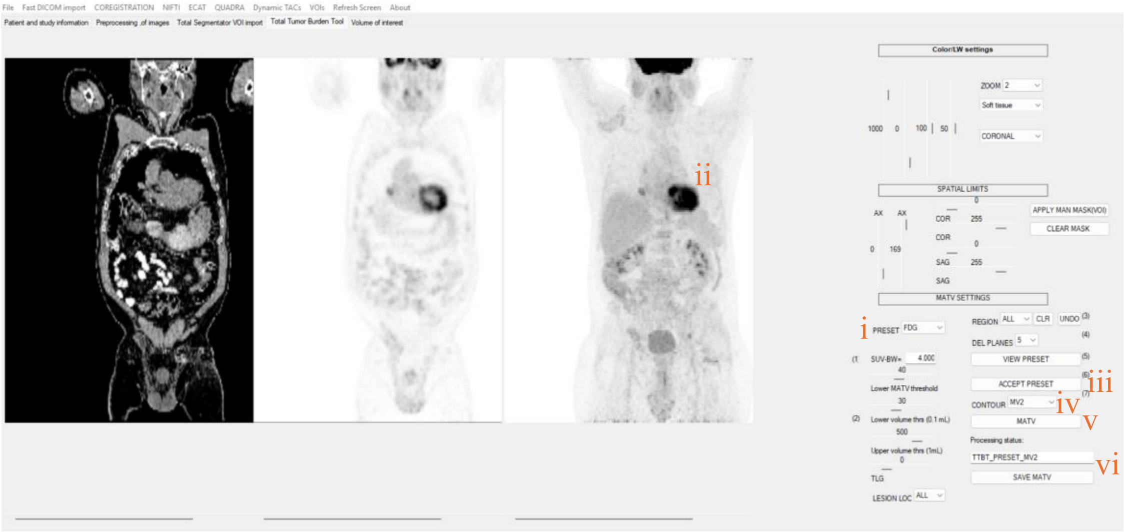
File Fast DICOM import COREGISTRATION NIFTI ECAT QUADRA Dynamic TACs VOIs Refresh Screen About
Patient and study information Preprocessing of images Total Segmentator VOI import Total Tumor Burden Tool Volume of interest
i
iv
iv
iii
ii
v
vi
S15
SUV40
AX
COR
SAG
MIP
VOI generation and saving settings
VOI load & save options
VOI LABLE
TTBT_FINAL_FIXE
SV

After the segmentation with SUV4, you should at least have the four files that are shown in the figure below.

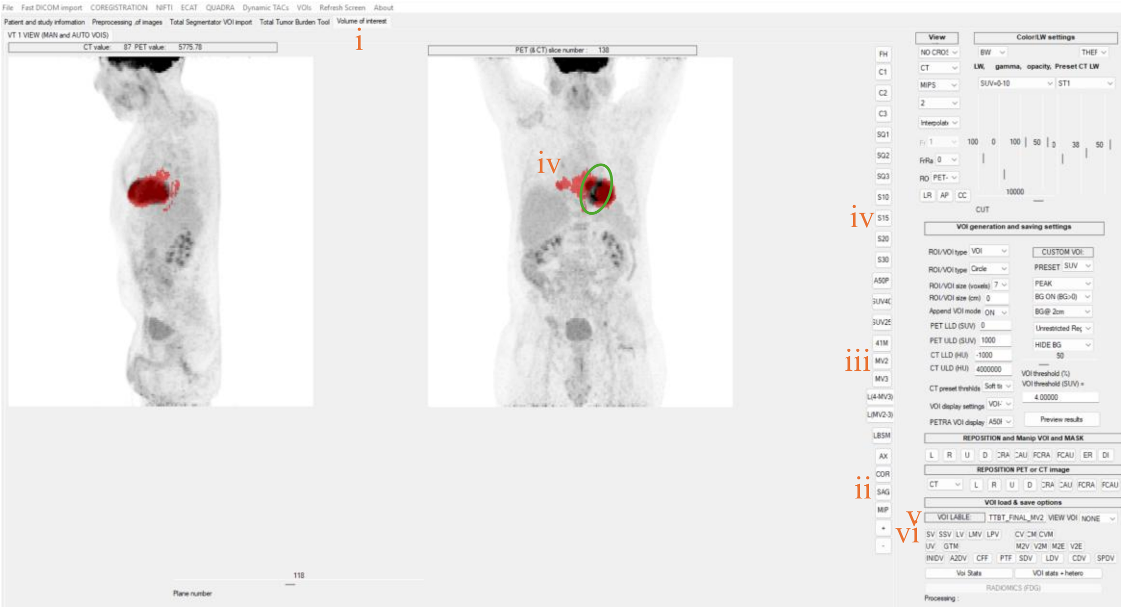


Patient_Anonymous_1-PROLUNG-021_20201028_105641@TTBT_FINAL_FIXED_SUV_40	17/02/2025 15:41	BMP-bestand	1.983 kB
Patient_Anonymous_1-PROLUNG-021_20201028_105641@TTBT_FINAL_FIXED_SUV_40.cormip	17/02/2025 15:41	BMP-bestand	1.983 kB
Patient_Anonymous_1-PROLUNG-021_20201028_105641@TTBT_FINAL_FIXED_SUV_40.nii	17/02/2025 15:41	NII-bestand	43.521 kB
Patient_Anonymous_1-PROLUNG-021_20201028_105641@TTBT_FINAL_FIXED_SUV_40.sagmip	17/02/2025 15:41	BMP-bestand	1.983 kB

Step 12A: To do the presegmentation for MV2, almost all steps are identical as those for SUV4. The only step that differs is step iv) where 'MV2' should be chosen as 'CONTOUR'. The steps are indicated with corresponding numbers on the figure below.

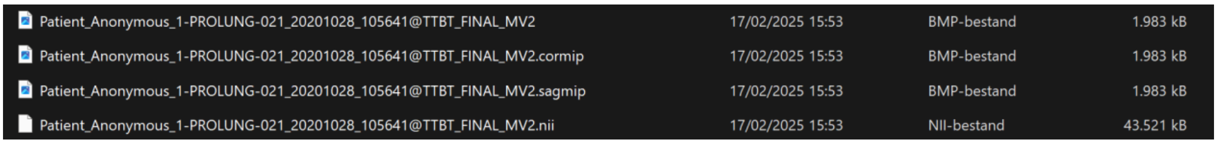


Step 12B: To do the final segmentation for MV2, almost all steps are identical as those for SUV4. The only step that differs is step iii) where the button 'MV2' should be pressed instead of the button SUV40. The steps are indicated with corresponding numbers on the figure below.


File Fast DICOM import COREGISTRATION NIFTI ECAT QUADRA Dynamic TACs VOIs Refresh Screen About
Patient and study information Preprocessing of images Total Segmentator VOI import Total Tumor Burden Tool Volume of interest
VT 1 VIEW (MAN and AUTO VOIS)
CT value: 87 PET value: 5775.78
i
PET (& CT) slice number: 138
iv
iv
FH
C1
C2
C3
SQ1
SQ2
SQ3
S10
S15
S20
S30
A50P
SUV40
SUV25
41M
iii
MV2
MV3
ii
AX
COR
SAG
MIP
v
vi
View
Color/LW settings
VOI generation and saving settings
REPOSITION and Manip VOI and MASK
REPOSITION PET or CT image
VOI load & save options
VOI LABLE TTBT_FINAL_MV2
Voi Stats
VOI stats + hetero
Plane number

Note: It is possible that an area larger than the tumor will be indicated when you right-click with your mouse. To remove a part of the indicated area, you can press the buttons 'C1', 'C2' or 'C3', with C meaning circle. Then hold shift while hovering over the part you want to remove. This action can be performed in all views accept the MIP view, meaning that the removal needs to be done for every slice. An example of a removed area is indicated with the green circle in the figure above.

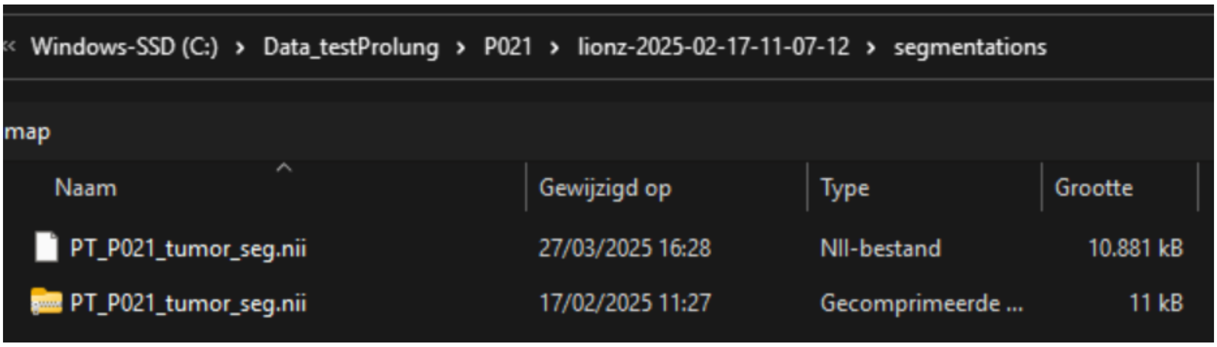
After the segmentation with MV2, you should at least have the four files that are shown in the figure below.



Patient_Anonymous_1-PROLUNG-021_20201028_105641@TTBT_FINAL_MV2	17/02/2025 15:53	BMP-bestand	1.983 kB
Patient_Anonymous_1-PROLUNG-021_20201028_105641@TTBT_FINAL_MV2.cormip	17/02/2025 15:53	BMP-bestand	1.983 kB
Patient_Anonymous_1-PROLUNG-021_20201028_105641@TTBT_FINAL_MV2.sagmip	17/02/2025 15:53	BMP-bestand	1.983 kB
Patient_Anonymous_1-PROLUNG-021_20201028_105641@TTBT_FINAL_MV2.nii	17/02/2025 15:53	NII-bestand	43.521 kB

Step 13: To do the segmentation for LIONZ, the presegmentation for LIONZ is also the final segmentation. Meaning that everything can be done in the tab ‘Total Tumor Burden Tool’. The steps needed for segmentation for LIONZ are indicated with corresponding numbers on the figure down below.

i) Set the preset to ‘load voi’. This will open a window. When you go to your patient map, you will see that because of step 8 there will be a map added. You follow the directory, as shown in the figure below, until you find a NII and/or Zip file. When selecting this file, the red spots will appear on the images.



« Windows-SSD (C:) › Data_testProlung › P021 › lionz-2025-02-17-11-07-12 › segmentations

map

Naam	Gewijzigd op	Type	Grootte
PT_P021_tumor_seg.nii	27/03/2025 16:28	NII-bestand	10.881 kB
PT_P021_tumor_seg.nii	17/02/2025 11:27	Gecomprimeerde ...	11 kB

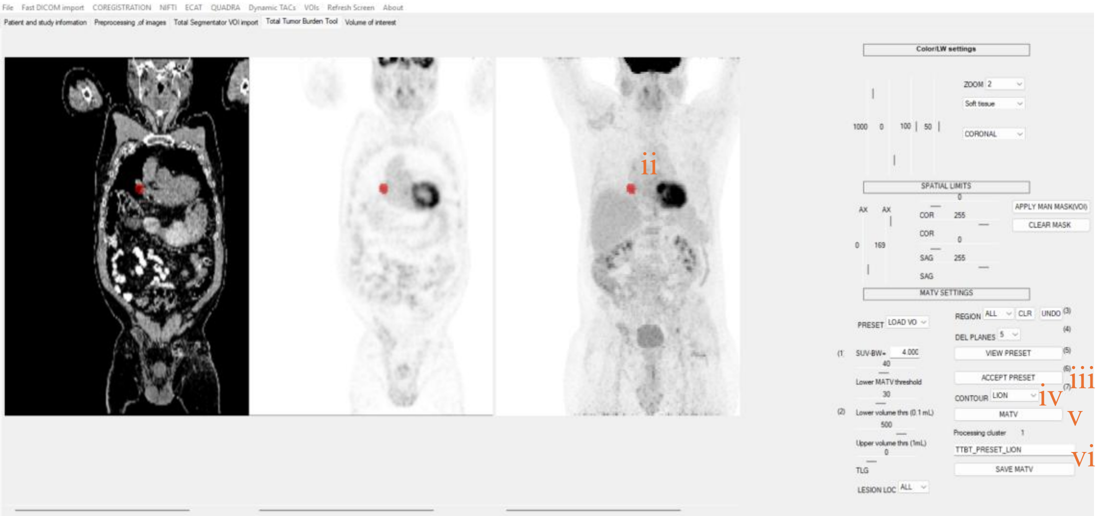
ii) Not all high-glucose-consuming parts are tumors. The non-tumors need to be unselected by clicking on them with the right mouse button.

iii) Press the button ‘ACCEPT PRESET’.

iv) Choose as ‘CONTOUR’ the option ‘LIONZ’

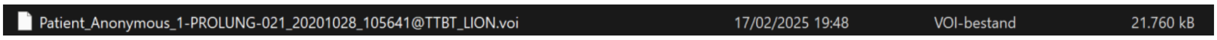
v) Press the button ‘MATV’

vi) Save the MATV by pressing the button ‘SAVE MATV’. It is recommended to keep the name that accurate suggests. This will likely be ‘TTBT_PRESET_LIONZ’. You can also change the name to ‘TTBT_LIONZ’.


File Fast DICOM import COREGISTRATION NIFTI ECAT QUADRA Dynamic TACs VOIs Refresh Screen About
Patient and study information Preprocessing of images Total Segmentator VOI import Total Tumor Burden Tool Volume of interest
ii
Color/LW settings
SPATIAL LIMITS
AX AX COR COR SAG SAG
APPLY MAN MASK(VOI)
CLEAR MASK
MATV SETTINGS
CLR
UNDO
(3)
(4)
VIEW PRESET
(5)
(6)
ACCEPT PRESET
iii
(7)
iv
MATV
v
vi
SAVE MATV
(1
(2)

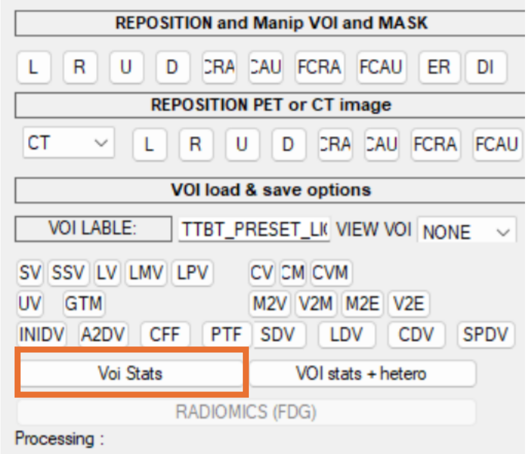
ZOOM 2
Soft tissue
1000 0 100 50
CORONAL
0 169
0 255 0 255
PRESET LOAD VO
REGION ALL
DEL PLANES 5
SUV-BW= 4.000
40
Lower MATV threshold
30
CONTOUR LION
Lower volume thrs (0.1 mL)
500
Upper volume thrs (1mL)
0
Processing cluster 1
TTBT_PRESET_LION
TLG
LESION LOC ALL

After the segmentation with LIONZ, you should at least have the file that is shown in the figure below.



Patient_Anonymous_1-PROLUNG-021_20201028_105641@TTBT_LION.voi 17/02/2025 19:48 VOI-bestand 21.760 kB

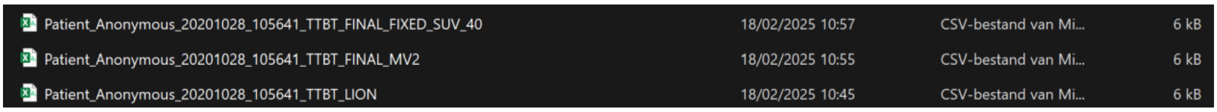
Step 14: When the VOIs of all three segmentation methods are acquired, VOI statistics can be extracted. This is simply done by pressing the ‘Voi Stats’ button and selecting the wanted VOIs of the patient. It is possible to select multiple VOIs at once, so all three VOIs from the different segmentation methods can be processed at the same time.



CT

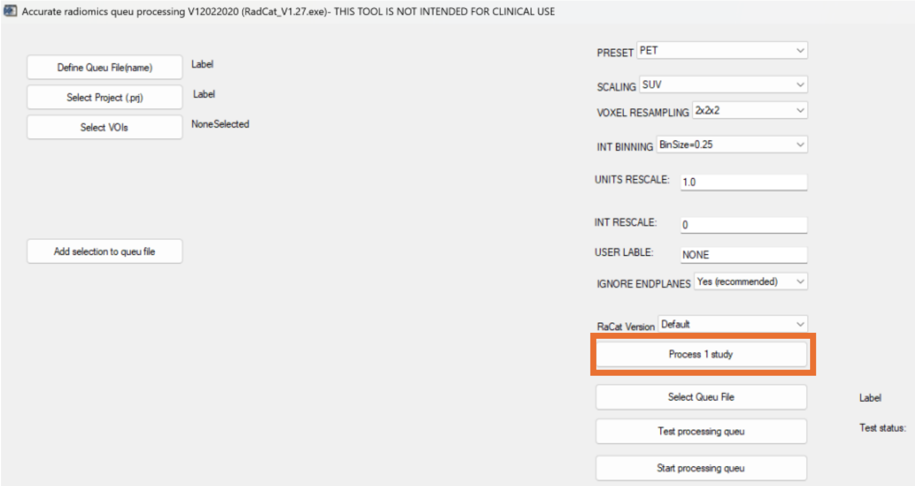
VOI LABLE: TTBT_PRESET_LI(VIEW VOI NONE

Now, there should be three Excel files added to the patient folder, such as shown in the figure below.



Patient_Anonymous_20201028_105641_TTBT_FINAL_FIXED_SUV_40	18/02/2025 10:57	CSV-bestand van Mi...	6 kB
Patient_Anonymous_20201028_105641_TTBT_FINAL_MV2	18/02/2025 10:55	CSV-bestand van Mi...	6 kB
Patient_Anonymous_20201028_105641_TTBT_LION	18/02/2025 10:45	CSV-bestand van Mi...	6 kB

Step 15: Close the ACCURATE tool and open the RADIOMICS tool. The RADIOMICS tool should be located on the C:\ disk of the computer and can be opened by launching ‘AccRadiomicsQueue_v127c_06082020. Then press ‘Click to continue’ and the application should open. When opened, the application should look like the image below.



Accurate radiomics queu processing V12022020 (RadCat_V1.27.exe)- THIS TOOL IS NOT INTENDED FOR CLINICAL USE

PRESET PET

SCALING SUV

VOXEL RESAMPLING 2x2x2

INT BINNING BinSize=0.25

UNITS RESCALE: 1.0

INT RESCALE: 0

USER LABLE: NONE

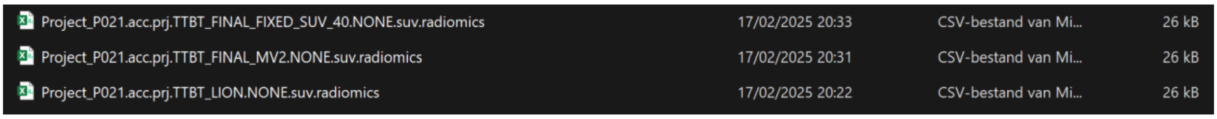
IGNORE ENDPLANES Yes (recommended)

RaCat Version Default

Note: You can either extract radiomic features for multiple patients at a time, via the option to define a que on the left side of the application. Or you can process one patient at a time via the right side of the application.

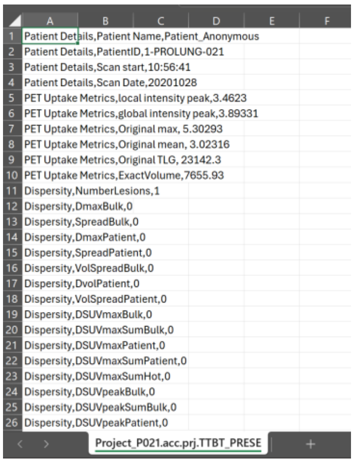
Step 16: To process one patient at a time, click ‘Process 1 study’. Then select the project file and next the different VOI-files you want to process.

Now, there should be three Excel files added to the patient folder, as shown in the figure below.



