Prognostication from Longitudinal Multisequence Brain MRI using Artificial Intelligence

Iris van der Loo

Prognostication from Longitudinal Multisequence Brain MRI using Artificial Intelligence

Author:
Iris van der Loo, BSc

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in

Technical Medicine

UNIVERSITY OF TWENTE.

Deventer ziekenhuis

NETHERLANDS CANCER INSTITUTE
ANTONI VAN LEEUWENHOEK

Abstract

Background: Brain tumors are among the deadliest cancers, are difficult to treat and often cause disease or treatment-related side effects. Therefore, evaluating tumor response to treatment plays an important role in treatment decision-making. Magnetic Resonance Imaging (MRI) is the most sensitive modality for the evaluation of brain tumors. Response Assessment in Neuro-Oncology (RANO) criteria are currently the most used to quantify treatment response. Important limitations of the RANO criteria are user-dependency, limited reproducibility and limited use of all information present in the scan. Prognostic artificial intelligence monitoring shows high potential in overcoming these limitations.

Methods: As a first step in the conversion to MRI 571 patients were included to make a prognosis based on the fluid-attenuated inversion recovery sequence. The algorithm was adjusted to multiple sequences ($n = 576$) and externally validated ($n = 119$). All patients with a primary malignancy of the brain ($n = 54$) were used to compare to the RANO criteria and volumetric assessments and externally validated ($n = 62$). Primary outcome measures were the concordance index and Kaplan Meier survival curves. Cox time-varying regression was used for associative analysis.

Results: For the fluid-attenuated inversion recovery sequence the logrank test on all three risk groups in the Kaplan Meier survival curves showed a significant difference ($p < 0.005$) with a concordance index of 0.61. Multisequence analysis of the complete dataset revealed a concordance index of 0.62 with a significant difference between the medium and high-risk patient groups ($p < 0.005$). The external validation showed a concordance index of 0.55 with a significant difference between medium and high-risk patient groups ($p = 0.01$). In both datasets the difference between low and medium-risk patients was insignificant. Comparison with other methods showed volumetric assessment had the best prognostic performance with a concordance index of 0.77.

Conclusion: This thesis has provided evidence for an alternative method of longitudinal monitoring of cancer patients from brain MRI. Unlike current response evaluation methods, this method works fully automatically and uses all information in the scan including disease or treatment-related side effects. Further improvements are needed to reach equal performance across datasets.

Graduation Committee

Chair
Prof.dr.ir. C.H. Slump
Robotics and Mechatronics, University of Twente, Enschede, The Netherlands

Clinical supervisor
Dr. A.L.T. Imholz
Internal Medicine, Deventer Ziekenhuis, Deventer, The Netherlands

Technical supervisor
Dr. C.O. Tan
Robotics and Mechatronics, University of Twente, Enschede, The Netherlands

Process supervisor
Drs. P.A. van Katwijk
Technical Medicine, University of Twente, Enschede, The Netherlands

Additional member and technical supervisor
Dr. S. Trebeschi
Postdoc researcher, The Netherlands Cancer Institute, Amsterdam, The Netherlands

External member
Dr. J.M. Wolterink
Mathematics of Imaging & AI, University of Twente, Enschede, The Netherlands

Acknowledgements

In the past year, I have learned to apply artificial intelligence to solve actual clinical problems at the department of radiology at the Netherlands Cancer Institute and the department of internal medicine at Deventer Ziekenhuis. I would like to thank prof. dr. Regina Beets-Tan and dr. Alex Imholz for this opportunity. I have further developed my clinical skills and have learned how to apply artificial intelligence in this clinical context. I am grateful to have had this opportunity and excited to continue the research in the future.

I would like to thank dr. Can Tan for your supervision and meaningful insights into the technical aspect of the thesis. Your supervision has kept me sharp throughout the thesis.

The guidance of dr. Alex Imholz has helped me develop and increase my confidence in my clinical work. Your enthusiasm and interest in the topic have been very helpful. Throughout the past year, you have reminded me to always keep an eye out for the clinician's perspective.

I would also like to express my gratitude to drs. Paul van Katwijk. You have helped me with my personal development and getting to know myself better. You always know how to ask the right questions which kept me thinking also after the intervision sessions.

Daily supervision was done by dr. Stefano Trebeschi. Thank you for all your time and effort. Our weekly meetings always steered me in the right direction and kept my focus sharp. Whenever I encountered a problem outside of those meetings you were always there to help me out.

I would also like to thank all of my co-workers at the NKI and Deventer Ziekenhuis. You were always open to questions and interesting discussions. It has been a pleasure to work with you.

Finally, I would like to thank my parents, Femke and Peter for always stimulating me to follow my interests. You have been very supportive and I know you will always have my back.

I hope you enjoy reading this thesis.

Contents

Abstract ii

Graduation Committee iii

Acknowledgements iv

1 General clinical background 2
- 1.1 Cancer and metastatic disease . . . 2
- 1.2 Cancer therapies . . . 3
- 1.3 Clinical decision making in oncology . . . 5
- 1.4 Proposed improvements . . . 8

2 General technological background 11
- 2.1 Magnetic Resonance Imaging . . . 11
- 2.2 Artificial intelligence . . . 15
- 2.3 Machine learning . . . 16
- 2.4 Deep learning . . . 17
- 2.5 Convolutional neural networks . . . 18
- 2.6 Prognostic AI monitoring . . . 19
- 2.7 Layers . . . 20
- 2.8 Loss function . . . 24
- 2.9 Training . . . 24

3 Extension of PAM to single sequence brain imaging 28
- 3.1 Introduction . . . 28
- 3.2 Technological background . . . 31
- 3.3 Methods . . . 35
- 3.4 Results . . . 42
- 3.5 Discussion . . . 52
- 3.6 Conclusion . . . 54

4 Extension of PAM to multisequence imaging of the brain 56
- 4.1 Introduction . . . 56
- 4.2 Methods . . . 61

4.3 Results 65
4.4 Discussion 74
4.5 Conclusion 77

5 Comparison with current methods for response monitoring 80
5.1 Introduction 80
5.2 Method 82
5.3 Results 84
5.4 Discussion 88
5.5 Conclusion 90

6 General conclusion and future outlook 92
6.1 General conclusion 92
6.2 Future outlook 93

List of Figures

1.1 Hallmarks of cancer . . . 3

2.1 Orientation of hydrogen nuclei . . . 12
2.2 Magnetization after an RF pulse . . . 13
2.3 Magnetic field gradient . . . 13
2.4 T1 relaxation . . . 14
2.5 T2 relaxation . . . 15
2.6 Artificial intelligence . . . 16
2.7 Schematic overview of a neuron . . . 17
2.8 Deep neural network architecture . . . 18
2.9 Convolutional neural network architecture . . . 19
2.10 Convolutional layer . . . 21
2.11 Rectified linear unit function . . . 22
2.12 Max pooling layer . . . 22
2.13 Group normalization . . . 23

3.1 The blood-brain barrier . . . 29
3.2 Network architecture for affine transformation . . . 32
3.3 Network architecture for elastic transformation . . . 33
3.4 Consort diagram FLAIR scans . . . 36
3.5 Overview of the prognostic AI monitor . . . 41
3.6 Including temporal information . . . 41
3.7 MRI histograms . . . 42
3.8 Visual effect of data harmonization . . . 43
3.9 Effect of data harmonization on validation loss . . . 43
3.10 Original fixed and moving image . . . 44
3.11 Transformed moving image . . . 44
3.12 Image overlays after registration . . . 45
3.13 Checkerboard image after image registration . . . 45
3.14 Associative analysis for computation of the risk score . . . 46
3.15 Associative analysis with confounders . . . 47
3.16 Associative analysis for primary brain cancer . . . 48
3.17 Kaplan Meier survival curves for all cancer types . . . 50

3.18 Kaplan Meier survival curves for primary brain tumors 51

4.1 T1-weighted image 57
4.2 T2-weighted image 57
4.3 FLAIR and postcontrast T1 image 58
4.4 Consort diagram multisequence scans 61
4.5 Dealing with missing sequences 63
4.6 Associative analysis for the risk score in multisequence scans . . 66
4.7 Associative analysis with confounders in multisequence scans . . 67
4.8 Kaplan Meier survival curves for multisequence images 69
4.9 Associative analysis for multisequence scans with primary brain tumors 70
4.10 Kaplan Meier survival curves for primary brain cancer 71
4.11 Cox regression for the external dataset 72
4.12 Kaplan Meier curves for external dataset 74