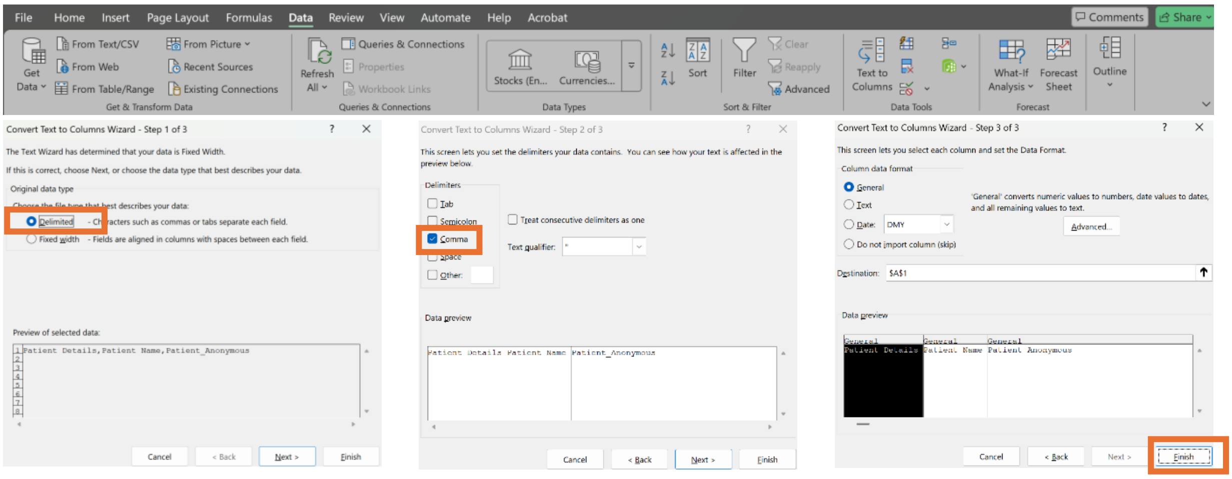
Project_P021.acc.prj.TTBT_FINAL_FIXED_SUV_40.NONE.suv.radiomics	17/02/2025 20:33	CSV-bestand van Mi...	26 kB
Project_P021.acc.prj.TTBT_FINAL_MV2.NONE.suv.radiomics	17/02/2025 20:31	CSV-bestand van Mi...	26 kB
Project_P021.acc.prj.TTBT_LION.NONE.suv.radiomics	17/02/2025 20:22	CSV-bestand van Mi...	26 kB

Step 17: Open the Excel files. The data will all be in one column, as it is actually a CSV file:



	A	B	C	D	E	F
1	Patient Details,Patient Name,Patient_Anonymous					
2	Patient Details,PatientID,1-PROLUNG-021					
3	Patient Details,Scan start,10:56:41					
4	Patient Details,Scan Date,20201028					
5	PET Uptake Metrics,local intensity peak,3.4623					
6	PET Uptake Metrics,global intensity peak,3.89331					
7	PET Uptake Metrics,Original max, 5.30293					
8	PET Uptake Metrics,Original mean, 3.02316					
9	PET Uptake Metrics,Original TLG, 23142.3					
10	PET Uptake Metrics,ExactVolume,7655.93					
11	Dispersity,NumberLesions,1					
12	Dispersity,DmaxBulk,0					
13	Dispersity,SpreadBulk,0					
14	Dispersity,DmaxPatient,0					
15	Dispersity,SpreadPatient,0					
16	Dispersity,VolSpreadBulk,0					
17	Dispersity,DvolPatient,0					
18	Dispersity,VolSpreadPatient,0					
19	Dispersity,DSUVmaxBulk,0					
20	Dispersity,DSUVmaxSumBulk,0					
21	Dispersity,DSUVmaxPatient,0					
22	Dispersity,DSUVmaxSumPatient,0					
23	Dispersity,DSUVmaxSumHot,0					
24	Dispersity,DSUVpeakBulk,0					
25	Dispersity,DSUVpeakSumBulk,0					
26	Dispersity,DSUVpeakPatient,0					

To separate the data in multiple columns, the ‘Text to column’ button under the ‘Data’ tab can be used as shown in the figures below. It is important to note that all cells must be selected.



Convert Text to Columns Wizard - Step 1 of 3

The Text Wizard has determined that your data is Fixed Width.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

● Delimited - Characters such as commas or tabs separate each field.

○ Fixed width - Fields are aligned in columns with spaces between each field.

Preview of selected data:

```
1 Patient Details,Patient Name,Patient_Anonymous
```

Convert Text to Columns Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

□ Tab

□ Semicolon

☑ Comma

□ Space

□ Other:

□ Treat consecutive delimiters as one

Text qualifier: "

Data preview

```
Patient Details Patient Name Patient_Anonymous
```

Convert Text to Columns Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

Column data format

● General

○ Text

○ Date: DMY

○ Do not import column (skip)

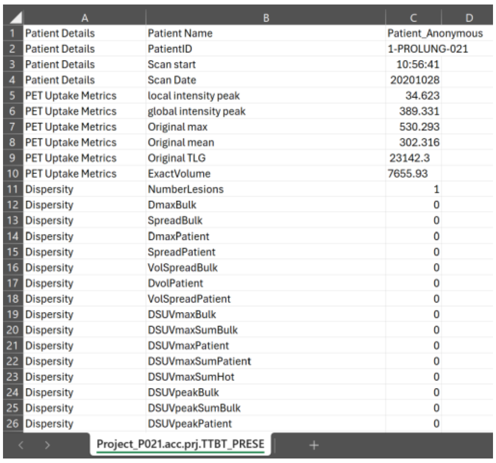
'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

Destination: A1

Data preview

```
General         General       General
Patient Details Patient Name  Patient Anonymous
```

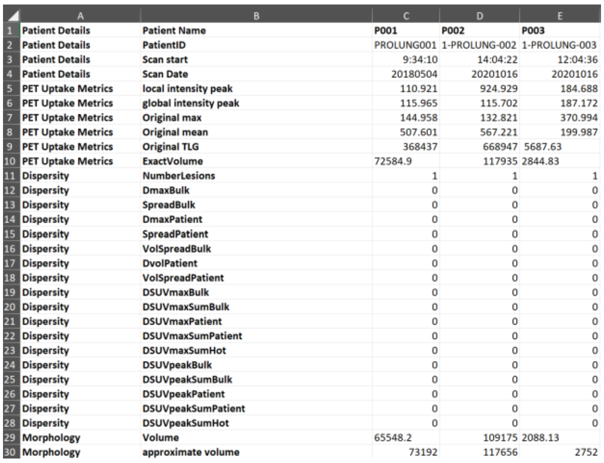
Now the data should be split into three columns.



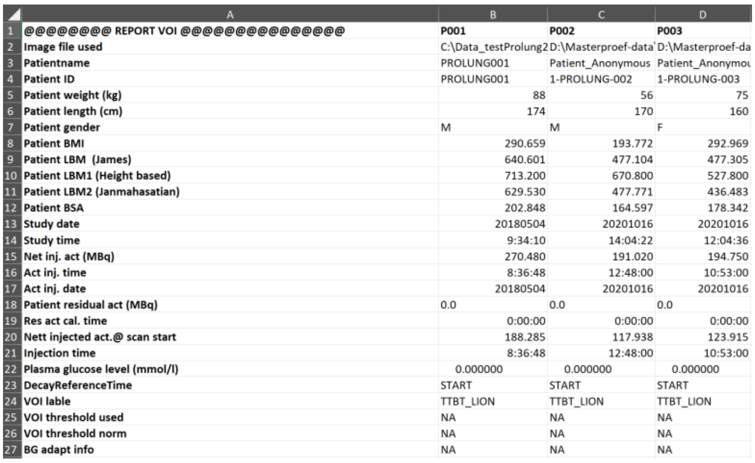
	A	B	C	D
1	Patient Details	Patient Name	Patient_Anonymous	
2	Patient Details	PatientID	1-PROLUNG-021	
3	Patient Details	Scan start	10:56:41	
4	Patient Details	Scan Date	20201028	
5	PET Uptake Metrics	local intensity peak	34.623	
6	PET Uptake Metrics	global intensity peak	389.331	
7	PET Uptake Metrics	Original max	530.293	
8	PET Uptake Metrics	Original mean	302.316	
9	PET Uptake Metrics	Original TLG	23142.3	
10	PET Uptake Metrics	ExactVolume	7655.93	
11	Dispersity	NumberLesions	1	
12	Dispersity	DmaxBulk	0	
13	Dispersity	SpreadBulk	0	
14	Dispersity	DmaxPatient	0	
15	Dispersity	SpreadPatient	0	
16	Dispersity	VolSpreadBulk	0	
17	Dispersity	DvolPatient	0	
18	Dispersity	VolSpreadPatient	0	
19	Dispersity	DSUVmaxBulk	0	
20	Dispersity	DSUVmaxSumBulk	0	
21	Dispersity	DSUVmaxPatient	0	
22	Dispersity	DSUVmaxSumPatient	0	
23	Dispersity	DSUVmaxSumHot	0	
24	Dispersity	DSUVpeakBulk	0	
25	Dispersity	DSUVpeakSumBulk	0	
26	Dispersity	DSUVpeakPatient	0	

This process needs to be repeated for every Excel file of the patient, including the VOI Excel files.

Step 18: Combine per segmentation method and per data type (VOI stats or Radiomics) the data of every patient in one big Excel file. Thus, in total, there must be 6 big Excel files with data from all patients. The files should look like the figures below for respectively Radiomics and VOI stats.



	A	B	C	D	E
1	Patient Details	Patient Name	P001	P002	P003
2	Patient Details	PatientID	PROLUNG001	1-PROLUNG-002	1-PROLUNG-003
3	Patient Details	Scan start	9:34:10	14:04:22	12:04:36
4	Patient Details	Scan Date	20180504	20201016	20201016
5	PET Uptake Metrics	local intensity peak	110.921	924.929	184.688
6	PET Uptake Metrics	global intensity peak	115.965	115.702	187.172
7	PET Uptake Metrics	Original max	144.958	132.821	370.994
8	PET Uptake Metrics	Original mean	507.601	567.221	199.987
9	PET Uptake Metrics	Original TLG	368437	668947	5687.63
10	PET Uptake Metrics	ExactVolume	72584.9	117935	2844.83
11	Dispersity	NumberLesions	1	1	1
12	Dispersity	DmaxBulk	0	0	0
13	Dispersity	SpreadBulk	0	0	0
14	Dispersity	DmaxPatient	0	0	0
15	Dispersity	SpreadPatient	0	0	0
16	Dispersity	VolSpreadBulk	0	0	0
17	Dispersity	DvolPatient	0	0	0
18	Dispersity	VolSpreadPatient	0	0	0
19	Dispersity	DSUVmaxBulk	0	0	0
20	Dispersity	DSUVmaxSumBulk	0	0	0
21	Dispersity	DSUVmaxPatient	0	0	0
22	Dispersity	DSUVmaxSumPatient	0	0	0
23	Dispersity	DSUVmaxSumHot	0	0	0
24	Dispersity	DSUVpeakBulk	0	0	0
25	Dispersity	DSUVpeakSumBulk	0	0	0
26	Dispersity	DSUVpeakPatient	0	0	0
27	Dispersity	DSUVpeakSumPatient	0	0	0
28	Dispersity	DSUVpeakSumHot	0	0	0
29	Morphology	Volume	65548.2	109175	2088.13
30	Morphology	approximate volume	73192	117656	2752



	A	B	C	D
1	@@@@@@@@ REPORT VOI @@@@@@@@@@@@@@@	P001	P002	P003
2	Image file used	C:\Data_testProlung2	D:\Masterproef-data'	D:\Masterproef-da
3	Patientname	PROLUNG001	Patient_Anonymous	Patient_Anonymou
4	Patient ID	PROLUNG001	1-PROLUNG-002	1-PROLUNG-003
5	Patient weight (kg)	88	56	75
6	Patient length (cm)	174	170	160
7	Patient gender	M	M	F
8	Patient BMI	290.659	193.772	292.969
9	Patient LBM (James)	640.601	477.104	477.305
10	Patient LBM1 (Height based)	713.200	670.800	527.800
11	Patient LBM2 (Janmahasatian)	629.530	477.771	436.483
12	Patient BSA	202.848	164.597	178.342
13	Study date	20180504	20201016	20201016
14	Study time	9:34:10	14:04:22	12:04:36
15	Net inj. act (MBq)	270.480	191.020	194.750
16	Act inj. time	8:36:48	12:48:00	10:53:00
17	Act inj. date	20180504	20201016	20201016
18	Patient residual act (MBq)	0.0	0.0	0.0
19	Res act cal. time	0:00:00	0:00:00	0:00:00
20	Nett injected act.@ scan start	188.285	117.938	123.915
21	Injection time	8:36:48	12:48:00	10:53:00
22	Plasma glucose level (mmol/l)	0.000000	0.000000	0.000000
23	DecayReferenceTime	START	START	START
24	VOI lable	TTBT_LION	TTBT_LION	TTBT_LION
25	VOI threshold used	NA	NA	NA
26	VOI threshold norm	NA	NA	NA
27	BG adapt info	NA	NA	NA

Appendix B: Results for overall survival of 24 months

SUV4 segmentation

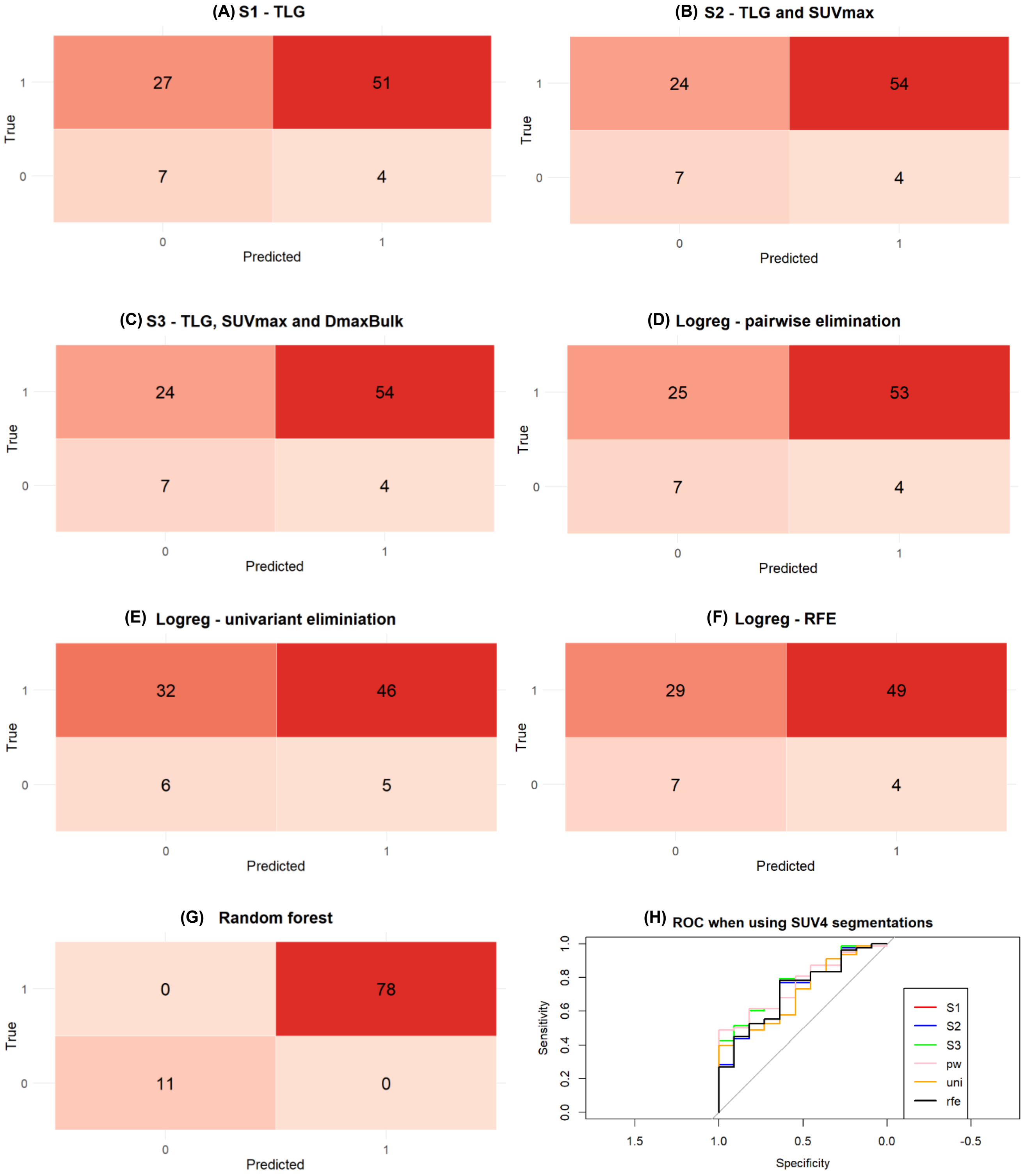


Figure S1: Results with SUV4 segmentation for the cross-validated prognostic models using OS24 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the cross-validated prognostic models using SUV4 segmentation. Logreg: logistic regression.

MV2 segmentation

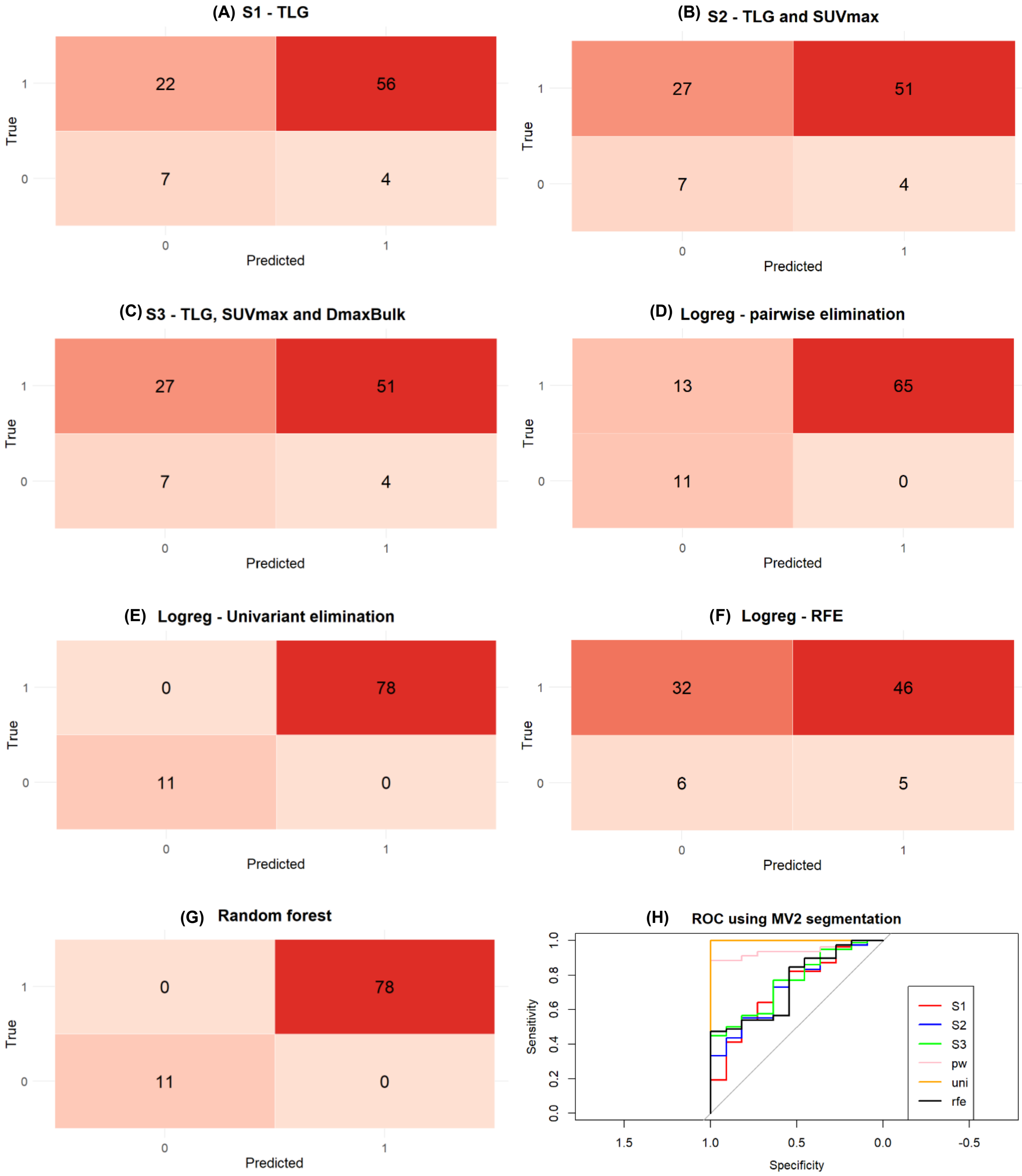


Figure S2: Results with MV2 segmentation for the cross-validated prognostic models using OS24 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the cross-validated prognostic models using MV2 segmentation. Logreg: logistic regression.