5.1 RANO measurements 81
5.2 Kaplan Meier curves for RANO assessments 84
5.3 Kaplan Meier curves for volumetric assessment 85
5.4 Kaplan Meier curves for PAM risk scores 86
5.5 Kaplan Meier curves for manual RANO assessments 87
5.6 Kaplan Meier curves for PAM risk scores 87

List of Tables

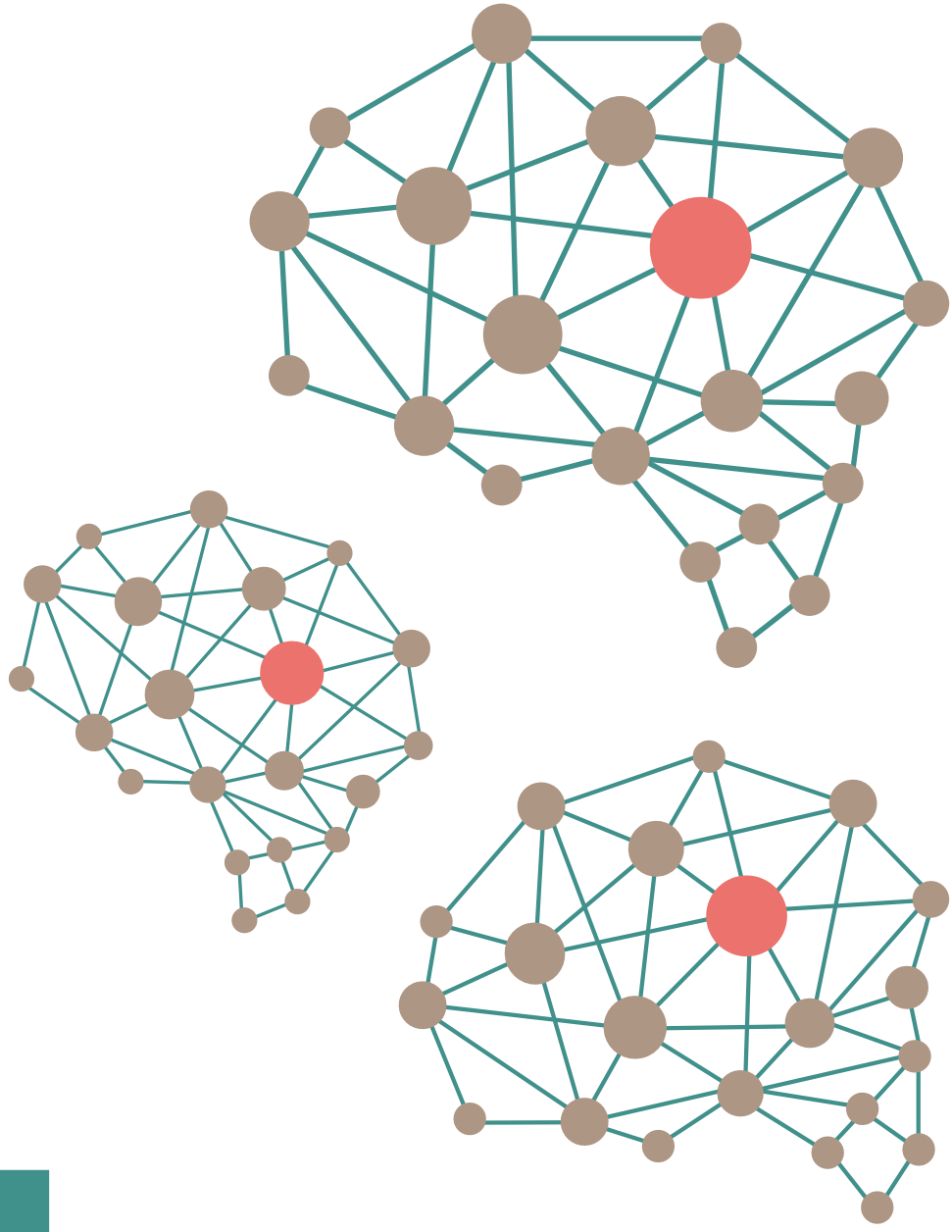
1.1 Current response criteria in brain imaging 7

3.1 Patient demographic FLAIR scans 37
3.2 Hazard ratios for computation of the risk score 47
3.3 Hazard ratios of PAM score and confounders 48
3.4 Hazard ratios of PAM score and confounders for primary brain cancer 49
3.5 C-indices for five computations of PAM risk score 49
3.6 C-indices subanalysis per cancer type 50
3.7 Logrank test for FLAIR pancancer dataset 51
3.8 Logrank test for the FLAIR primary brain cancer dataset 52

4.1 MRI image characteristics in RANO 60
4.2 Available scans in TCIA dataset 61
4.3 Patient demographic for NKI multisequence scans 62
4.4 Patient demographic for external validation set 62
4.5 Hazard ratios for the risk score in multisequence scans 66
4.6 Hazard ratios for confounders in multisequence scans 67
4.7 C-indices for the risk score for multisequence scans 68
4.8 C-indices subanalysis per cancer type 68
4.9 Logrank test for the risk groups in multisequence PAM 69
4.10 Hazard ratios for confounder analysis in multisequence scans with primary brain tumors 70
4.11 Logrank test for the risk groups in primary brain cancer 72
4.12 Hazard ratios for confounder analysis in the external dataset . . 73
4.13 C-indices for the external dataset 73
4.14 Logrank test for the external dataset 74

5.1 RANO high grade glioma 80
5.2 Logrank test for the response categories as determined by RANO assessment 84
5.3 Logrank test for the risk groups in the volumetric assessment . . 85
5.4 Logrank test for the risk groups in PAM 86

5.5 C-indices PAM and volumetric asssessment 86
5.6 Logrank test for the risk groups as determined by PAM for the data from Deventer Ziekenhuis 88



1

Chapter 1

General clinical background

1.1 Cancer and metastatic disease

Cancer is a disease caused by genetic mutation in cancer-susceptible genes. However, one genetic mutation will not immediately cause growth of cancer. For the progression of a tumor more mutations are required. These mutations, along with a favorable environment for its growth, are the reason a malignancy progresses and invades surrounding tissue. Two important classes of genes play a role in tumor progression: tumor-suppressor genes and oncogenes.[1] Further progression of a cancer may lead to metastatic disease. Hanahan and Weinberg have published the six hallmarks of cancer.[2] These hallmarks are distinctive and complementary concepts that enable tumor growth and metastatic dissemination. The six hallmarks are: sustaining proliferative signaling, evading growth suppressors, activating invasion and metastasis, enabling replicative immortality, inducing angiogenesis and resisting cell death. These hallmarks allow cancer cells to survive, proliferate and disseminate. Tumor types acquire these six functions through distinct mechanisms at various time points in the tumorgenesis. After the first publication in 2000 Hanahan and Weinberg have revisited their six hallmarks and added two emerging hallmarks: deregulating cellular energetics and avoiding immune destruction. They also added two enabling characteristics: genome instability and mutation and tumor-promoting inflammation. An overview of these new hallmarks can be found in figure 1.1.

The hallmarks, amongst other things, allow for dissemination of cancer cells to distant sites. It has long been assumed metastatic dissemination is one of the last steps in a multistep tumor progression process.[3] However, recent research shows dissemination of cancer cells from the primary tumor to distant sites occurs early in the progression of the disease.[2,4] Often this means before the discovery of the primary tumor. This is why patients often receive a systemic therapy. This is aimed to minimize residual disease. These systemic therapies can either

be provided adjuvant (after local removal of the tumor) or neoadjuvant (before local removal).[4]

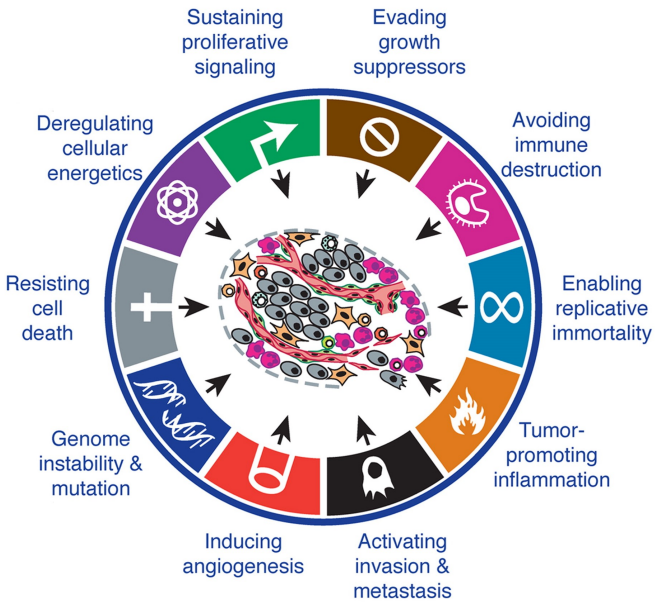


Figure 1.1: Overview of acquired capabilities necessary for tumor growth and progression. Adapted from Hanahan and Weinberg.[3]

1.2 Cancer therapies

There are various treatment options for cancer. The relevant options depend on the type and stage of a cancer. The most important distinction is the distinction between local and systemic treatments. Local treatments are used to locally treat the tumor (and possibly surrounding tissue). Examples of local therapies are surgery and radiation therapy. On the other hand, systemic treatment affects the whole body. Most drug therapies such as chemotherapy, immunotherapy or targeted therapy (including endocrine therapy) are therefore considered systemic treatments.

1.2.1 Local therapy

Surgical excision of a tumor can be a relatively simple method to treat patients with early stage solid tumors that are still confined to the anatomic site of origin.[5] However, due to dissemination of cancer cells as described in 1.1 a

neoadjuvant or adjuvant systemic therapy may be required.

Another option for local therapy is radiotherapy. Radiation may be administered as photons or particles (protons, neutrons or electrons). The interaction of photons with the tissue leads to ionizations. These ionizations either interact directly with subcellular structures (like the DNA) or they interact with water. After interaction with water free radicals are formed which in turn interact with the subcellular structures. Both pathways eventually lead to DNA damage. This DNA damage may have the following consequences: normal cell division, DNA damage-induced senescence, apoptosis or mitotic-linked cell death. In the end this leads to destruction of tumor cells.[5] In radiotherapy there is a trade-off between providing a maximal dosage of radiation to the tumor cells while limiting the radiation dose for surrounding healthy tissue. The two main types of radiation therapy are external beam radiation and brachytherapy (internal radiation).[6]

1.2.2 Systemic therapy

Chemotherapy is a systemic therapy which means it does not only target the primary tumor, but it is also able to target metastases. The goal of chemotherapy is to decrease the tumor burden. Traditional chemotherapeutics act by killing rapidly dividing cells through targetting DNA or processes critical for cell division. Non-traditional therapies target vulnerabilities specific to cancer cells.[7] The side effects of chemotherapy are toxicity to healthy tissues and the development of cellular resistance to chemotherapeutic agents. Chemotherapy may be given as the initial treatment of cancer. In this case it can either be induction chemotherapy or neoadjuvant chemotherapy. Induction chemotherapy is given to patients who present with advanced cancer and have no alternative treatment options in the hope other treatment options become available after shrinkage of the tumor. If there are no alternative treatment options and the prognosis is poor than palliative chemotherapy may be started to extend the life of a patient. Chemotherapy may also be delivered after local treatment (adjuvant). Adjuvant chemotherapy is used to tackle micrometastases as much as possible.[8]

Immunotherapy is capable of targeting tumor cells. Most studies on immunotherapy thus far have focused on enhancing the antitumor response of T cells. These T cells are able to recognize cancer antigens and can start an immune response. Another pathway is using human antibodies that recognize growth factors on the surface of tumor cells. The antibodies can interact with these growth factors leading to tumor regression. This is not caused by the

direct destruction of tumor cells, but by the interference with cell growth. Generally, immunotherapy either focuses on stimulating the patient's own immune system or on passively administering already activated immune cells.[5,9] Immune therapy is not without immune-related side effects, but in general these side effects are better tolerated than the side effects of chemotherapy.[9]

Targeted therapies may target or block any of the malfunctioning molecules of pathways that cause the development of cancer. It is a new generation of drugs that interfere with a specific molecular target which has a role in tumor growth or progression.[10] The first generation of targeted therapies was endocrine therapy. Endocrine therapy was proved to be effective in breast cancer patients expressing the oestrogen receptor.[11] In general, targeted therapies interfere with one of the specific hallmarks shown in figure 1.1. Targeted therapies can be divided in small-molecule inhibitors and monoclonal antibodies.[12] An example of small-molecule inhibitors is targeting mutant kinases. This has shown good response rates in chronic myeloid leukaemia, gastrointestinal tumors and lung cancer.[10] However, not all cancers show a mutant kinase which means this therapy can only be used for specific cancers. Another example of a targeted therapy is the targeting of the tumor microenvironment. This method is able to cut off blood supply to tumors by using antiangiogenic agents. This is potentially applicable to a lot more cancer types.[10]

1.3 Clinical decision making in oncology

Medical imaging is an important factor in the evaluation of cancer therapies. Imaging is necessary to properly evaluate treatment making it one of the cornerstones of current cancer care. It also plays an important role in screening, diagnosis and staging.[13] Magnetic resonance imaging (MRI) is commonly used for localized disease in the pelvis, abdomen and brain. It is safer than computed tomography (CT) as it does not use ionising radiation. This is especially an advantage for cancer patients who routinely undergo scans. MRI also provides finer details of the brain, spinal cord and vascular anatomy, because of the good soft-tissue contrast. That is why MRI is commonly used in diagnostic imaging of the brain. In brain MRI T1-weighted, T2-weighted and FLAIR images are often used. Using these sequences allows for distinction between white and gray matter and differentiation of cerebrospinal fluid (CSF). The FLAIR sequence specifically is used to distinguish between CSF and brain abnormalities.[14]

Clinical decision making in oncology is not solely based on medical images. Instead it is based on a combination of variables. This means it often requires

collaboration between various medical disciplines. In that case a multidisciplinary tumor board (MTB) is set up to discuss patients.[15] These MTB's lead to an improved quality of care for cancer patients.[16,17] Soukup et al. demonstrated the input of radiologists in MTB's significantly affects clinical decision making.[18] Radiologists use tumor response evaluation criteria to quantify a patient's response to treatment. Generally, the most important evaluation criteria are the Response Evaluation Criteria in Solid tumors (RECIST). Specifically for the brain the Response Assessment in Neuro-Oncology (RANO) criteria have been developed.

1.3.1 Response Evaluation Criteria in Solid tumors

Evaluation of tumor response plays an important role in treatment decision making. The World Health Organisation attempted to introduce a standardised treatment response evaluation. This method uses two dimensional measurements of target lesions and categorised responses in one of four categories. It was the basis for the Response Evaluation Criteria in Solid tumors (RECIST) that were developed in 2000.[19] In 2009 RECIST 1.1 was published[20], and in 2017 an adapted version for immunotherapy (iRECIST) was published[21]. RECIST is still the standard for response evaluation of solid tumors in the abdomen.

RECIST describes a standardized method to estimate the total tumor burden. To do so RECIST makes a distinction between target and non-target lesions. Target lesions are lesions with a maximal axial diameter of at least 10 mm, should allow for reproducible and repeated measurements and should be representative of all organs. A maximum of 5 target lesions can be selected of which there are maximal 2 target lesions per organ. The total tumor burden is determined by the sum of diameters of the target lesions and the minimal diameter of pathological lymph nodes. A pathological lymph node is defined as having a minimal diameter of at least 15 mm. Depending on the change of the sum of diameters between the baseline scan and the follow-up scan the response may be classified as complete response, partial response, progressive disease or stable disease. A similar workflow exists to use non-target lesions to classify tumor response as complete response, non-complete response/non-progressive disease or progressive disease, although no indication exists for these to be measured.[20]

1.3.2 Response Assessment in Neuro-Oncology Criteria

Historically the response evaluation of brain tumors was performed by either MacDonald Criteria or RECIST. However, both methods revealed shortcomings

in the evaluation of tumors in the central nervous system (CNS). That is why the Response Assessment in Neuro-Oncology (RANO) criteria have been developed. The main differences between these methods are described in table 1.1.[22] Some therapeutic agents are known to reduce contrast enhancement and therefore patients seem to be responding well. However, although these patients show a longer period of progression-free survival it does not change overall survival.[23] This is known as pseudoresponse. RANO criteria try to take this into account by repeating the MRI at least 4 weeks later to prove a durable response.[22] The RANO criteria for low grade gliomas are based on the percent change in T2/FLAIR images, new lesions, corticosteroids and clinical status. There are also RANO criteria for brain metastases. These criteria use elements of RECIST and the RANO criteria for high grade glioma. Lesions of 10 mm or larger are generally considered measurable in these RANO criteria. The criteria also give guidance for the number of target lesions, use of corticosteroids and pseudoprogression.[22]