MV2 segmentation external validation

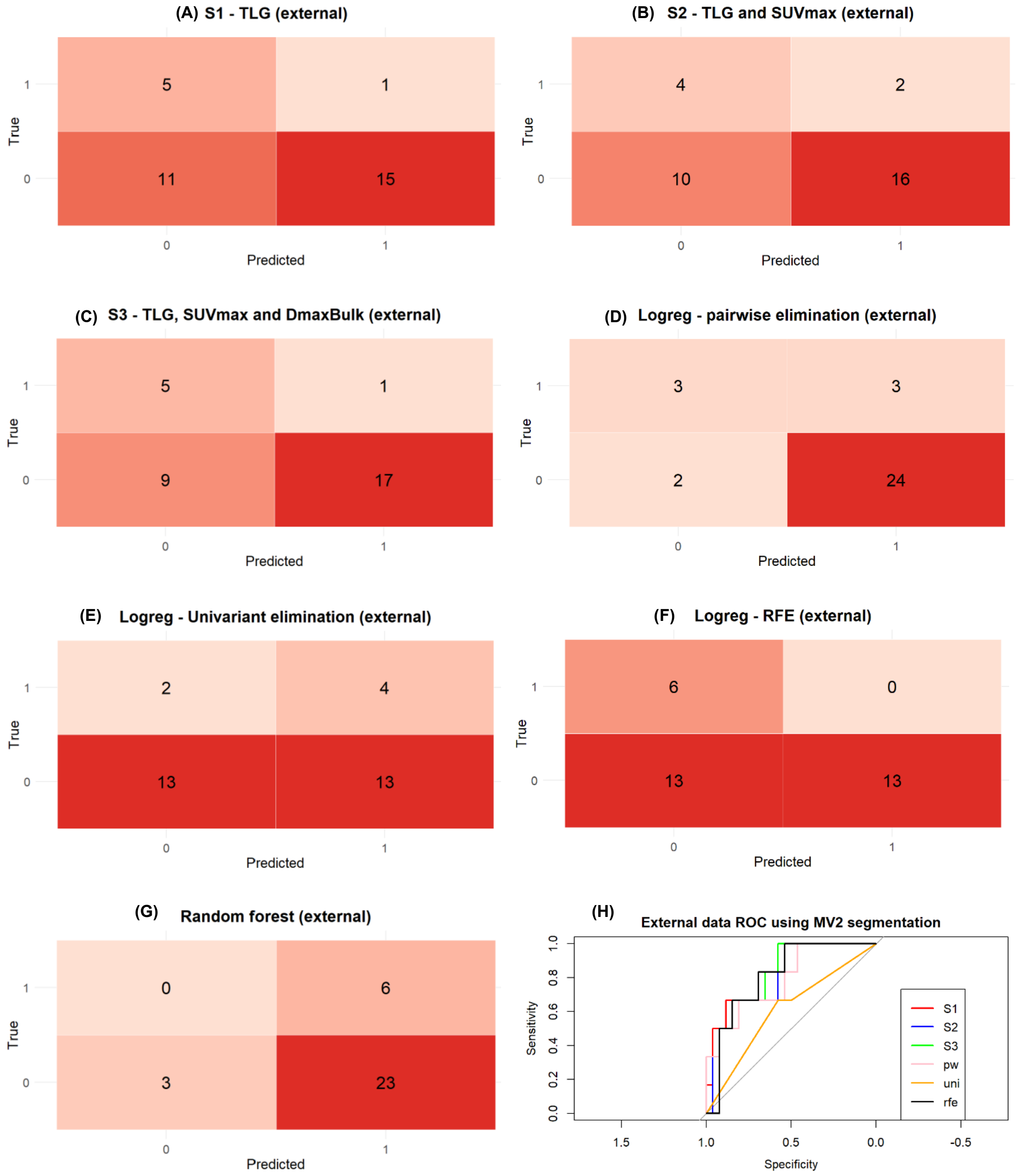


Figure S3: Results with MV2 segmentation for the externally validated prognostic models using OS24 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the externally validated prognostic models using MV2 segmentation. Logreg: logistic regression.

LIONZ segmentation

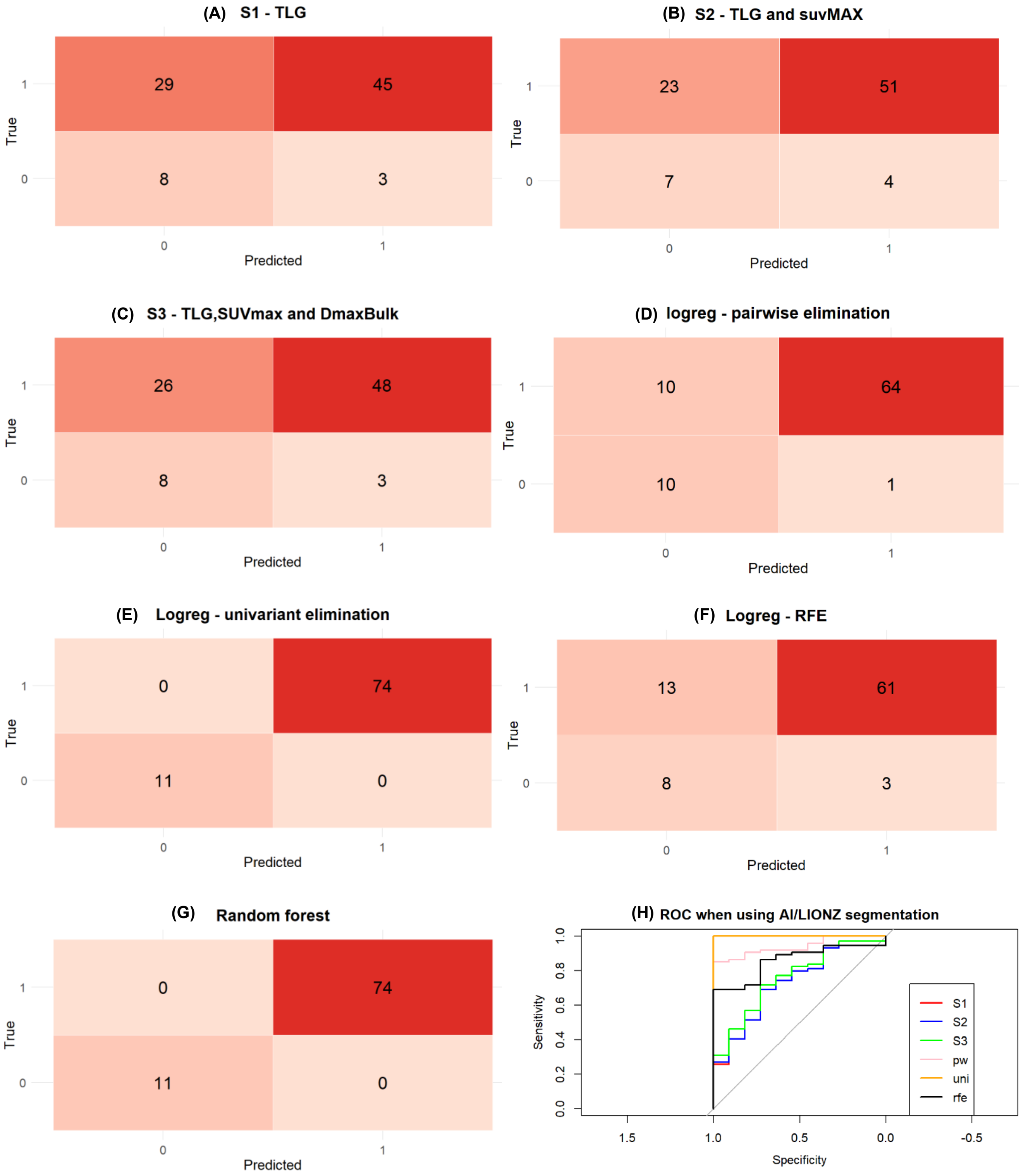


Figure S4: Results with LIONZ segmentation for the cross-validated prognostic models using OS24 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the cross-validated prognostic models using LIONZ segmentation. Logreg: logistic regression.

LIONZ segmentation external validation

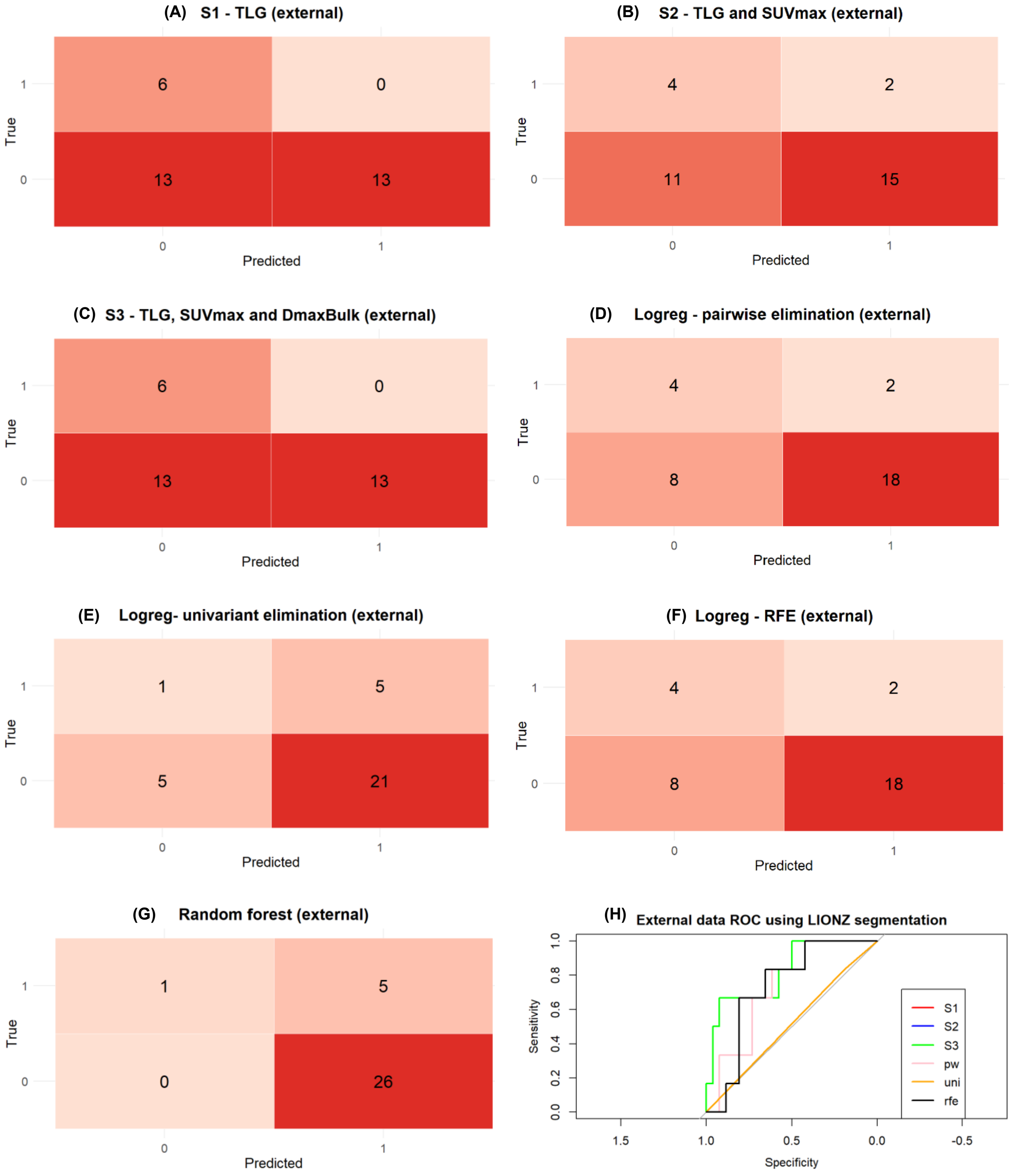


Figure S5: Results with LIONZ segmentation for the externally validated prognostic models using OS24 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the externally validated prognostic models using LIONZ segmentation. Logreg: logistic regression.

General information

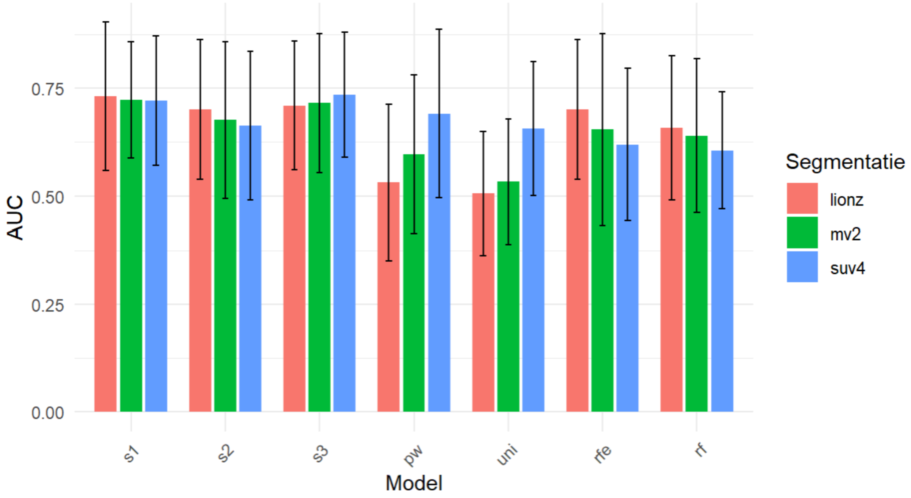


Figure S6: Histogram comparing the AUC values of all cross-validated models using LIONZ, MV2, and SUV4 segmentations with an overall survival of 24 months as prognostic outcome. Orange bars represent models using LIONZ segmentations, green bars represent models using MV2 segmentations, and blue bars represent models using SUV4 segmentations.

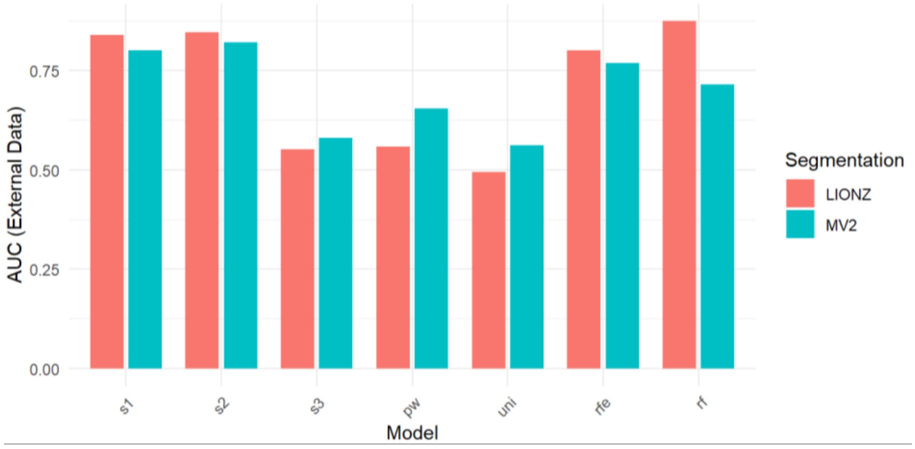


Figure S7: Histogram comparing the AUC values of all externally validated models obtained using LIONZ and MV2 segmentations with an overall survival of 24 months as prognostic outcome. Orange bars represent models using LIONZ segmentations, blue bars represent models using MV2 segmentation.

Table S1: Cross-validated AUC values with standard deviation and externally validated AUC values of prognostic models using SUV4, MV2, and LIONZ as segmentation methods. All models use an overall survival of 24 months as prognostic outcome. AUC sd: area under the curve standard deviation, AUC ext: AUC of external validation.

OS24	s1	s2	s3	pw	uni	rfe	rf
	SUV4						
AUC	0.72	0.66	0.74	0.69	0.66	0.62	0.59
AUC sd	0.15	0.17	0.14	0.20	0.16	0.18	0.14
	MV2						
AUC	0.72	0.68	0.72	0.60	0.53	0.65	0.62
AUC sd	0.14	0.18	0.16	0.18	0.15	0.22	0.19
AUC ext	0.82	0.81	0.81	0.89	0.61	0.81	0.62
	LIONZ						
AUC	0.73	0.70	0.71	0.53	0.51	0.70	0.66
AUC sd	0.17	0.16	0.15	0.18	0.14	0.16	0.17
AUC ext	0.82	0.82	0.82	0.72	0.51	0.73	0.81

DeLong test conversion matrices

SUV4

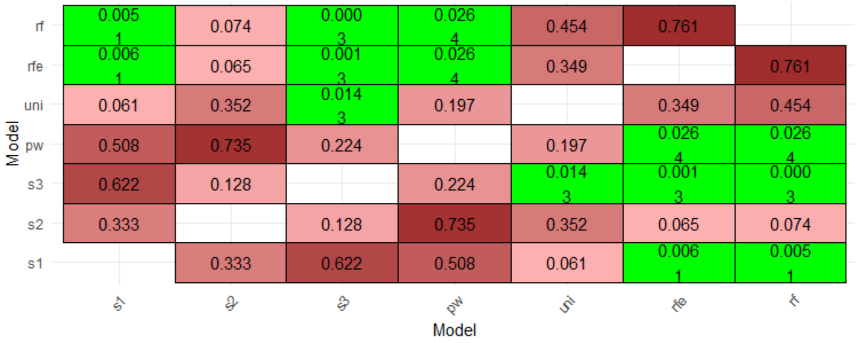


Figure S8: Comparison matrix of the DeLong test with SUV4 segmentation for the cross-validation of the training cohort, with OS24 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

MV2

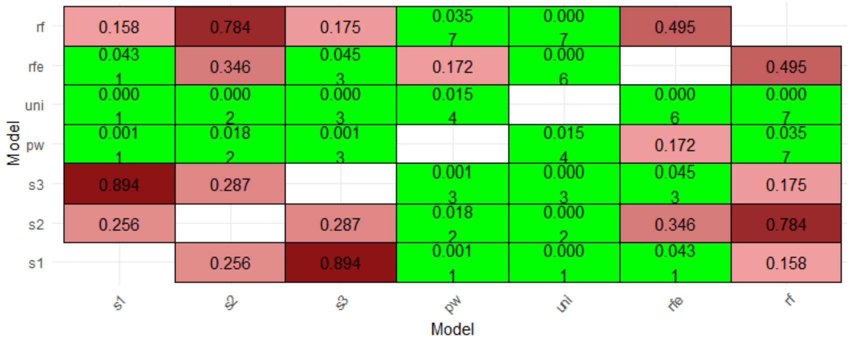


Figure S9: Comparison matrix of the DeLong test with MV2 segmentation for the cross-validation of the training cohort, with OS24 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

MV2 external validation

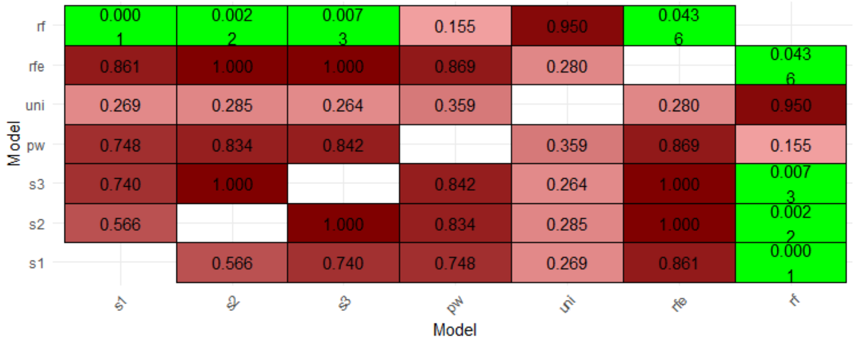
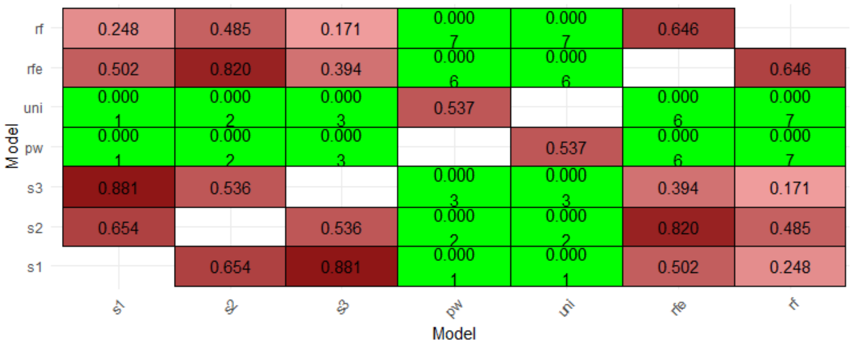


Figure S10: Comparison matrix of the DeLong test with MV2 segmentation for the external validation with the external cohort, with OS24 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

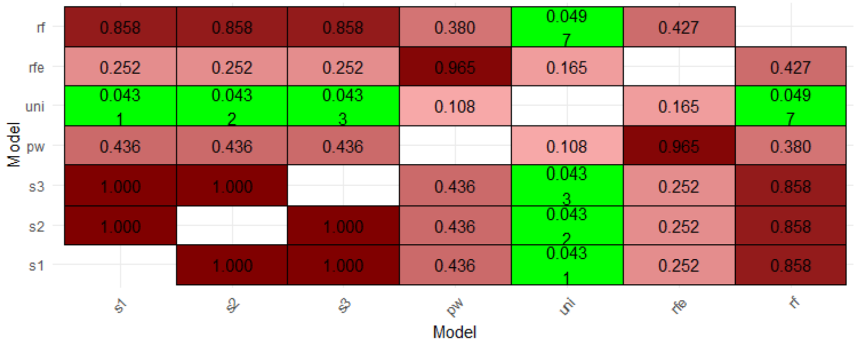
LIONZ



	s1	s2	s3	pw	uni	rfe	rf
rf	0.248	0.485	0.171	0.000 7	0.000 7	0.646	
rfe	0.502	0.820	0.394	0.000 6	0.000 6		0.646
uni	0.000 1	0.000 2	0.000 3	0.537		0.000 6	0.000 7
pw	0.000 1	0.000 2	0.000 3		0.537	0.000 6	0.000 7
s3	0.881	0.536		0.000 3	0.000 3	0.394	0.171
s2	0.654		0.536	0.000 2	0.000 2	0.820	0.485
s1		0.654	0.881	0.000 1	0.000 1	0.502	0.248

Figure S11: Comparison matrix of the DeLong test with LIONZ segmentation for the cross-validation of the training cohort, with OS24 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

LIONZ external validation



	s1	s2	s3	pw	uni	rfe	rf
rf	0.858	0.858	0.858	0.380	0.049 7	0.427	
rfe	0.252	0.252	0.252	0.965	0.165		0.427
uni	0.043 1	0.043 2	0.043 3	0.108		0.165	0.049 7
pw	0.436	0.436	0.436		0.108	0.965	0.380
s3	1.000	1.000		0.436	0.043 3	0.252	0.858
s2	1.000		1.000	0.436	0.043 2	0.252	0.858
s1		1.000	1.000	0.436	0.043 1	0.252	0.858

Figure S12: Comparison matrix of the DeLong test with LIONZ segmentation for the external validation with the validation cohort, with OS24 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

Appendix C: Results for overall survival of 36 months

SUV4 segmentation

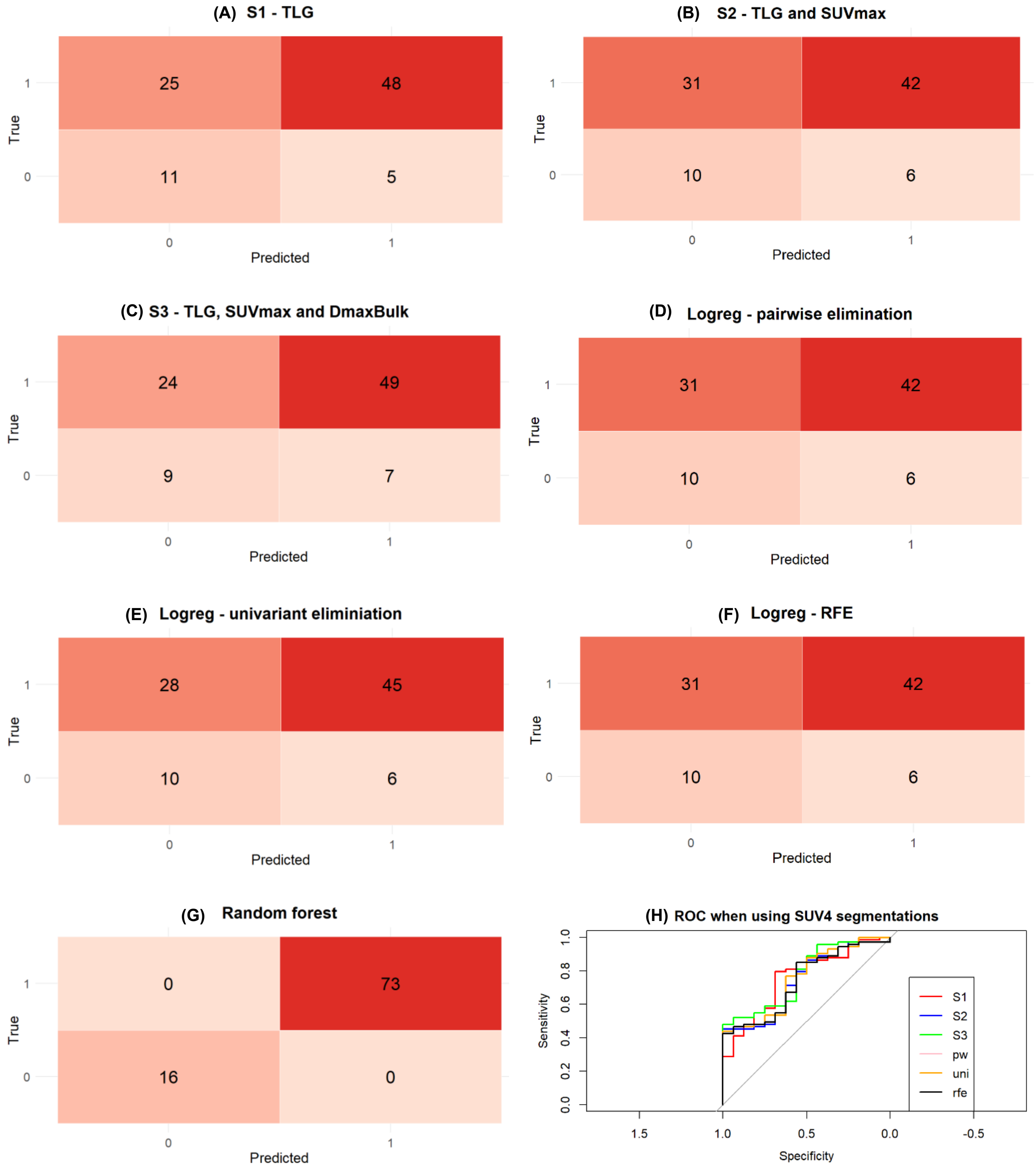


Figure S13: Results with SUV4 segmentation for the cross-validated prognostic models using OS36 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the cross-validated prognostic models using SUV4 segmentation. Logreg: logistic regression.

MV2 segmentation

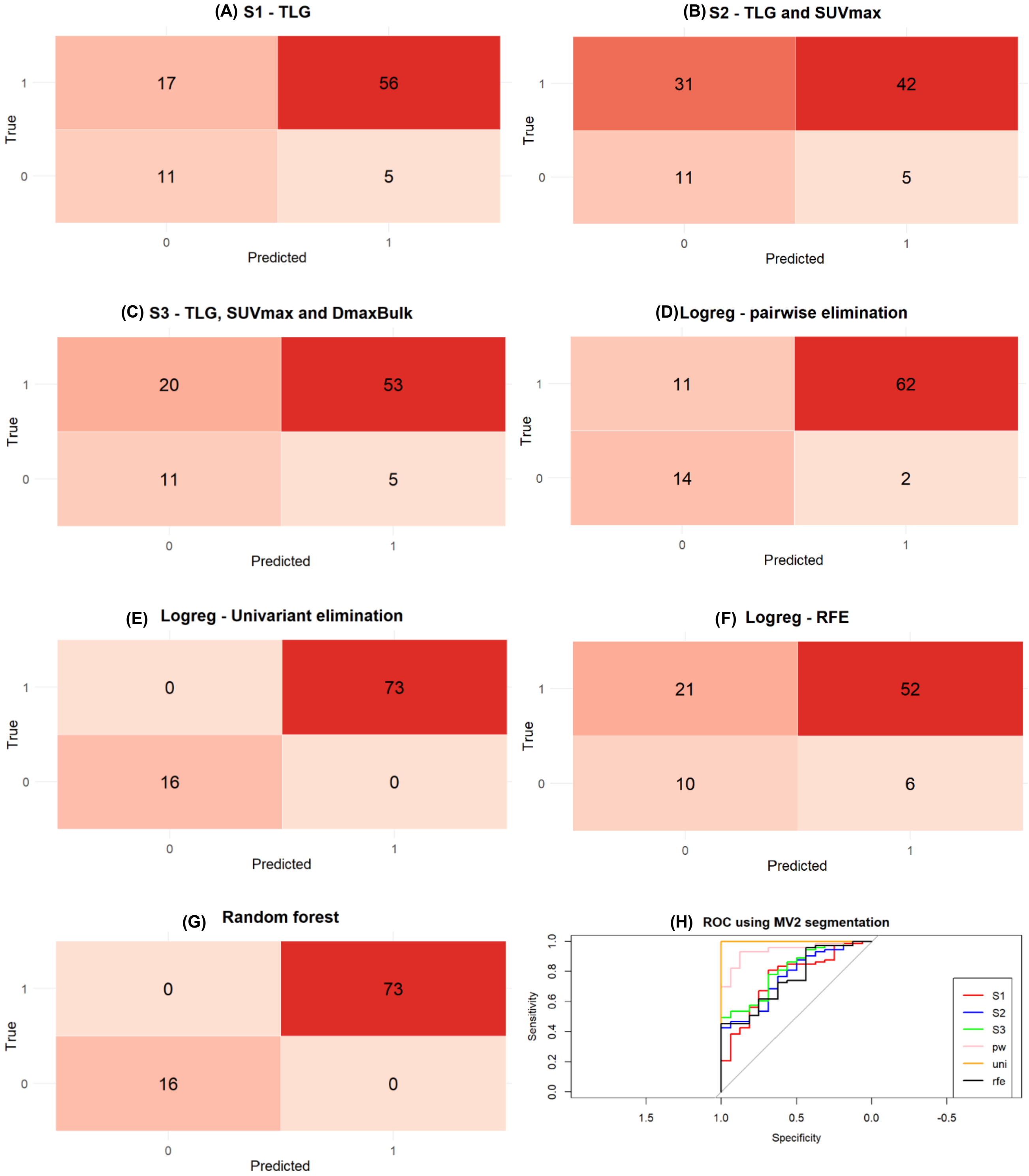


Figure S14: Results with MV2 segmentation for the cross-validated prognostic models using OS36 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the cross-validated prognostic models using MV2 segmentation. Logreg: logistic regression.

MV2 segmentation external validation

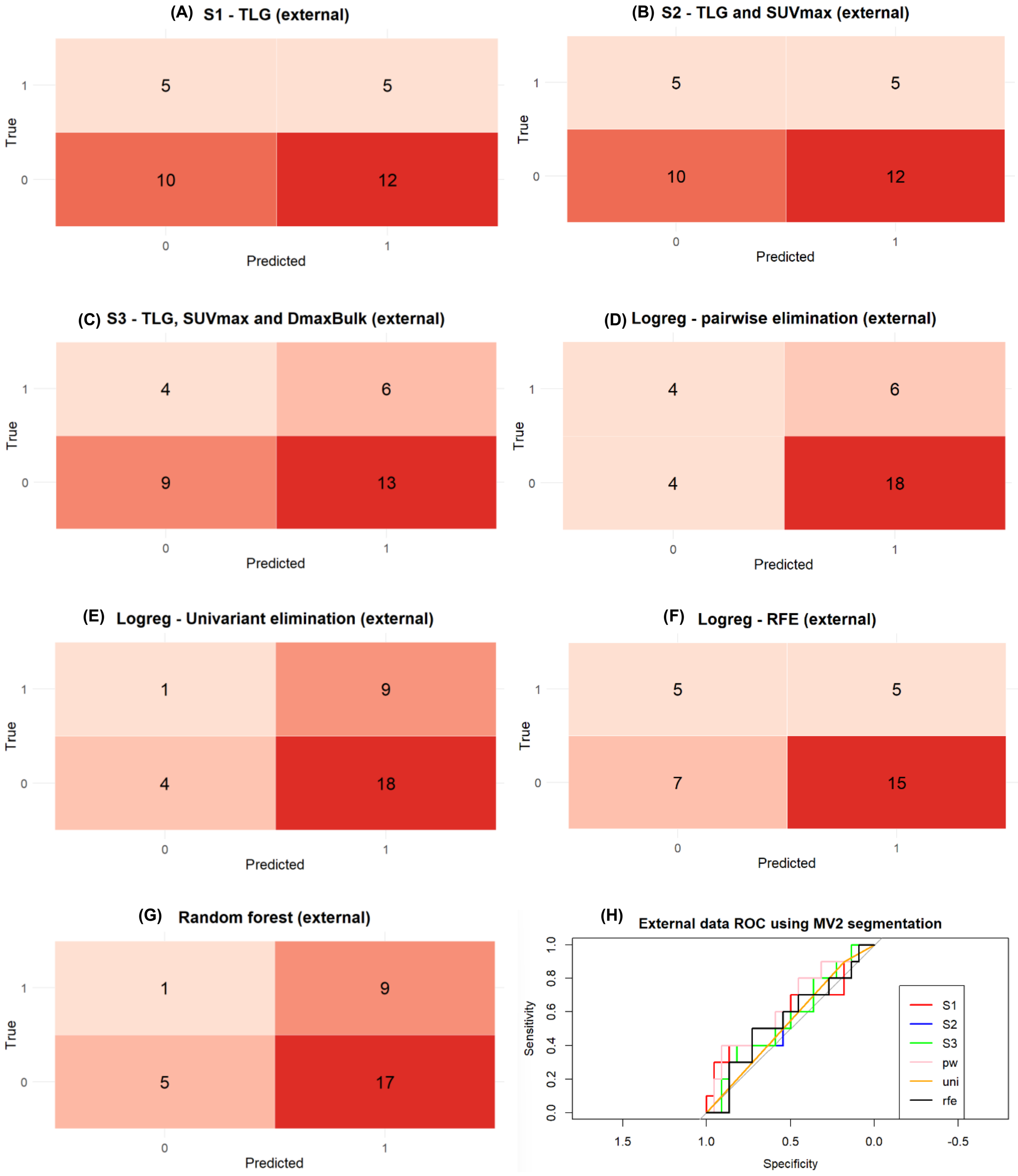


Figure S15: Results with MV2 segmentation for the externally validated prognostic models using OS36 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the externally validated prognostic models using MV2 segmentation. Logreg: logistic regression.

LIONZ segmentation

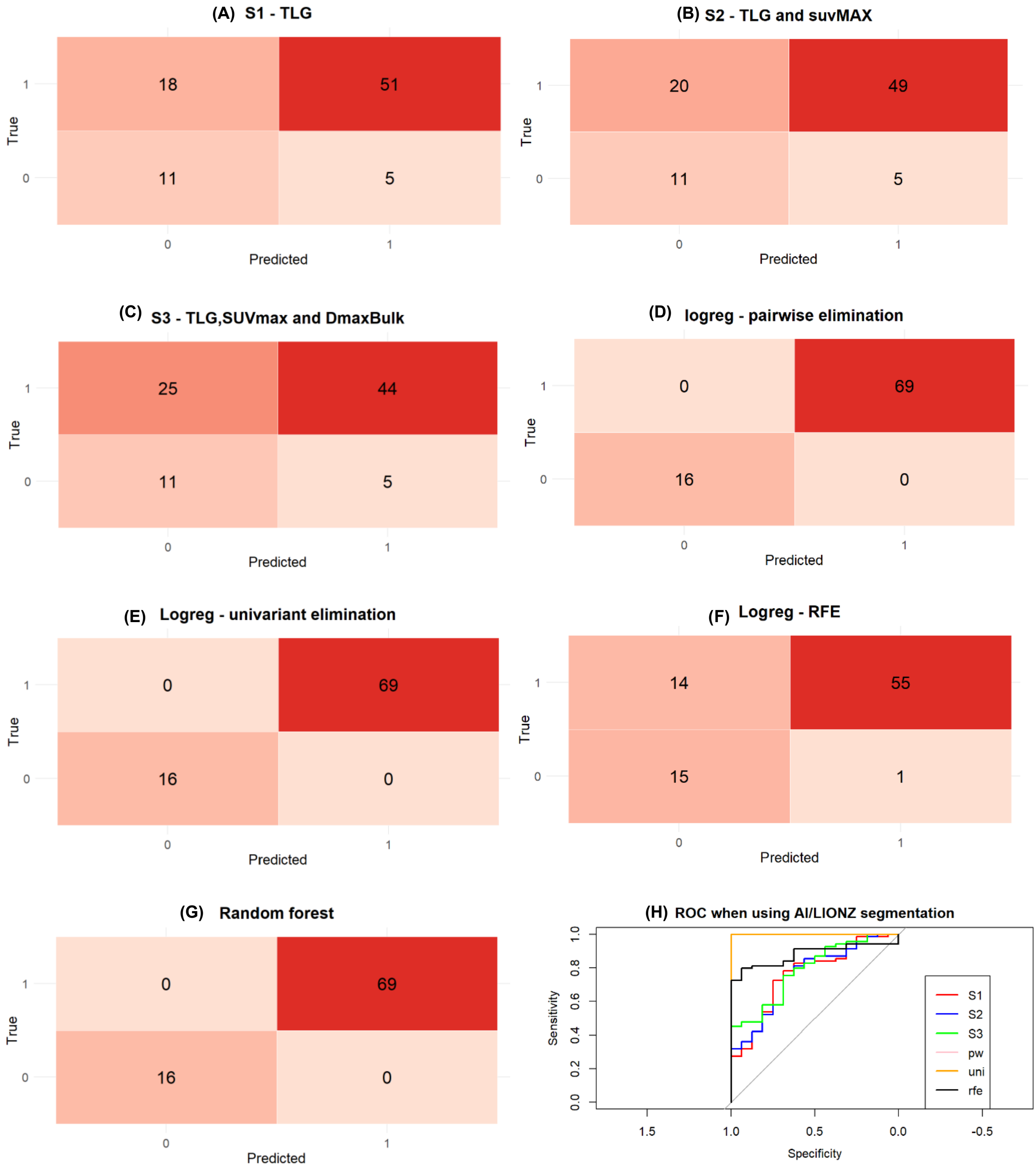


Figure S16: Results with LIONZ segmentation for the cross-validated prognostic models using OS36 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the cross-validated prognostic models using LIONZ segmentation. Logreg: logistic regression.