Table 1.1: Summary of current response criteria in brain imaging[22]

Criterion	RECIST	MacDonald	RANO
Measurement	1 direction contrast enhancement	2 directions contrast enhancement	2 directions contrast enhancement + T2/FLAIR
Progression	$\geq$ 20% increase in sum of lesions	$\geq$ 25% increase in product of perpendicular diameter	$\geq$ 25% increase in product of perpendicular diameter
Response	$\geq$ 30% decrease in sum of lesions	$\geq$ 50% decrease in product of perpendicular diameter	$\geq$ 50% decrease in product of perpendicular diameter
Durability of response	Optional	Yes (at least 4 weeks)	Yes (at least 4 weeks)
Definition of measurability	Yes	No	Yes
Number of target lesions	Up to 5	None specified	Up to 5
T2/FLAIR	Not evaluated	Not evaluated	Evaluated
Corticosteroids considered	No	Yes	Yes
Clinical status considered	No	Yes	Yes
Pseudo-progression considered	No	No	Yes

1.3.3 Limitations

An important limitation of RECIST/RANO is the broad response categories. Tumor response is classified in one of the four earlier described categories. A more detailed and extensive description of changes in tumor sizes is likely to

provide additional prognostic information.[24] A decrease in the sum of diameters does not always mean a decrease in viable tumor cells and vice versa. For example, intratumoral necrosis and hemorrhage could induce a mismatch between the actual viable tumor cells and the sum of tumor diameters.[25] Also taking into account size-independent tumor morphology is likely to provide additional information. For hepatic metastases Chun et al. have already shown that a morphological response evaluation predicted survival better than a RECIST evaluation. They divided the hepatic metastases in three groups. Group 1 metastases had a homogeneous low attenuation and sharply defined tumor-liver interface. Group 3 lesions were heterogeneously attenuated and had thick and poorly defined tumor-liver interfaces. Group 2 were the lesions that did not fit in group 1 or 3.[26] Additionally, the limited number of target lesions does not always give an accurate response evaluation. These target lesions may be selected arbitrarily which has a significant impact on the response evaluation. Due to arbitrary selection different radiologists might select different target lesions.[27] Another important limitation of RECIST is the high variability (24%) in the sum of diameters.[28] Lesions that have unclear margins and complex shapes are difficult to measure.[27] This may cause the variability to exceed the 20% cut-off margin between SD and PD, resulting in potential misclassification of the response to treatment.[27,29]

A specific shortcoming for the RANO criteria is the quantification of nonenhancing disease progression. Since no quantifiable measure has currently been provided the usefulness of nonenhancing disease progression is limited. The nonenhancing disease progression is also difficult to assess in patients receiving immunotherapy. Immunotherapy often causes peritumoral edema which is difficult to differentiate from nonenhancing disease progression.[22]

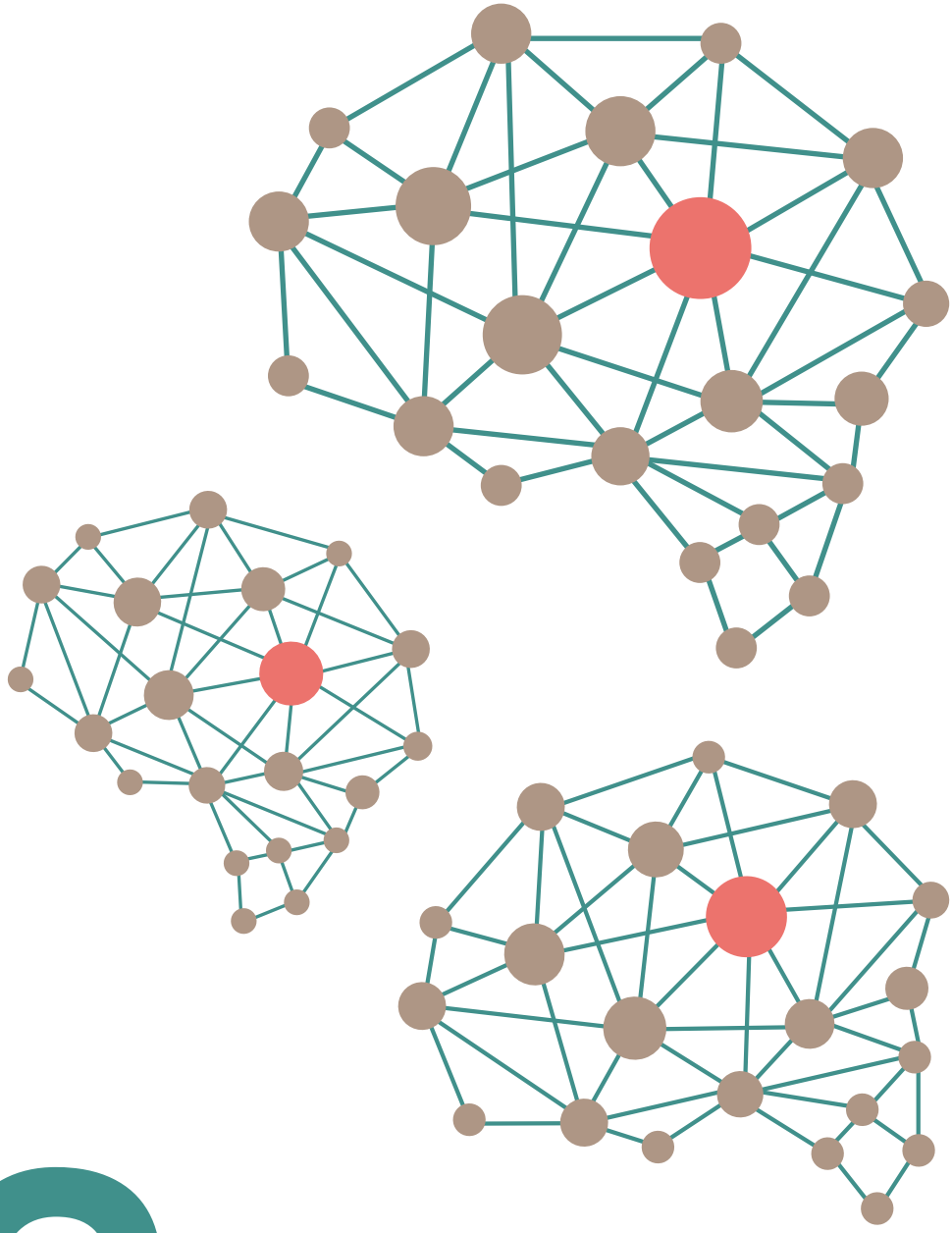
1.4 Proposed improvements

The limitations associated with current response evaluation ask for a new method to evaluate treatment response. This method should be reproducible, user-independent and take into account the complete tumor burden and possible tumor or treatment related side-effects. A promising tool that suits all the aforementioned requirements is the prognostic artificial intelligence (AI) monitor (PAM).[30]

PAM is a method to assess morphological changes between follow-up scans of the same patient. It works fully automatic and is independent of user input. PAM thereby overcomes most limitations of current methods for

radiological evaluation. On top of that PAM has already been compared to radiological assessment and blood values (hemoglobine, erythrocytes, leukocytes, and thrombocytes) in an earlier publication.[30] In univariate analysis PAM performed better than the radiological assessment and blood values. In multivariate analysis PAM remained statistically significant as well as radiological progression, leukocyte count and age. This demonstrates PAM is of additional value over current methods for prognostication.[30] Prognostication plays a key role in personal and clinical decisions in patients with advanced cancer.[31] Therefore, after further development PAM might play a significant role in clinical decision making in the future.

This research will focus on improving prognostication based on brain MRI. Important limitations of the current method are user-dependency, limited reproducibility and limited use of all information present in the scan. PAM overcomes these limitations and shows promising results. The ultimate goal is to provide the clinician with better information on an individual patient's prognosis, so individualized treatment decisions can be optimized. This research will therefore address the following question: How can the prognostic AI-monitor be adapted to MRI to be an accurate predictor of patient survival eventually leading to optimization of treatment decisions for the individual patient?