LIONZ segmentation external validation

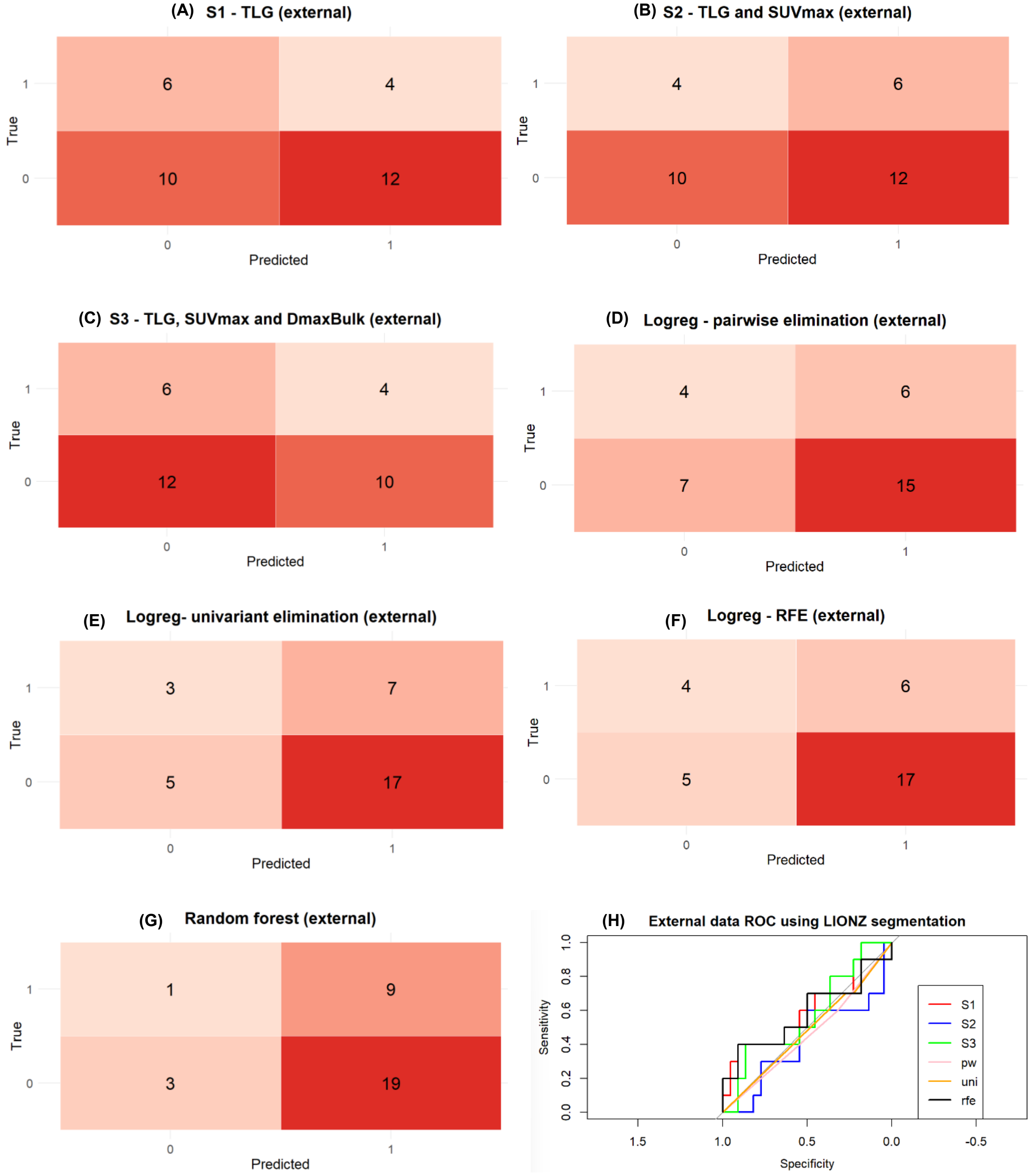


Figure S17: Results with LIONZ segmentation for the externally validated prognostic models using OS36 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the externally validated prognostic models using LIONZ segmentation. Logreg: logistic regression.

General graphs

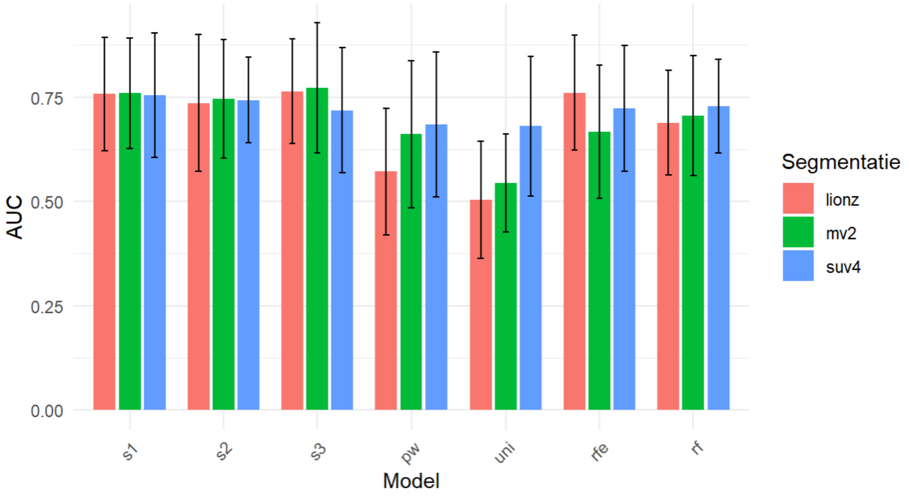


Figure S18: Histogram comparing the AUC values of all cross-validated models using LIONZ, MV2, and SUV4 segmentations with an overall survival of 36 months as prognostic outcome. Orange bars represent models using LIONZ segmentations, green bars represent models using MV2 segmentations, and blue bars represent models using SUV4 segmentations.

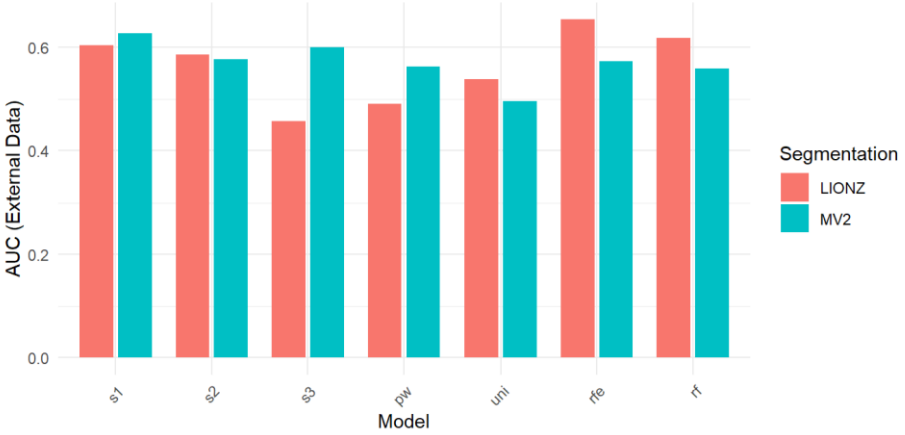


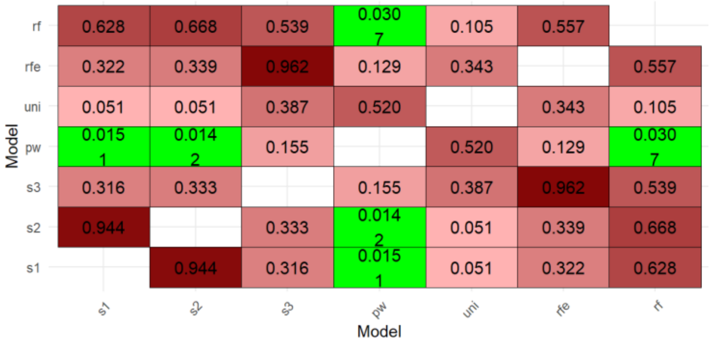
Figure S19: Histogram comparing the AUC values of all externally validated models obtained using LIONZ and MV2 segmentations with an overall survival of 36 months as prognostic outcome. Orange bars represent models using LIONZ segmentations, blue bars represent models using MV2 segmentation.

Table S2: Cross-validated AUC values with standard deviation and externally validated AUC values of prognostic models using SUV4, MV2, and LIONZ as segmentation methods. All models use an overall survival of 36 months as prognostic outcome. AUC sd: area under the curve standard deviation, AUC ext: AUC of external validation.

OS36	s1	s2	s3	pw	uni	rfe	rf
	SUV4						
AUC	0.75	0.74	0.72	0.68	0.68	0.72	0.72
AUCsd	0.15	0.10	0.15	0.17	0.17	0.15	0.12
	MV2						
AUC	0.76	0.75	0.77	0.66	0.54	0.67	0.70
AUC sd	0.13	0.14	0.16	0.18	0.12	0.16	0.13
AUC ext	0.59	0.57	0.57	0.64	0.54	0.55	0.53
	LIONZ						
AUC	0.76	0.74	0.76	0.57	0.50	0.76	0.69
AUC sd	0.14	0.16	0.13	0.15	0.14	0.14	0.13
AUC ext	0.60	0.42	0.57	0.46	0.48	0.58	0.58

DeLong test conversion matrices

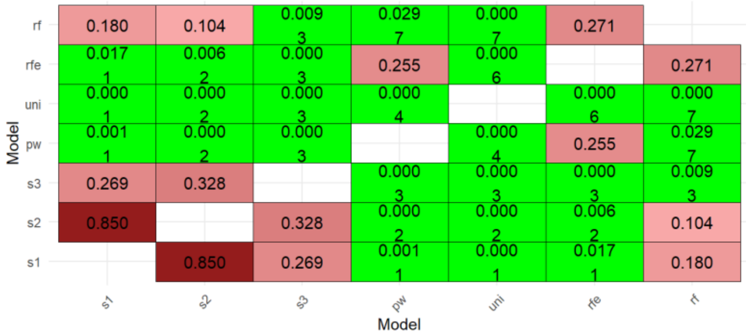
SUV4



Model	s1	s2	s3	pw	uni	rfe	rf
rf	0.628	0.668	0.539	0.030 7	0.105	0.557	
rfe	0.322	0.339	0.962	0.129	0.343		0.557
uni	0.051	0.051	0.387	0.520		0.343	0.105
pw	0.015 1	0.014 2	0.155		0.520	0.129	0.030 7
s3	0.316	0.333		0.155	0.387	0.962	0.539
s2	0.944		0.333	0.014 2	0.051	0.339	0.668
s1		0.944	0.316	0.015 1	0.051	0.322	0.628

Figure S20: Comparison matrix of the DeLong test with SUV4 segmentation for the cross-validation of the training cohort, with OS36 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

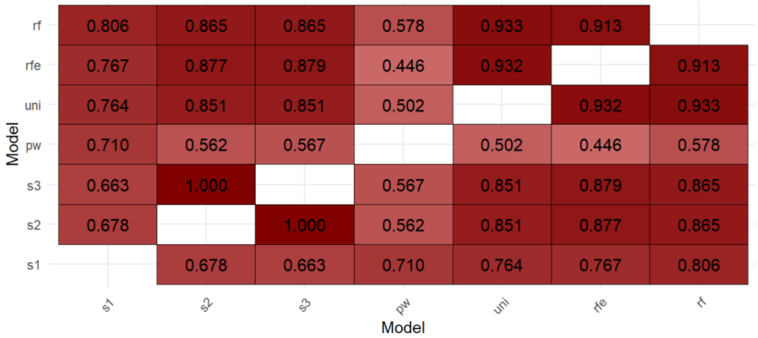
MV2



Model	s1	s2	s3	pw	uni	rfe	rf
rf	0.180	0.104	0.009 3	0.029 7	0.000 7	0.271	
rfe	0.017 1	0.006 2	0.000 3	0.255	0.000 6		0.271
uni	0.000 1	0.000 2	0.000 3	0.000 4		0.000 6	0.000 7
pw	0.001 1	0.000 2	0.000 3		0.000 4	0.255	0.029 7
s3	0.269	0.328		0.000 3	0.000 3	0.000 3	0.009 3
s2	0.850		0.328	0.000 2	0.000 2	0.006 2	0.104
s1		0.850	0.269	0.001 1	0.000 1	0.017 1	0.180

Figure S21: Comparison matrix of the DeLong test with MV2 segmentation for the cross-validation of the training cohort, with OS36 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

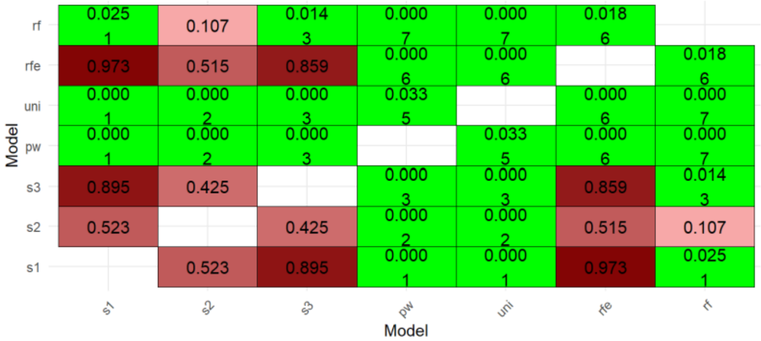
MV2 external validation



Model	s1	s2	s3	pw	uni	rfe	rf
rf	0.806	0.865	0.865	0.578	0.933	0.913	
rfe	0.767	0.877	0.879	0.446	0.932		0.913
uni	0.764	0.851	0.851	0.502		0.932	0.933
pw	0.710	0.562	0.567		0.502	0.446	0.578
s3	0.663	1.000		0.567	0.851	0.879	0.865
s2	0.678		1.000	0.562	0.851	0.877	0.865
s1		0.678	0.663	0.710	0.764	0.767	0.806

Figure S22: Comparison matrix of the DeLong test with MV2 segmentation for the external validation with the validation cohort, with OS36 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

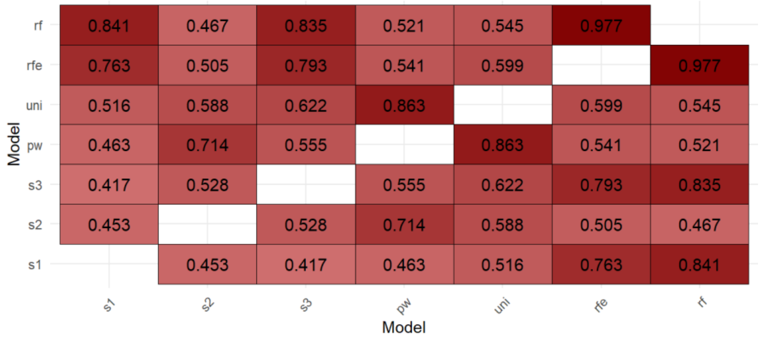
LIONZ



	s1	s2	s3	pw	uni	rfe	rf
rf	0.025 1	0.107	0.014 3	0.000 7	0.000 7	0.018 6	
rfe	0.973	0.515	0.859	0.000 6	0.000 6		0.018 6
uni	0.000 1	0.000 2	0.000 3	0.033 5		0.000 6	0.000 7
pw	0.000 1	0.000 2	0.000 3		0.033 5	0.000 6	0.000 7
s3	0.895	0.425		0.000 3	0.000 3	0.859	0.014 3
s2	0.523		0.425	0.000 2	0.000 2	0.515	0.107
s1		0.523	0.895	0.000 1	0.000 1	0.973	0.025 1

Figure S23: Comparison matrix of the DeLong test with LIONZ segmentation for the cross-validation of the training cohort, with OS36 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

LIONZ external validation



	s1	s2	s3	pw	uni	rfe	rf
rf	0.841	0.467	0.835	0.521	0.545	0.977	
rfe	0.763	0.505	0.793	0.541	0.599		0.977
uni	0.516	0.588	0.622	0.863		0.599	0.545
pw	0.463	0.714	0.555		0.863	0.541	0.521
s3	0.417	0.528		0.555	0.622	0.793	0.835
s2	0.453		0.528	0.714	0.588	0.505	0.467
s1		0.453	0.417	0.463	0.516	0.763	0.841

Figure S24: Comparison matrix of the DeLong test with LIONZ segmentation for the external validation with the external cohort, with OS36 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

Appendix D: Results for time to progression of 24 months

SUV4 segmentation

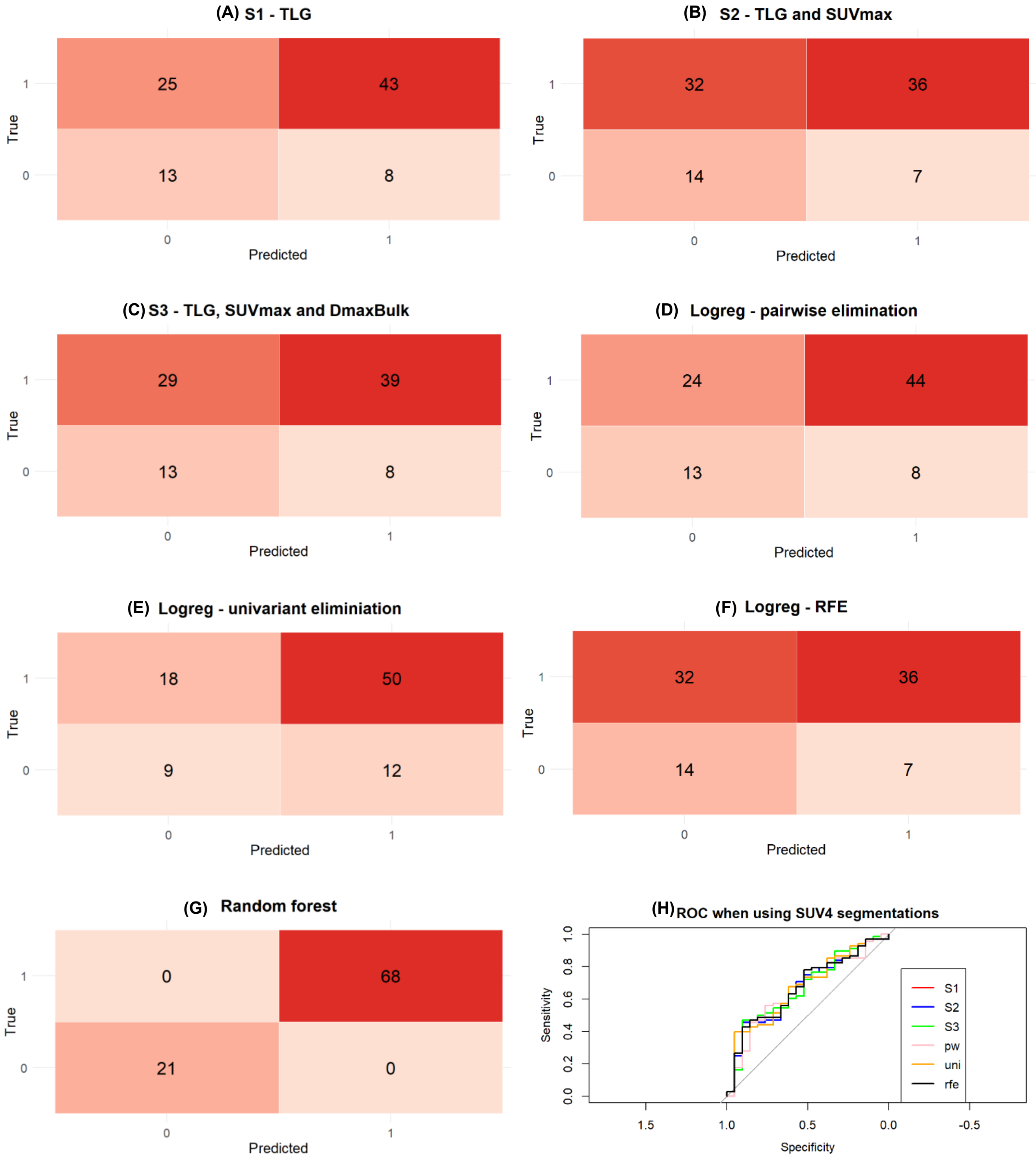


Figure S25: Results with SUV4 segmentation for the cross-validated prognostic models using TTP24 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the cross-validated prognostic models using SUV4 segmentation. Logreg: logistic regression.