2

Chapter 2

General technological background

2.1 Magnetic Resonance Imaging

2.1.1 Signal acquisition

Magnetic Resonance Imaging (MRI) is based on the magnetic properties of the hydrogen nucleus. This magnetic property is known as *spin* as the nuclei spin around their axis.[32,33,34] This spin leads to interaction with electromagnetic fields. Normally, the spins of various hydrogen nuclei cancel each other out leaving a net magnetic vector of 0. However, when an external magnetic field (B_0) is generated the nuclei align in parallel or perpendicular to the magnetic field (figure 2.1). Alignment parallel to the magnetic field is a low-energy state and alignment perpendicular to the magnetic field is a high-energy state. The low energy state is preferred. Therefore the net magnetization (M_0) will be along the B_0 field in the direction of the z-axis.[32,35] The hydrogen nucleus is not statically aligned in the direction of B_0 as it has an angular momentum, meaning it will precess around the axis of the magnetic field.[32] The speed of the rotation is described by the Larmor frequency (ω_0) as in equation 2.1. The variable γ represents the gyromagnetic ratio, which is 42.6 MHz/T for the proton.[32,33,34,35]

$$\omega_0 = \gamma B_0 \tag{2.1}$$

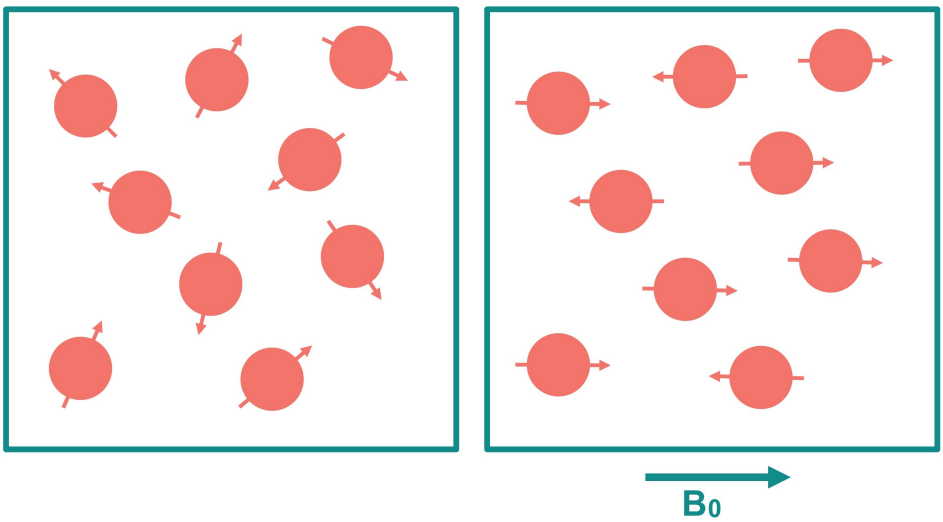


Figure 2.1: Orientation of hydrogen nuclei with spin. The left image shows a situation without an external magnetic field. The right image shows the orientation when an external magnetic field is generated. Adapted from van Geuns et al.[35]

A second radiofrequency (RF) magnetic field B_1 is applied perpendicular to B_0 at the resonance frequency. This excites the nuclei and allows for the absorption of energy which causes a transition from lower to higher energy levels. The opposite transition happens upon relaxation. The energy needed to induce transition depends on B_0 as described by equation 2.2.

$$\Delta E = \frac{\gamma h B_0}{2\pi} \tag{2.2}$$

The B_1 field is typically applied in short pulses. These pulses cause the net magnetization to move away from its alignment along the B_0 field. At this point in time the magnetization can be split in M_z and M_{xy}. The component of magnetization in the xy-plane (M_{xy}) generates a detectable signal. This process is visualized in figure 2.2. The signal is largest when a 90° pulse flips the net magnetization in the xy-plane.[34] When these pulses stop the system will return to the original equilibrium through relaxation. This process creates a voltage that can be detected by a suitable coil. This signal is amplified and displayed as the free-induction decay (FID). To improve the signal-to-noise ratio multiple FID's are obtained by multiple pulses. These various signals are then averaged. A Fourier transformation can be used to resolve the FID into an image or a frequency spectrum.[32,34]

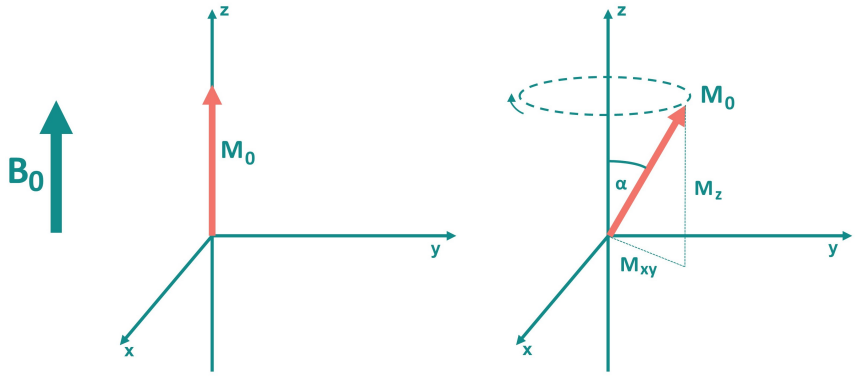


Figure 2.2: The left image shows the net magnetization at equilibrium. When an RF pulse is applied the magnetization makes an angle with the z-axis. This causes the signal to be divided in two components: M_{xy} and M_z. Adapted from Ridgway.[34]

A gradient may be used to distinguish from what position a specific signal is originating. By applying a gradient the magnetic field strength varies which leads to a different Larmor frequency. The variation in frequency measurements can then be used to select a slice and create a three-dimensional image reconstruction. This principle is visualized in figure 2.3.

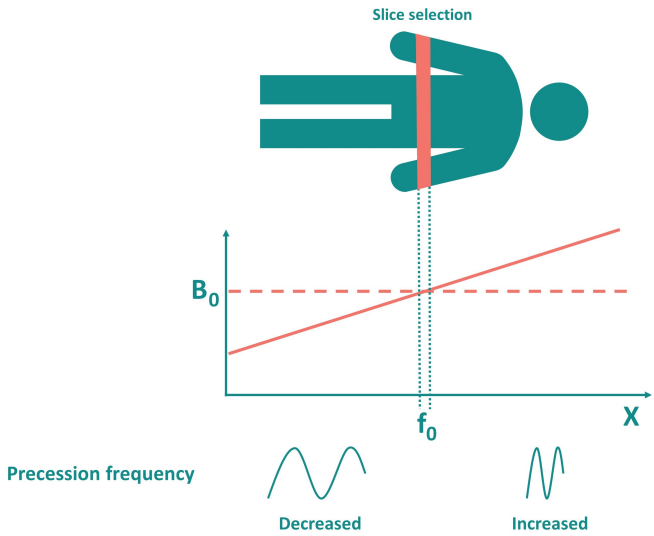


Figure 2.3: The magnetic field gradient is applied simultaneously to the excitation pulse. This causes the larmor frequency to vary with the magnetic field gradient which can be used for slice selection. Adapted from Grover et al.[32] and Ridgway.[34]

2.1.2 Signal characteristics

MR images are affected by tissue parameters such as T1, T2 and proton density. T1 and T2 are time constants that describe the longitudinal (z) and transverse (xy) relaxations, respectively. T1 is a time constant that describes the recovery of the magnetization vector to the original value at equilibrium (figure 2.4).

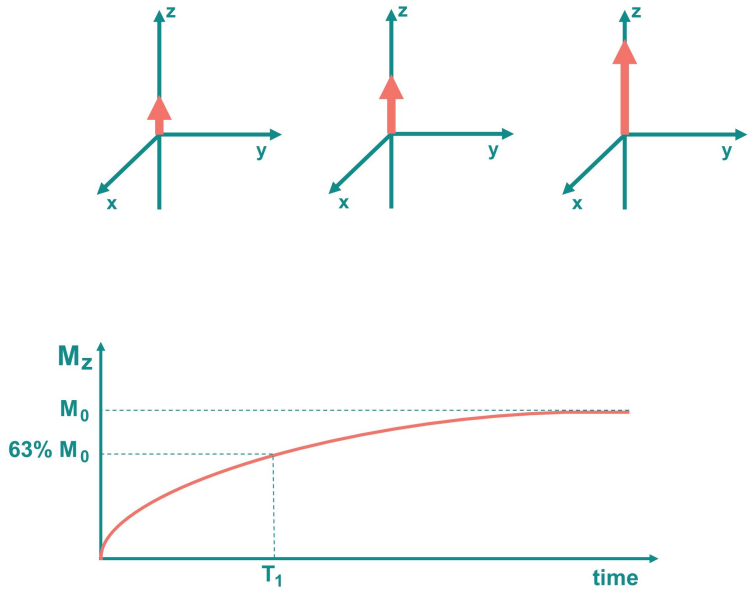


Figure 2.4: The process of T1 relaxation is visualised as the increasing magnetization in the z-axis. The graph allows for a different visualization of the same principle. T1 is the time constant at which 63% of M_0 has been recovered. Adapted from Ridgway.[34]

T2 relaxation describes the decay of the xy-component as the magnetization vector returns to its equilibrium. T2 relaxation is a faster process than T1 relaxation. The decay of the transverse signal is related to the phase of the spins. After an RF pulse the spins initially rotate in phase, but over time there will be loss of coherence and spins will move out of phase. This causes for the net magnetisation to decrease over time. The loss of phase coherence can be explained by interactions with other protons known as T2 relaxation. Neighbouring protons may slightly alter the magnetic field experience by a proton. This slightly changes the precession frequency and thus causes the spins to move out of phase. Another cause for spins moving out of phase is related to static magnetic field inhomogeneities. T2* describes the T2 relaxation as well as the magnetic field inhomogeneities (figure 2.5).[34]

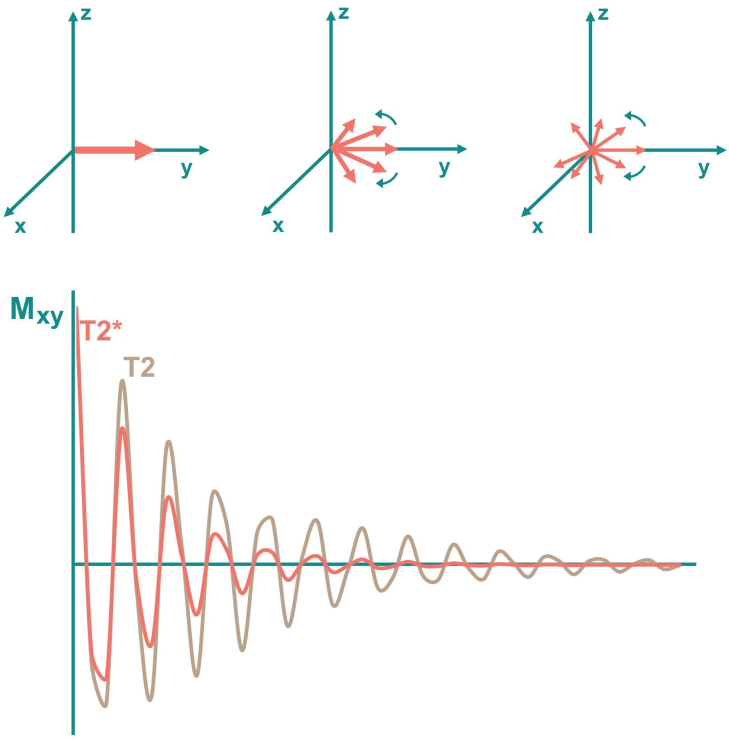


Figure 2.5: The process of T2 relaxation is visualised as the dephasing of spins in the xy-plane. The graph allows for a different visualization of the same principle. T2 is the time constant at which 37% of M_0 is left. Adapted from Ridgway.[34]

Next to T1, T2 and proton density MRI image contrast is affected by repetition time (TR) and echo time (TE). TR is defined as the time between an RF excitation pulse and the subsequent RF pulse. TE is the time between the RF pulse and the maximally detected echo. Varying these parameters will affect tissue contrast, because they change sensitivity to relaxation times. An example of this is using a relatively low TR which allows for better contrast between fat (T1 = 250 ms) and water (T1 = 4000 ms). When T1 is longer both fat and water have returned to equilibrium and contrast is no longer present. T2 relaxation is a faster process, so image contrast can be manipulated by TE in T2-weighted images.[33]

2.2 Artificial intelligence

As explained in chapter 1 PAM is a possibly promising tool to improve patient prognostication. PAM uses artificial intelligence (AI) for prognostication. AI is

an overarching term for anything related to devices that are able to perceive the environment and take adequate action to maximize the chance to reach a certain goal. AI is a self-explanatory term for the intelligence exhibited by machines. Within AI there is the field of machine learning (ML). This is defined as the ability of a computer to learn to complete tasks without being literally programmed to do so. Within machine learning there is the concept of deep learning. Deep learning is a particular group of ML algorithms based on artificial neural networks - resembling the biological counterpart. A specific type of neural network that can be used within deep learning is the convolutional neural network (CNN). PAM uses a convolutional neural network for image-to-image registration. The relationship between all aforementioned terms is explained in figure 2.6. All concepts will be explained in more detail in this chapter.

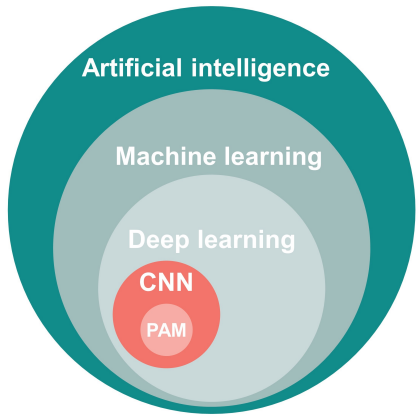


Figure 2.6: The relation between artificial intelligence, machine learning, deep learning, convolutional neural networks and the prognostic AI-monitor.

2.3 Machine learning

Machine learning is a field within AI. It is is defined as the ability of a computer to learn without being literally programmed to do so. Machine learning can be divided into supervised and unsupervised learning.[36] Supervised learning uses a labeled dataset for training. Each input is accompanied by the corresponding target variable or outcome. The algorithm learns from comparing its own output with the actual output. It then modifies internal parameters (weights) to reduce the error.[37] Unsupervised learning uses an unlabeled dataset. The algorithm aims to find a structure within the data.[36]

2.4 Deep learning

Within machine learning there is the concept of deep learning. Deep learning is based on artificial neural networks - resembling biological neural networks found in brains. A neural network consists of various neurons. These neurons are simple processing units that receive multiple inputs, and produce an activation. These inputs can either be from the environment or from previous neurons. A neural network is a network of interconnected neurons that activate each other through weighted activations. An overview of an artificial neuron is provided in figure 2.7. A network learns by adjusting the weights to approximate the desired output.[38]

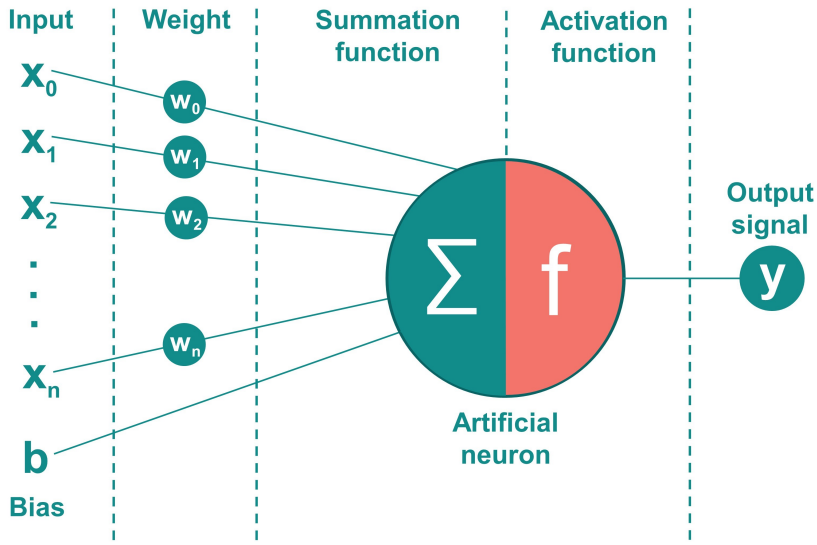


Figure 2.7: Schematic overview of a single processing unit (neuron) in a neural network. The input is shown as x_i, the weights as w_i, the bias as b and the output signal as y. Adapted from Sarker.[39]