MV2 segmentation

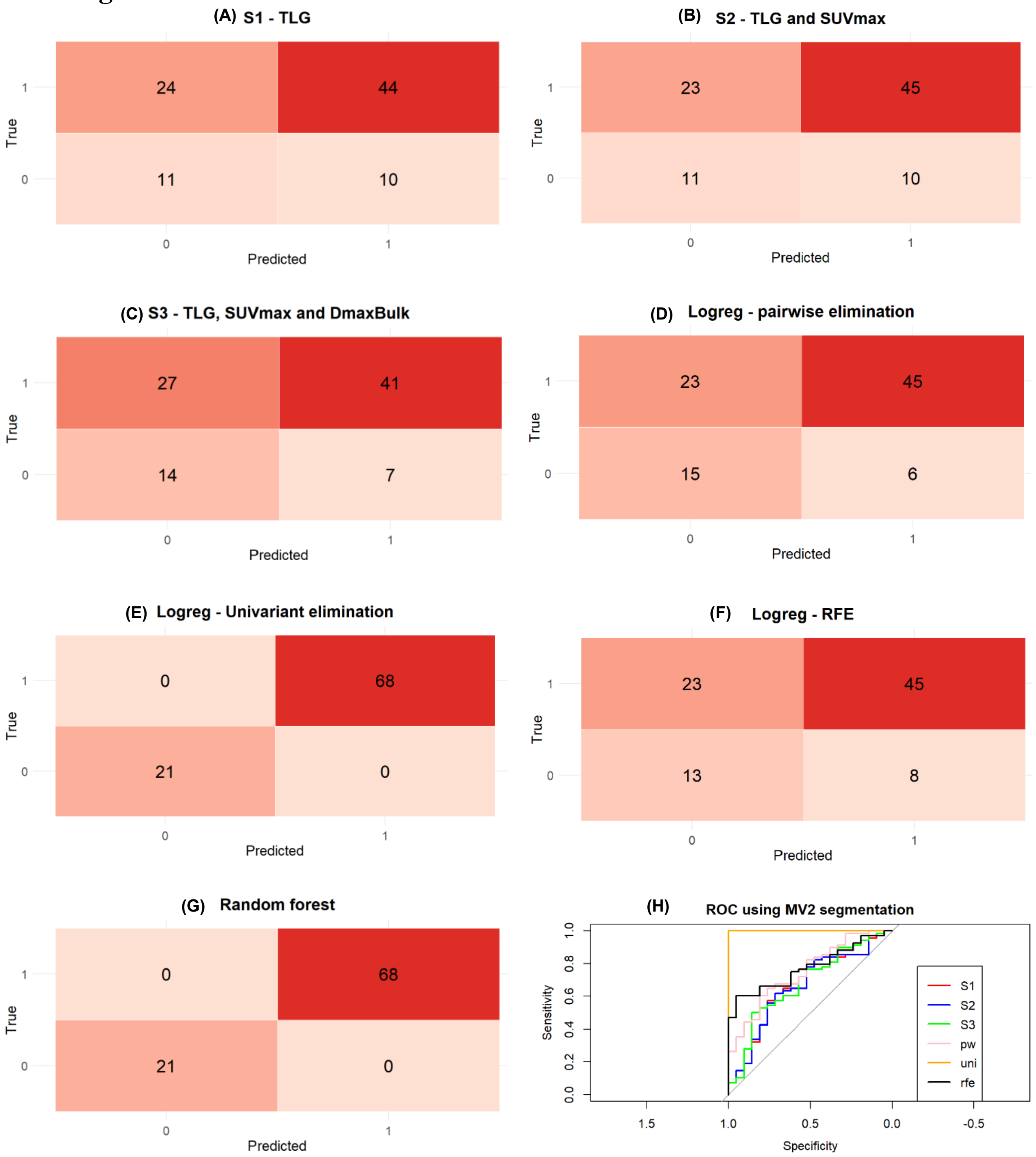


Figure S26: Results with MV2 segmentation for the cross-validated prognostic models using TTP24 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the cross-validated prognostic models using MV2 segmentation. Logreg: logistic regression.

MV2 segmentation external validation

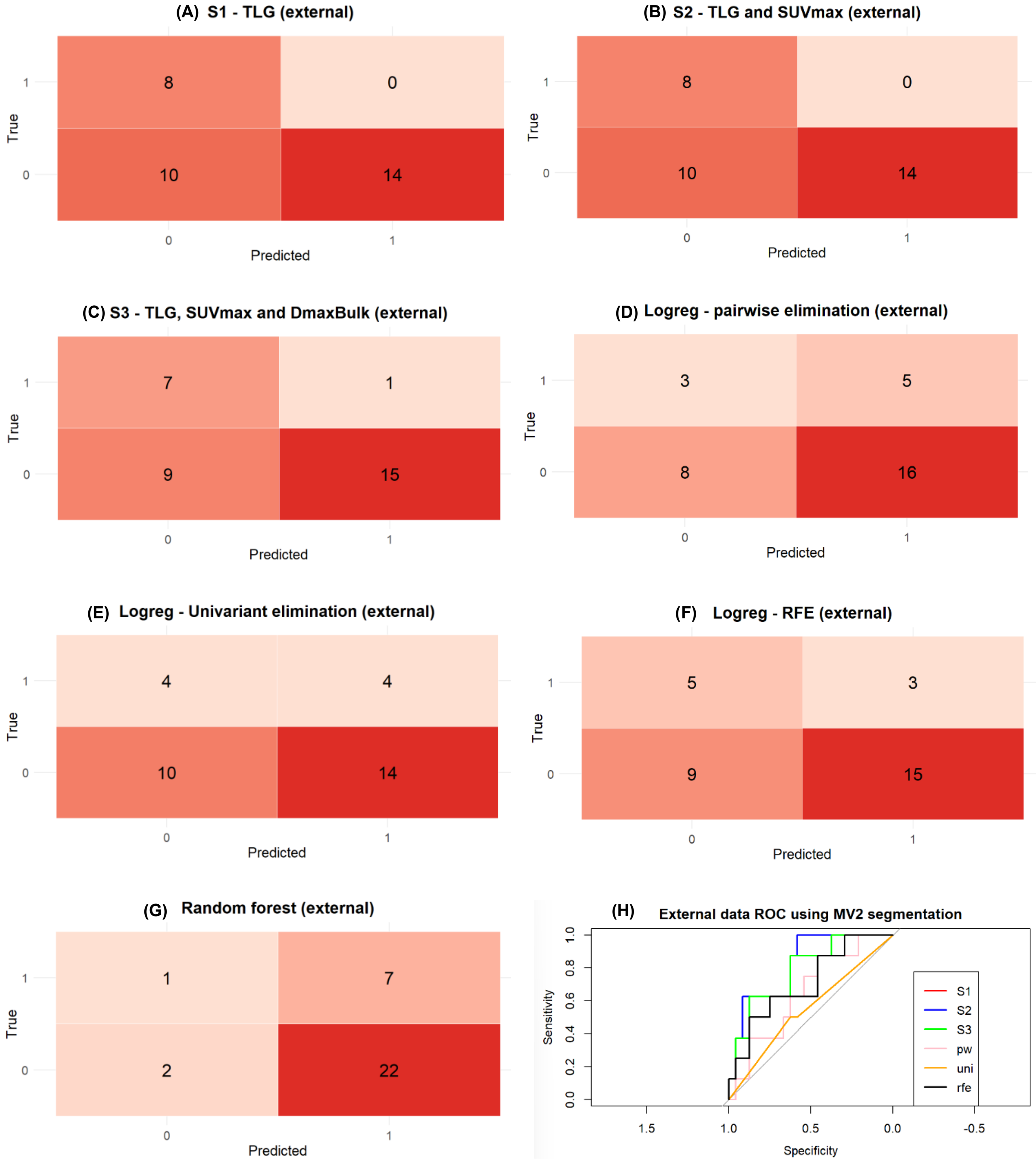


Figure S27: Results with MV2 segmentation for the externally validated prognostic models using TTP24 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the externally validated prognostic models using MV2 segmentation. Logreg: logistic regression.

LIONZ segmentation

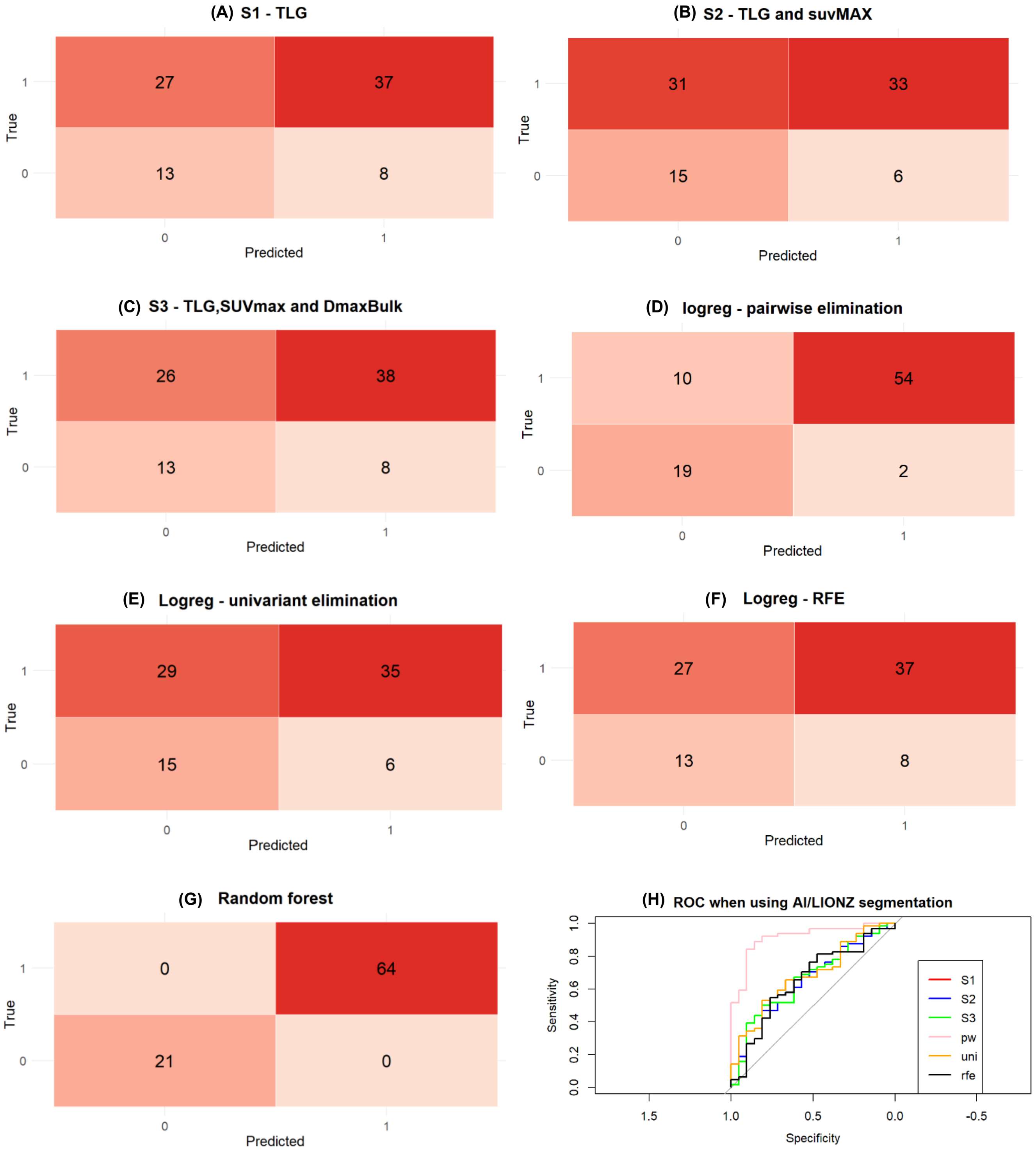


Figure S28: Results with LIONZ segmentation for the cross-validated prognostic models using TTP24 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the cross-validated prognostic models using LIONZ segmentation. Logreg: logistic regression.

LIONZ segmentation external validation

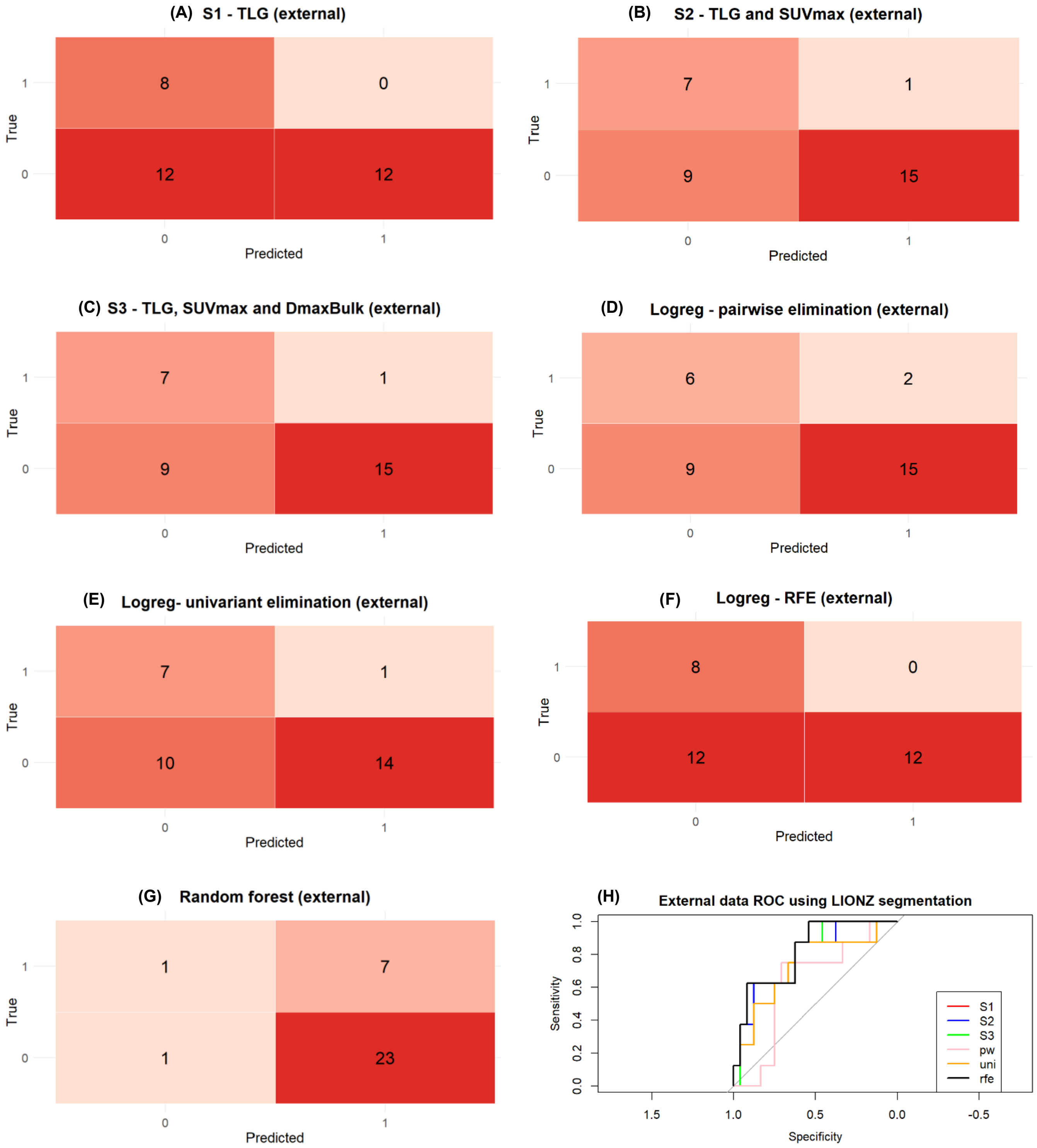


Figure S29: Results with LIONZ segmentation for the externally validated prognostic models using TTP24 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the externally validated prognostic models using LIONZ segmentation. Logreg: logistic regression.

General information

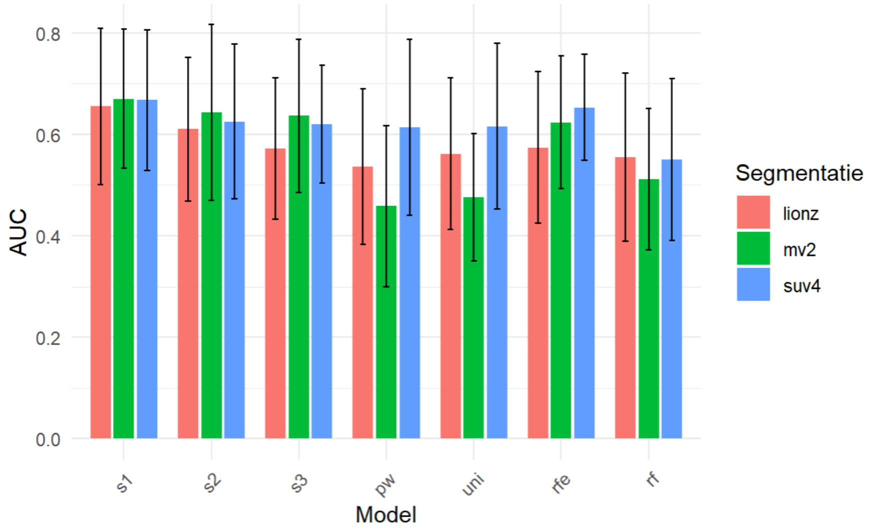


Figure S30: Histogram comparing the AUC values of all cross-validated models using LIONZ, MV2, and SUV4 segmentations with a time to progression of 24 months as prognostic outcome. Orange bars represent models using LIONZ segmentations, green bars represent models using MV2 segmentations, and blue bars represent models using SUV4 segmentations.

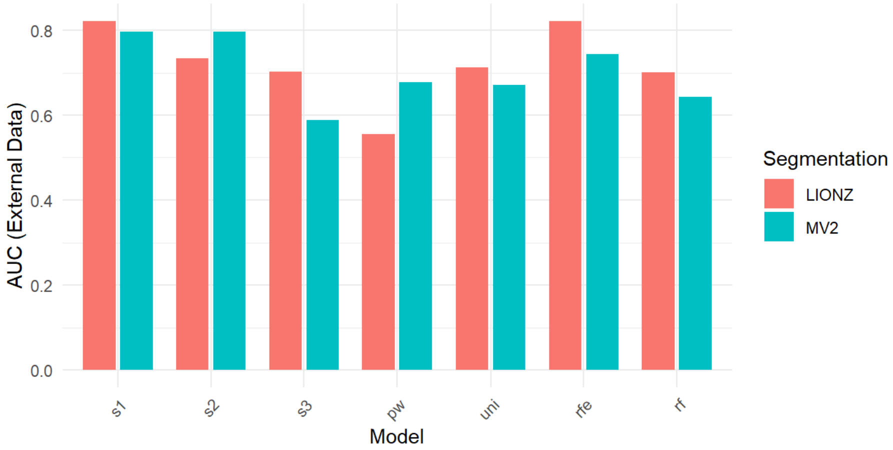


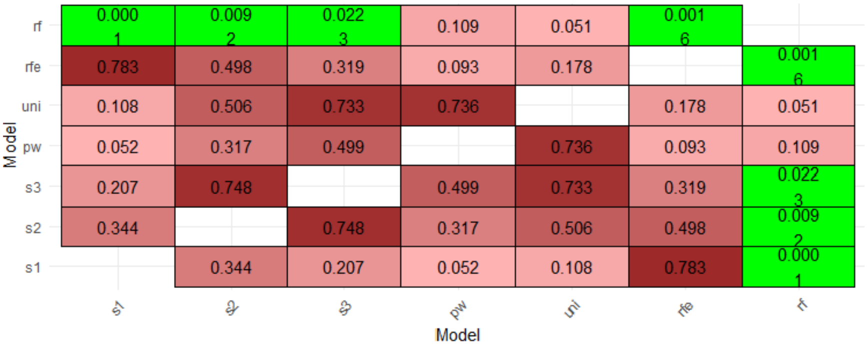
Figure S31: Histogram comparing the AUC values of all externally validated models obtained using LIONZ and MV2 segmentations with a time to progression of 24 months as prognostic outcome. Orange bars represent models using LIONZ segmentations, blue bars represent models using MV2 segmentation.

Table S3: Cross-validated AUC values with standard deviation and externally validated AUC values of prognostic models using SUV4, MV2, and LIONZ as segmentation methods. All models use a time to progression of 24 months as prognostic outcome. AUC sd: area under the curve standard deviation, AUC ext: AUC of external validation.