Deep learning is essentially a composition of simple but non-linear modules (neurons) that transform the representation at one level (input) into a representation at a higher level (output). Depending on the number of these non-linear modules an increasingly more complex representation can be learned. In the case of image processing adding layers to the network amplifies those aspects of the image necessary to differentiate between different outcomes while suppressing the irrelevant details of the image. A deep neural network consists of multiple hidden layers. This is schematically depicted in figure 2.8. A typical feedforward network is composed by different functions eventually leading to

an output. For example, a network with three layers can be represented by functions $f^{(1)}$, $f^{(2)}$ and $f^{(3)}$ connected in a chain to form the output of the network $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$. The length of this chain is the depth of the network. In the training set a label $y \approx f^*(x)$ is provided for a given input x. While training $f(x)$ must match $f^*(x)$. The network uses the labels in the training set to learn what the output layer must do. The behavior of the individual layers in the network is not directly specified. The learning algorithm must lead to the most optimal use of each hidden layer, so the desired output is close to the ground truth label.[40]

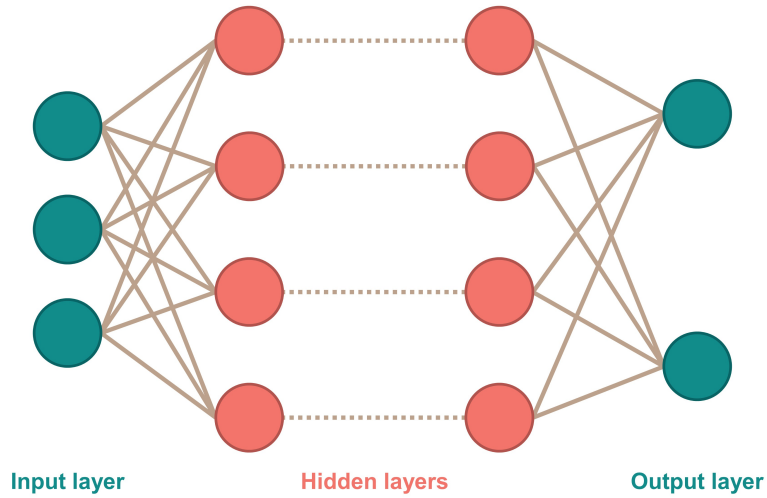


Figure 2.8: The general architecture of a deep neural network with more than one hidden layer.

2.5 Convolutional neural networks

Convolutional neural networks (CNNs) are in many ways similar to standard artificial neural networks, but are designed to deal with input on a regular grid. Feeding a neural network with grid-shaped data leads to substantially less input for neurons in the first layer. The standard artificial neural networks are not designed to process this kind of data. This would require too much computational power and may lead to overfitting. The architecture of CNNs is adjusted to receiving multilayered data as input. In case of 2D imaging, layers are comprised of neurons organised in the three dimensions (height, width and depth). There are various motivations for using a CNN. The first one is the use of sparse interactions. Unlike traditional neural networks CNNs are not

fully connected. This means that in a CNN not every output neuron interacts with every input neuron. CNNs achieve this by using a kernel size smaller than the input size. In the case of images or scans this means instead of using each individual pixel as input it uses small meaningful features in the images such as edges. This reduces memory requirements and increases computational efficiency. Another advantage is that CNNs can use the same parameter for more functions in the model. This concept is known as parameter sharing. The value of a weight at one input is tied to a weight applied elsewhere in the network. This further reduces the storage requirements of the model. In a CNN the particular form of parameter sharing leads to equivariance to translation of a given layer. This is mathematically represented as $f(g(x)) = g(f(x))$. In the case of an image it would mean shifting all pixels one to the right and subsequently applying the convolution would yield the same result as applying the convolution and subsequently shifting all pixels on to the right.[40] Amongst others these properties make CNNs very succesful in image detection, segmentation and object recognition.[37] There are various layers that can be used in a convolutional neural network.[37,41] A typical CNN is comprised of three kinds of layers: convolutional layers, pooling layers and fully-connected layers. Convolutional and pooling layers implement feature extraction. A fully connected layer maps the features into a final output.[42] An overview of a basic convolutional neural network is provided in figure 2.9.

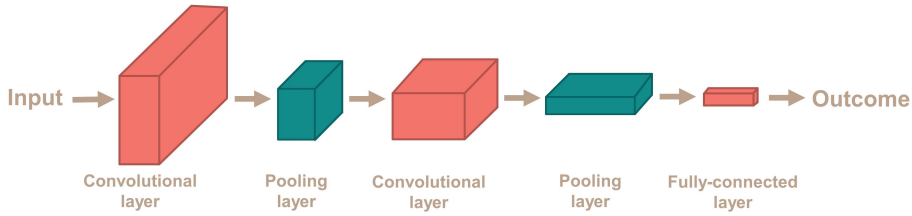


Figure 2.9: The general architecture of a convolutional neural networks with convolutional layers, pooling layers and fully-connected layers.

2.6 Prognostic AI monitoring

Prognostic AI monitoring (PAM) is a technique that uses CNNs to asses morphological changes by performing a deep-learning based image-to-image registration.[30] Image-to-image registration is essentially the process of establishing a voxel-wise match between two scans. PAM matches the anatomical landmarks and shapes of two scan volumes and thereby quantifies anatomical differences between scans.[43] These quantifications are extracted as

imaging feature vectors from the latent space of the network and fed into a classifier. This classifier is trained to predict survival. [30]

The input for PAM consists of a fixed and a moving image. Image registration is applied to the moving image to match the fixed image. The input is processed in two subsequent parts. Firstly, a CNN is used to determine the 12 parameters of affine image transform. The affine image transform corrects for patient position. The second part of the network is fed with the outcome of the first part: the fixed image and the affine warped moving image. The network processes these inputs to regress a displacement field. The displacement field contains a 3D vector for each voxel. This vector indicates the necessary displacement of a voxel in the moving image to match the anatomical position in the fixed image. This part of the network is used to assess the gross morphological changes between the moving and fixed image. [30]

One of the advantages of PAM is that it can be trained on unlabeled data. [30] Since the process of labeling data is labour intensive, using unlabeled data generally means using larger datasets.

2.7 Layers

There are various important layers for PAM which will be explained in this section. These layers are the convolutional layer, activation layer, pooling layer, fully-connected layer, group normalization layer and the last activation layer. The way these layers make up the final architecture of PAM as used in this research will be described in chapter 3.

2.7.1 Convolutional layer

In a convolutional layer a kernel is applied across an input. The input is an array of numbers called a tensor. The input is element-wise multiplied with the kernel and then summed. The result of this operation is put at the corresponding position in the output tensor. This process is visualized in figure 2.10. The output tensor is also called a feature map. In a convolutional layer the kernel contains the learnable parameters in a training process. There are three important hyperparameters in a convolutional layer: kernel size, stride, the number of kernels and zero-padding. Kernel size, stride and zero-padding determine the size of the output. The kernel size is usually 3×3, but can be different sizes as well. Stride is the number of nodes a kernel moves each time. In figure 2.10 the stride is 1. A disadvantage of using a convolutional layer may be loss of information, because information at the edges of a matrix is lost when

using a kernel. This can be resolved by zero-padding. This puts zeros at the edge of the matrix and therefore the output does not necessarily shrink anymore, but this also depends on the stride and kernel size. The influence of input size (N), stride (S), kernel size (K) and zero-padding (P) on the output size (O) is described in equation 2.3.[44] The number of kernels is directly correlated with the depth of the resulting feature maps.[42]

$$O = 1 + \frac{N + 2P - K}{S} \tag{2.3}$$

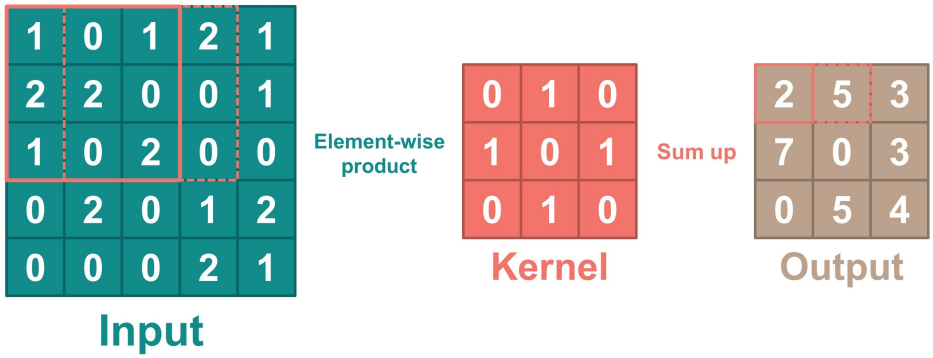


Figure 2.10: Visualization of the convolutional layer. The green matrix on the right is the input. The kernel in this example is size 3×3, the stride is 1 and there is no zero-padding. The kernel and corresponding part of the input image are element-wise multiplied and subsequently summed. This results in the beige output matrix of size 3×3.