TTP24	s1	s2	s3	pw	uni	rfe	rf
	SUV4						
AUC	0.67	0.62	0.62	0.61	0.62	0.65	0.54
AUCsd	0.14	0.15	0.12	0.17	0.16	0.10	0.16
	MV2						
AUC	0.67	0.64	0.64	0.46	0.48	0.62	0.51
AUC sd	0.14	0.17	0.15	0.16	0.12	0.13	0.14
AUC ext	0.82	0.82	0.78	0.65	0.55	0.71	0.66
	LIONZ						
AUC	0.65	0.61	0.57	0.54	0.56	0.57	0.55
AUC sd	0.15	0.14	0.14	0.15	0.15	0.15	0.17
AUC ext	0.82	0.78	0.80	0.63	0.73	0.82	0.63

DeLong test conversion matrices

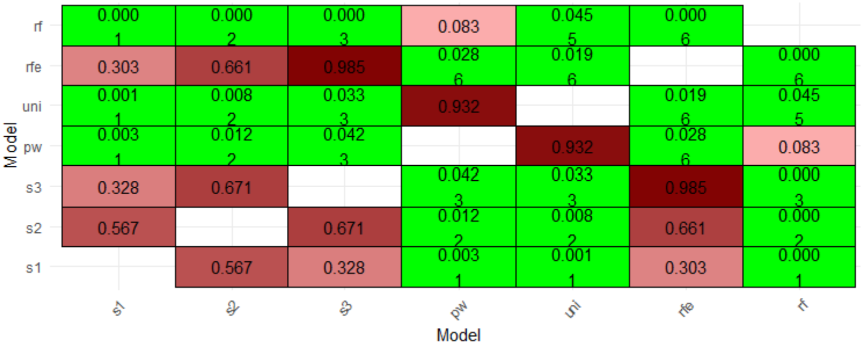
SUV4



	s1	s2	s3	pw	uni	rfe	rf
rf	0.000 1	0.009 2	0.022 3	0.109	0.051	0.001 6	
rfe	0.783	0.498	0.319	0.093	0.178		0.001 6
uni	0.108	0.506	0.733	0.736		0.178	0.051
pw	0.052	0.317	0.499		0.736	0.093	0.109
s3	0.207	0.748		0.499	0.733	0.319	0.022 3
s2	0.344		0.748	0.317	0.506	0.498	0.009 2
s1		0.344	0.207	0.052	0.108	0.783	0.000 1

Figure S32: Comparison matrix of the DeLong test with SUV4 segmentation for the cross-validation of the training cohort, with TTP24 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

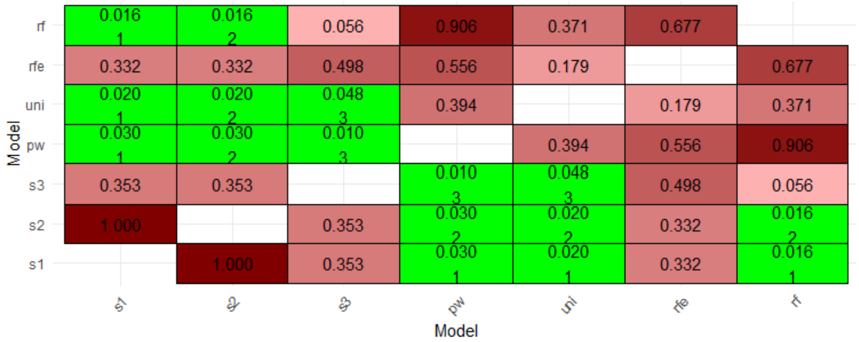
MV2



	s1	s2	s3	pw	uni	rfe	rf
rf	0.000 1	0.000 2	0.000 3	0.083	0.045 5	0.000 6	
rfe	0.303	0.661	0.985	0.028 6	0.019 6		0.000 6
uni	0.001 1	0.008 2	0.033 3	0.932		0.019 6	0.045 5
pw	0.003 1	0.012 2	0.042 3		0.932	0.028 6	0.083
s3	0.328	0.671		0.042 3	0.033 3	0.985	0.000 3
s2	0.567		0.671	0.012 2	0.008 2	0.661	0.000 2
s1		0.567	0.328	0.003 1	0.001 1	0.303	0.000 1

Figure S33: Comparison matrix of the DeLong test with MV2 segmentation for the cross-validation of the training cohort, with TTP24 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

MV2 external validation



	s1	s2	s3	pw	uni	rfe	rf
rf	0.016 1	0.016 2	0.056	0.906	0.371	0.677	
rfe	0.332	0.332	0.498	0.556	0.179		0.677
uni	0.020 1	0.020 2	0.048 3	0.394		0.179	0.371
pw	0.030 1	0.030 2	0.010 3		0.394	0.556	0.906
s3	0.353	0.353		0.010 3	0.048 3	0.498	0.056
s2	1.000		0.353	0.030 2	0.020 2	0.332	0.016 2
s1		1.000	0.353	0.030 1	0.020 1	0.332	0.016 1

Figure S34: Comparison matrix of the DeLong test with MV2 segmentation for the external validation with the external cohort, with TTP24 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

LIONZ

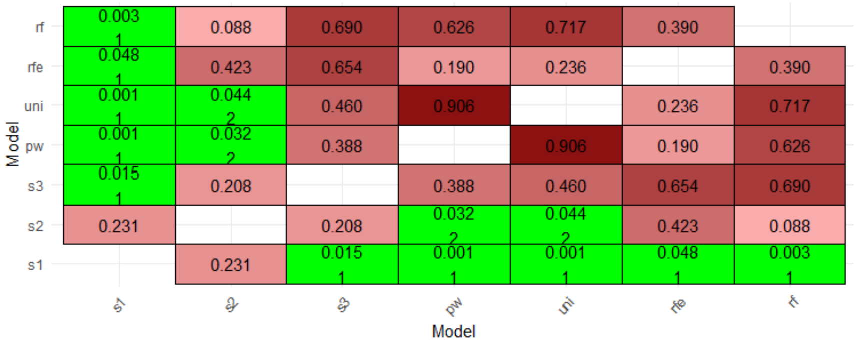


Figure S35: Comparison matrix of the DeLong test with LIONZ segmentation for the cross-validation of the training cohort, with TTP24 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

LIONZ external validation

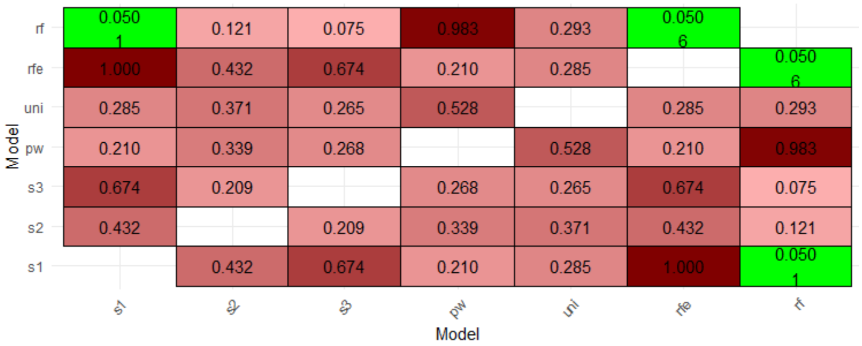


Figure S36: Comparison matrix of the DeLong test with LIONZ segmentation for the external validation with the external cohort, with TTP24 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

Appendix E: Results for time to progression of 36 months

SUV4 segmentation

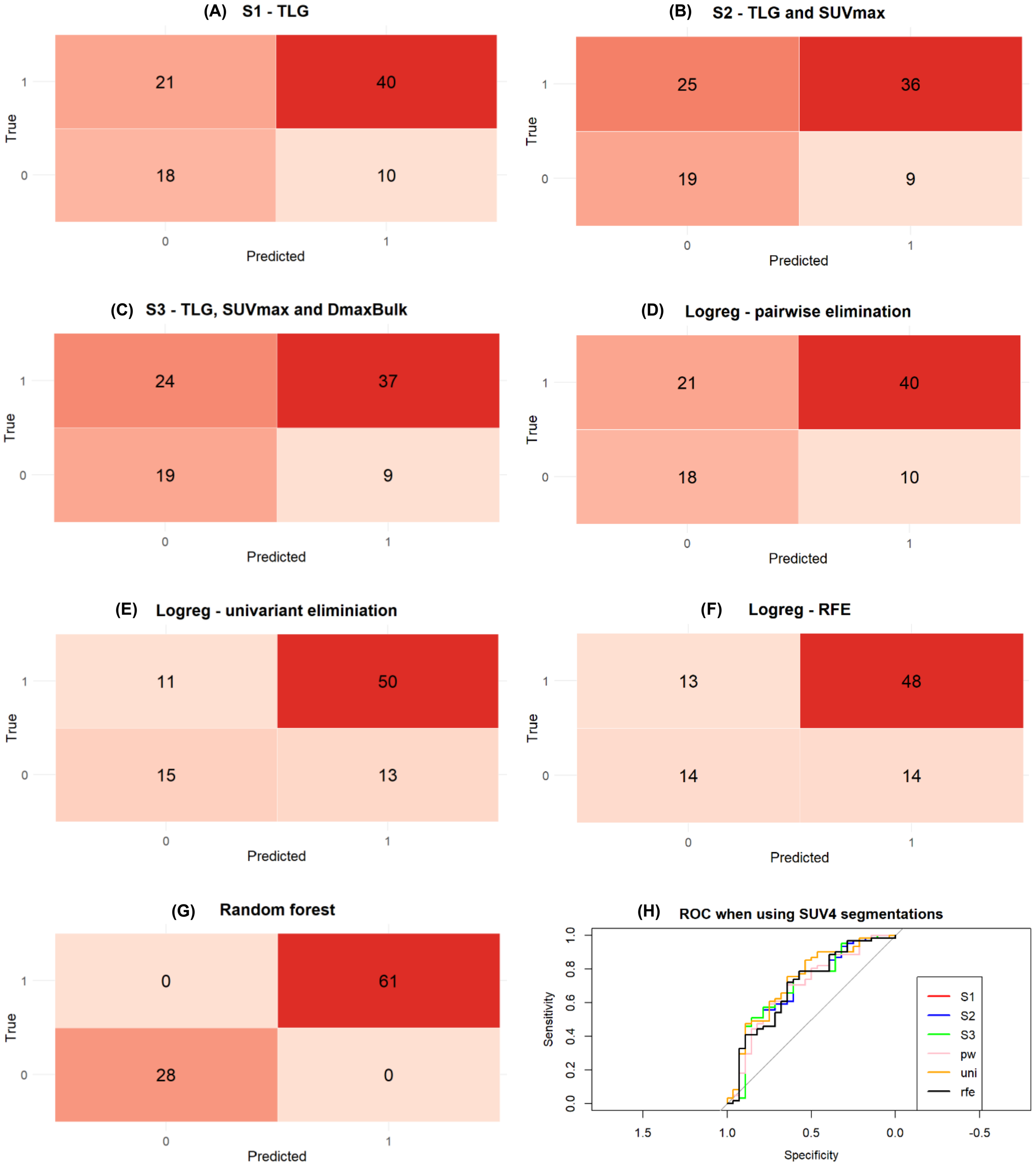


Figure S37: Results with SUV4 segmentation for the cross-validated prognostic models using TTP36 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the cross-validated prognostic models using SUV4 segmentation. Logreg: logistic regression.

MV2 segmentation

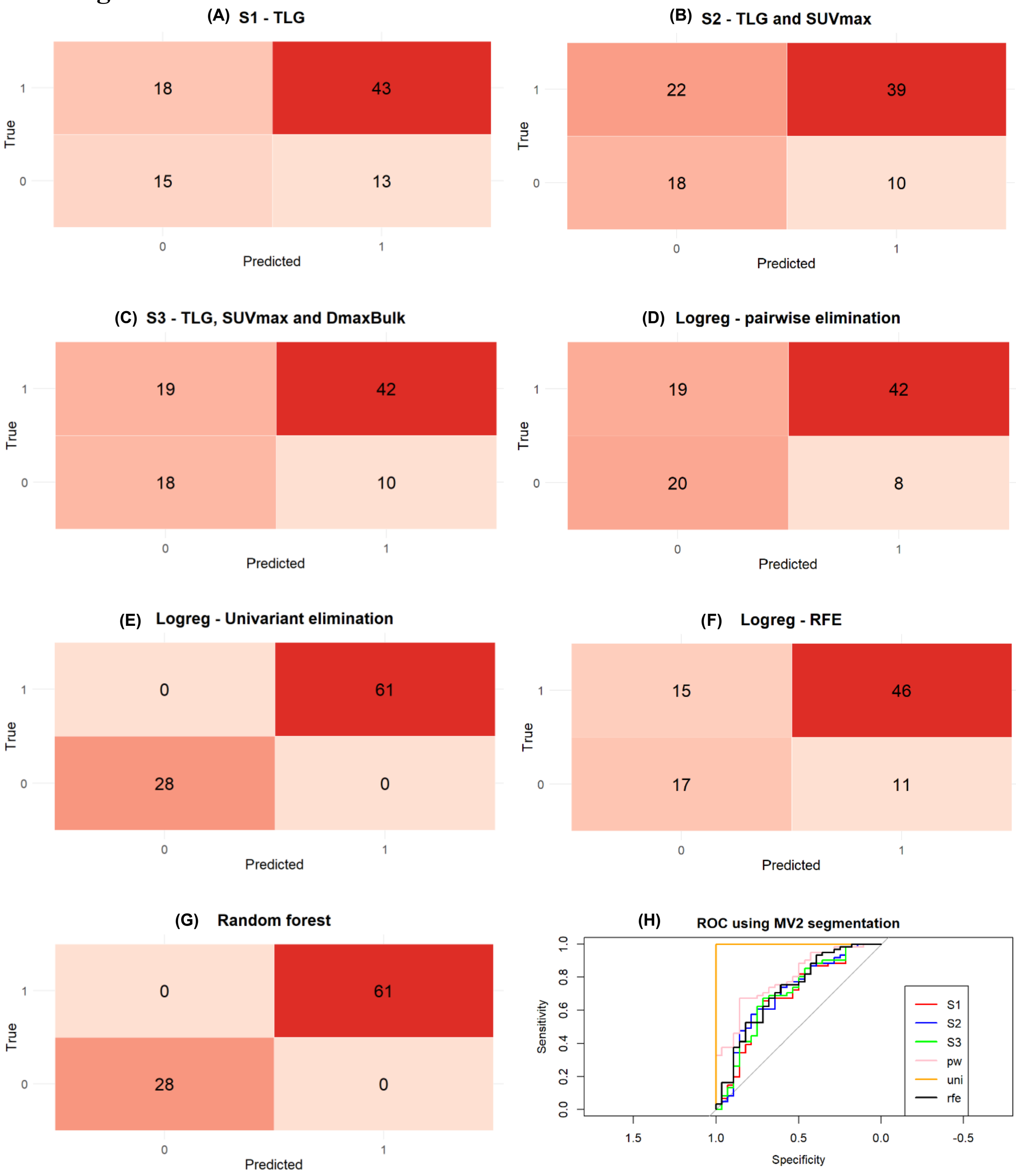


Figure S38: Results with MV2 segmentation for the cross-validated prognostic models using TTP36 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the cross-validated prognostic models using MV2 segmentation. Logreg: logistic regression.

MV2 segmentation external validation

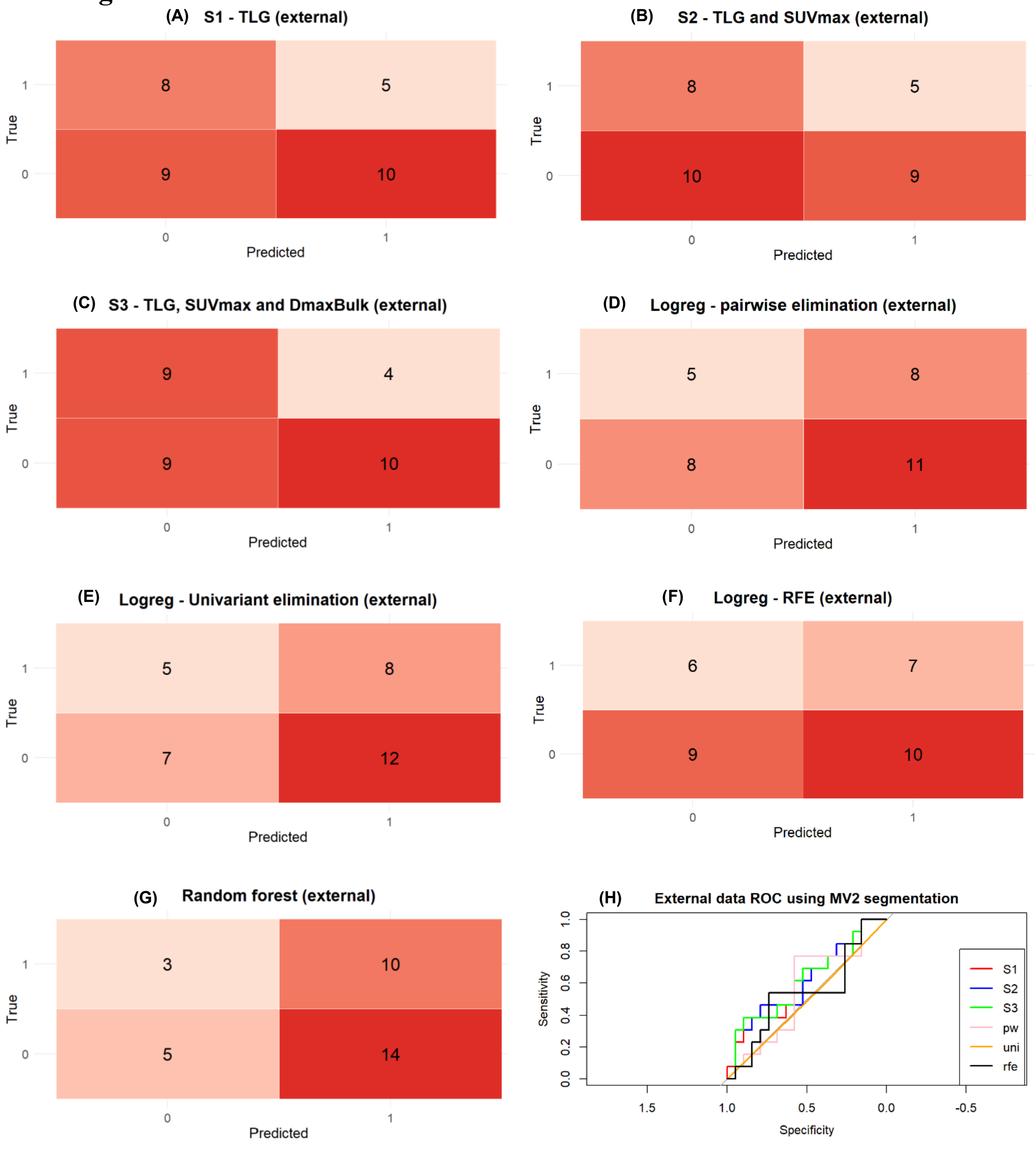


Figure S39: Results with MV2 segmentation for the externally validated prognostic models using TTP36 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the externally validated prognostic models using MV2 segmentation. Logreg: logistic regression..

LIONZ segmentation

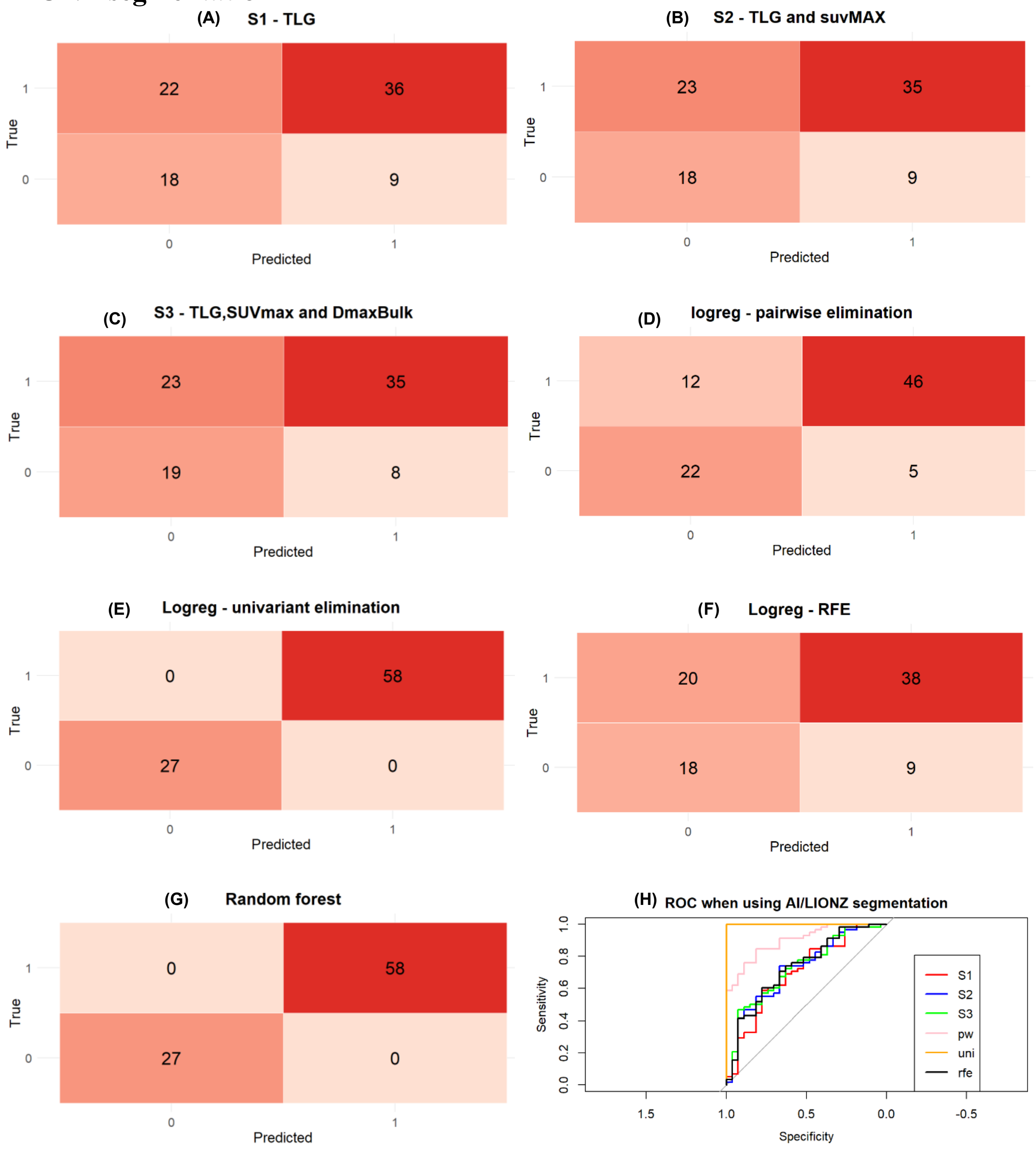


Figure S40: Results with LIONZ segmentation for the cross-validated prognostic models using TTP36 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the cross-validated prognostic models using LIONZ segmentation. Logreg: logistic regression.

LIONZ segmentation external validation

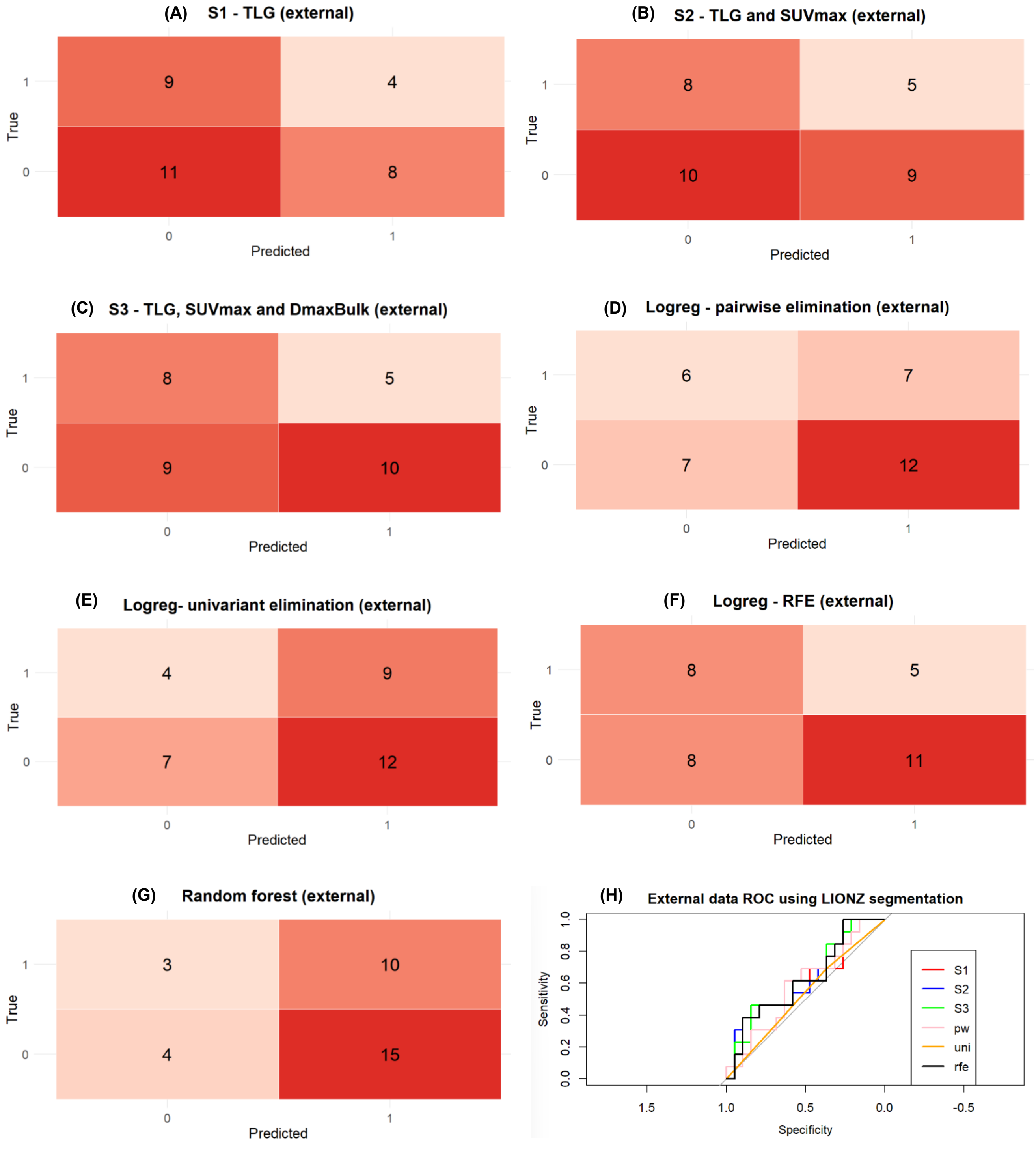


Figure S41: Results with LIONZ segmentation for the externally validated prognostic models using TTP36 as the prognostic outcome. (A) The confusion matrix for the s1 model is a TLG-only logistic regression model; (B) The confusion matrix for the s2 model is a logistic regression model with the TLG and SUVmax; (C) The confusion matrix for the s3 model is a logistic regression model with the TLG, SUVmax and DmaxBulk; (D) The confusion matrix for a logistic regression model after pairwise elimination; (E) The confusion matrix for a logistic regression model after univariate selection; (F) the confusion matrix for a logistic regression model after recursive feature elimination; (G) The confusion matrix for a random forest model; (H) The ROC curve for the externally validated prognostic models using LIONZ segmentation. Logreg: logistic regression.

General information

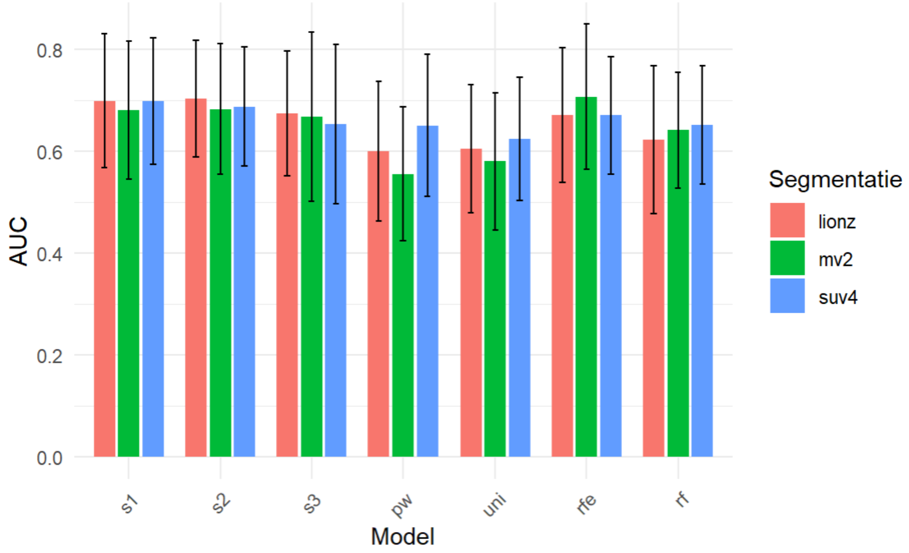


Figure S42: Histogram comparing the AUC values of all cross-validated models using LIONZ, MV2, and SUV4 segmentations with a time to progression of 36 months as prognostic outcome. Orange bars represent models using LIONZ segmentations, green bars represent models using MV2 segmentations, and blue bars represent models using SUV4 segmentations.

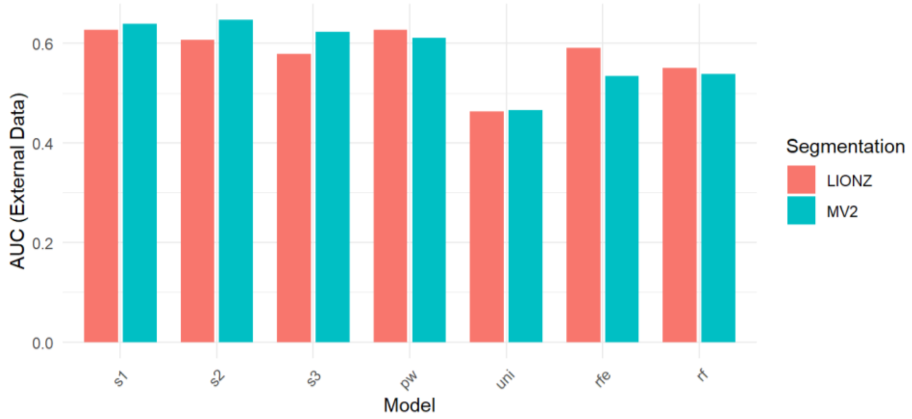


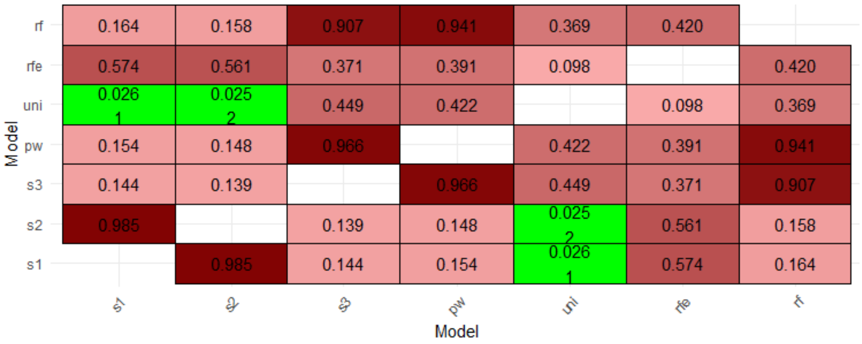
Figure S43: Histogram comparing the AUC values of all externally validated models obtained using LIONZ and MV2 segmentations with a time to progression of 36 months as prognostic outcome. Orange bars represent models using LIONZ segmentations, blue bars represent models using MV2 segmentation.

Table S4: Cross-validated AUC values with standard deviation and externally validated AUC values of prognostic models using SUV4, MV2, and LIONZ as segmentation methods. All models use a time to progression of 36 months as prognostic outcome. AUC sd: area under the curve standard deviation, AUC ext: AUC of external validation.

TTP36	s1	s2	s3	pw	uni	rfe	rf
	SUV4						
AUC	0.70	0.69	0.65	0.65	0.62	0.67	0.65
AUCsd	0.12	0.12	0.16	0.14	0.12	0.12	0.11
	MV2						
AUC	0.68	0.68	0.67	0.56	0.58	0.71	0.64
AUC sd	0.14	0.13	0.17	0.13	0.13	0.14	0.11
AUC ext	0.61	0.62	0.62	0.56	0.49	0.54	0.53
	LIONZ						
AUC	0.70	0.70	0.67	0.60	0.61	0.67	0.62
AUC sd	0.13	0.11	0.12	0.14	0.13	0.13	0.14
AUC ext	0.63	0.63	0.62	0.59	0.53	0.62	0.49

DeLong test conversion matrices

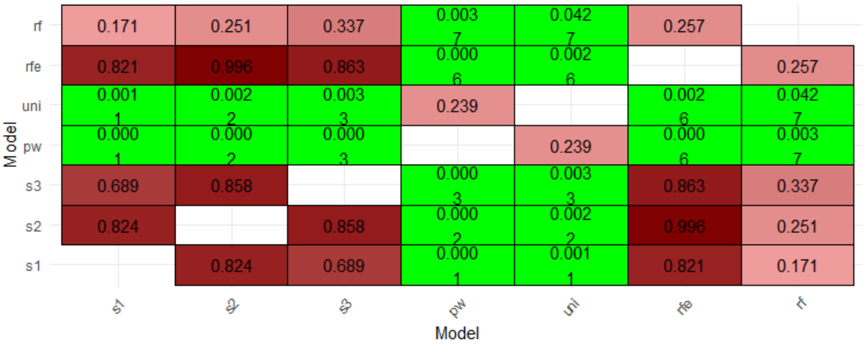
SUV4



	s1	s2	s3	pw	uni	rfe	rf
rf	0.164	0.158	0.907	0.941	0.369	0.420	
rfe	0.574	0.561	0.371	0.391	0.098		0.420
uni	0.026 1	0.025 2	0.449	0.422		0.098	0.369
pw	0.154	0.148	0.966		0.422	0.391	0.941
s3	0.144	0.139		0.966	0.449	0.371	0.907
s2	0.985		0.139	0.148	0.025 2	0.561	0.158
s1		0.985	0.144	0.154	0.026 1	0.574	0.164

Figure S44: Comparison matrix of the DeLong test with SUV4 segmentation for the cross-validation of the training cohort, with TTP36 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

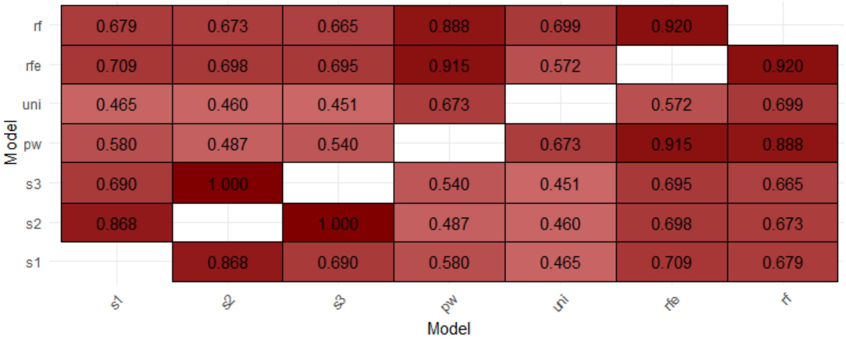
MV2



	s1	s2	s3	pw	uni	rfe	rf
rf	0.171	0.251	0.337	0.003 7	0.042 7	0.257	
rfe	0.821	0.996	0.863	0.000 6	0.002 6		0.257
uni	0.001 1	0.002 2	0.003 3	0.239		0.002 6	0.042 7
pw	0.000 1	0.000 2	0.000 3		0.239	0.000 6	0.003 7
s3	0.689	0.858		0.000 3	0.003 3	0.863	0.337
s2	0.824		0.858	0.000 2	0.002 2	0.996	0.251
s1		0.824	0.689	0.000 1	0.001 1	0.821	0.171

Figure S45: Comparison matrix of the DeLong test with MV2 segmentation for the cross-validation of the training cohort, with TTP36 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

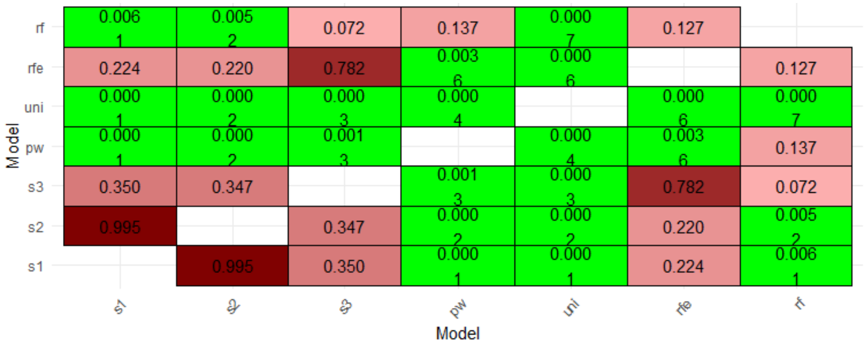
MV2 external validation



	s1	s2	s3	pw	uni	rfe	rf
rf	0.679	0.673	0.665	0.888	0.699	0.920	
rfe	0.709	0.698	0.695	0.915	0.572		0.920
uni	0.465	0.460	0.451	0.673		0.572	0.699
pw	0.580	0.487	0.540		0.673	0.915	0.888
s3	0.690	1.000		0.540	0.451	0.695	0.665
s2	0.868		1.000	0.487	0.460	0.698	0.673
s1		0.868	0.690	0.580	0.465	0.709	0.679

Figure S46: Comparison matrix of the DeLong test with MV2 segmentation for the external validation with the validation cohort, with TTP36 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

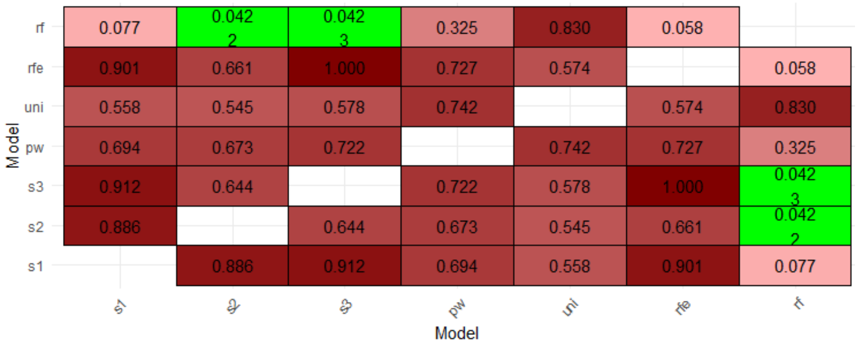
LIONZ



Model	s1	s2	s3	pw	uni	rfe	rf
rf	0.006 1	0.005 2	0.072	0.137	0.000 7	0.127	
rfe	0.224	0.220	0.782	0.003 6	0.000 6		0.127
uni	0.000 1	0.000 2	0.000 3	0.000 4		0.000 6	0.000 7
pw	0.000 1	0.000 2	0.001 3		0.000 4	0.003 6	0.137
s3	0.350	0.347		0.001 3	0.000 3	0.782	0.072
s2	0.995		0.347	0.000 2	0.000 2	0.220	0.005 2
s1		0.995	0.350	0.000 1	0.000 1	0.224	0.006 1

Figure S47: Comparison matrix of the DeLong test with LIONZ segmentation for the cross-validation of the training cohort, with TTP36 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.

LIONZ external validation



Model	s1	s2	s3	pw	uni	rfe	rf
rf	0.077	0.042 2	0.042 3	0.325	0.830	0.058	
rfe	0.901	0.661	1.000	0.727	0.574		0.058
uni	0.558	0.545	0.578	0.742		0.574	0.830
pw	0.694	0.673	0.722		0.742	0.727	0.325
s3	0.912	0.644		0.722	0.578	1.000	0.042 3
s2	0.886		0.644	0.673	0.545	0.661	0.042 2
s1		0.886	0.912	0.694	0.558	0.901	0.077

Figure S48: Comparison matrix of the DeLong test with LIONZ segmentation for the external validation with the validation cohort, with TTP36 as prognostic outcome. Green boxes indicate a statistically significant difference between the two models, with the number of the best model shown. Red boxes indicate a non-statistically significant difference between the two models.