2.7.2 Activation layer

The outputs of the convolutional layer are passed on to the next layer. Usually this is a non-linear activation layer. The most common non-linear activation function is the rectified linear unit (ReLU) which computes the function in equation 2.4.[42] The derivative of the ReLU function equals 1 if x is more than 0, otherwise it equals 0 as expressed in equation 2.5. Thus, the ReLU function will output the input as long as the input consists of only positive numbers. A negative number in the input will be converted to 0 in the output. This is also visualized in figure 2.11.

$$ReLU(x) = max(0, x) \tag{2.4}$$

$$\frac{d}{dx}(x) = \{\, 1\, if\ x > 0;\ 0\, otherwise\} \tag{2.5}$$

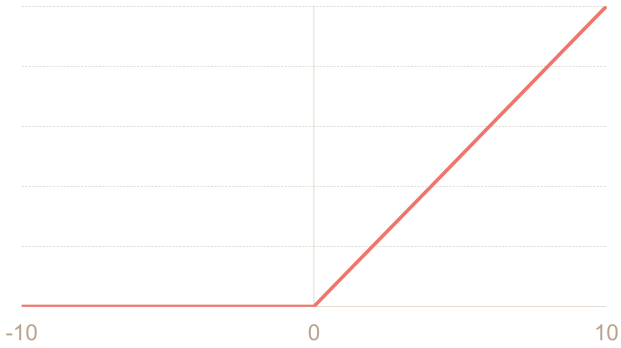


Figure 2.11: Visualization of the rectified linear unit function (ReLU). Negative values are converted to 0, positive values are not converted.

2.7.3 Pooling layer

A pooling layer is aimed at downsampling the input. Downsampling introduces invariance to small changes. There are no learnable parameters in the pooling layer. The hyperparameters in a pooling layer are stride, padding and filter size. Unlike in a convolutional layer the kernel in a pooling layer does not contain any learnable parameters. There are two popular variants of the pooling layer: max pooling and global average pooling. Max pooling filters sub-regions of the input feature maps and returns the maximum value (figure 2.12). A common filter size in max pooling is 2×2 with a stride of 2. Max pooling changes the height and width of a feature map, the depth remains unchanged.[42,44] Global average pooling performs downsampling to a 1×1 array. It does so by calculating the average of all elements in the feature map. Its advantages are a reduction of learnable parameters and it enables the network to process inputs of variable size.[42]

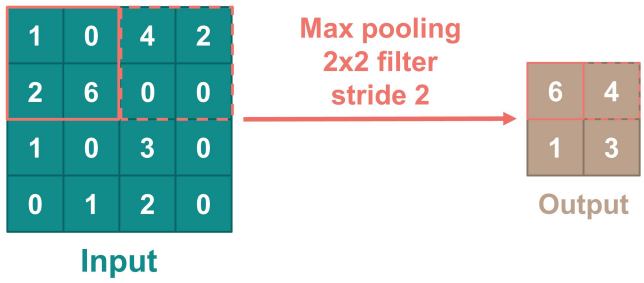


Figure 2.12: Visualization of max pooling layer. A filter of size 2x2 with stride 2 computes the maximum of sub-regions of the input. This leads to downsampling of the input.

2.7.4 Fully-connected layer

As discussed before in section 2.5 a CNN consists out of multiple convolutional and pooling layers followed by a fully-connected layer. The output feature map of the last convolutional and pooling layers is often transformed into a 1 dimensional array. This array is connected to one or more fully-connected layers which are also known as dense layers. These layers map the output of the last convolutional layer to final outputs.[42] The term fully-connected is self-explanatory: a neuron in one layer is connected to all neurons in the previous and next layer. This design is similar to the design of a standard artificial neural network as in figure 2.8. The fully-connected layer contains a lot of parameters and is therefore computationally expensive.[44]

2.7.5 Group normalization layer

The group normalization layer was developed by Wu and He.[45] Usually batch normalization (BN) is used. Features are normalized by mean and variance of a batch. However, BN only works with larger batch sizes. Group normalization (GN) on the other hand divides channels into groups and normalizes within these groups. GN is very stable over a range of batch sizes, while BN shows an increasing error for smaller batch sizes. Feature normalization in general is mathematically expressed as:

$$\hat{x}_i = \frac{1}{\sigma_i}(x_i - \mu_i) \tag{2.6}$$

In equation 2.6 x is the feature computed by a layer, i is the index, μ is the mean and σ the standard deviation. Generally there are three axes. N is the batch axis, C is the channel axis and H and W are the spatial dimensions (height and width for a 2D image). The difference between BN and GN is the set that is used for normalization. BN normalizes along the batch dimension. However, GN does not exploit the batch dimension. This is visualized in figure 2.13.

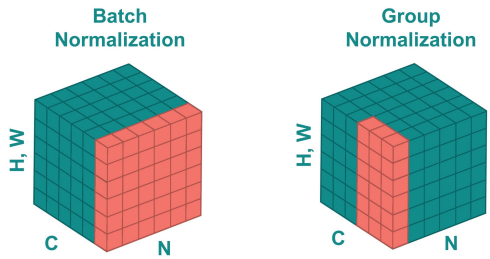


Figure 2.13: Feature map tensors with N the batch axis, C the channel axis and H, W the spatial dimensions (height and width). The group of pixels in red are normalized with the given method. Adapted from Wu and He.[45]

2.7.6 Last activation layer

Usually, the activation layer after the last fully-connected layer is different from other activation layers used in the CNN. For example, for multi-class classification tasks a softmax function is commonly used. The softmax functions normalizes all values from the last fully connected layers and thereby converts them to class probabilities which are all between 0 and 1 and sum to 1.

2.8 Loss function

The loss function measures the error between the predicted outcome and ground truth. It is used to update the weights in PAM through backpropagation. The loss function needs to be carefully chosen depending on the task at hand. Examples are using a cross-entropy loss for multi-class classification problems while mean squared error loss is more common for regression tasks.[42] In case of PAM the loss function must be able to measure the (dis)similarity between two images. A correlation coefficient loss is suitable for this purpose. In addition regularization losses are used to prevent overfitting or unrealistic deformations in the resulting deformation field.[46] The exact function used to compute the loss in PAM in this thesis will be further explored in section 3.2.1.

2.9 Training

The trainable parameters in a CNN are the kernels in the convolutional layer and the weights in the fully-connected layers. During training PAM uses the loss function to compute an error to quantify the deviation between the predicted and expected output. The loss function is used to adjust the internal parameters of the network, the weights and biases. Gradients are commonly used for this purpose.[42] In multilayer deep learning networks, such as PAM, gradients can be computed using backpropagation. Backpropagation is based on the chain rule for derivatives. The derivative can be computed by working backwards from the output of a module to the input of a module. By using backpropagation repeatedly derivatives (gradients) for all neurons can be computed.[37] The principle of back propagation can be explained by four equations.[47] Backpropagation goes from the output layer back to the hidden layers hence the first equation must describe the error (δ) in the output layer (L).

$$\delta_j^L = \frac{\partial C}{\partial a_j^L} \sigma'(z_j^L) \tag{2.7}$$

Equation 2.8 shows the error for neuron j in the output layer L. The first term $\frac{\partial C}{\partial a_j^L}$ describes how fast the cost function C would change as a consequence of the output activation a. The second term $\sigma'(z_j^L)$ provides a measure for how fast the activation function (σ) changes at weighted input z_j^L. For back propagation we use the matrix-based form of equation 2.8 which is:

$$\delta^L = \nabla_a C \odot \sigma'(z_j^L) \tag{2.8}$$

The second equation uses the error in the next layer (δ^{l+1}) to compute the error in the previous layer (δ^l):

$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l) \tag{2.9}$$

In this equation $(w^{l+1})^T$ represents the weight matrix for the $l + 1^{th}$ layer in transposed form. This equation essentially uses the information from the $l+1^{th}$ layer to compute the error in the l^{th} layer. It therefore can be used to compute the error in any layer in the network. After using equations 2.8 and 2.9 all that is needed is to use the computed errors to quantify the rate of change of the cost with respect to the weights (w) and the biases (b) in the network. For the bias this means that for a given neuron the following equation applies:

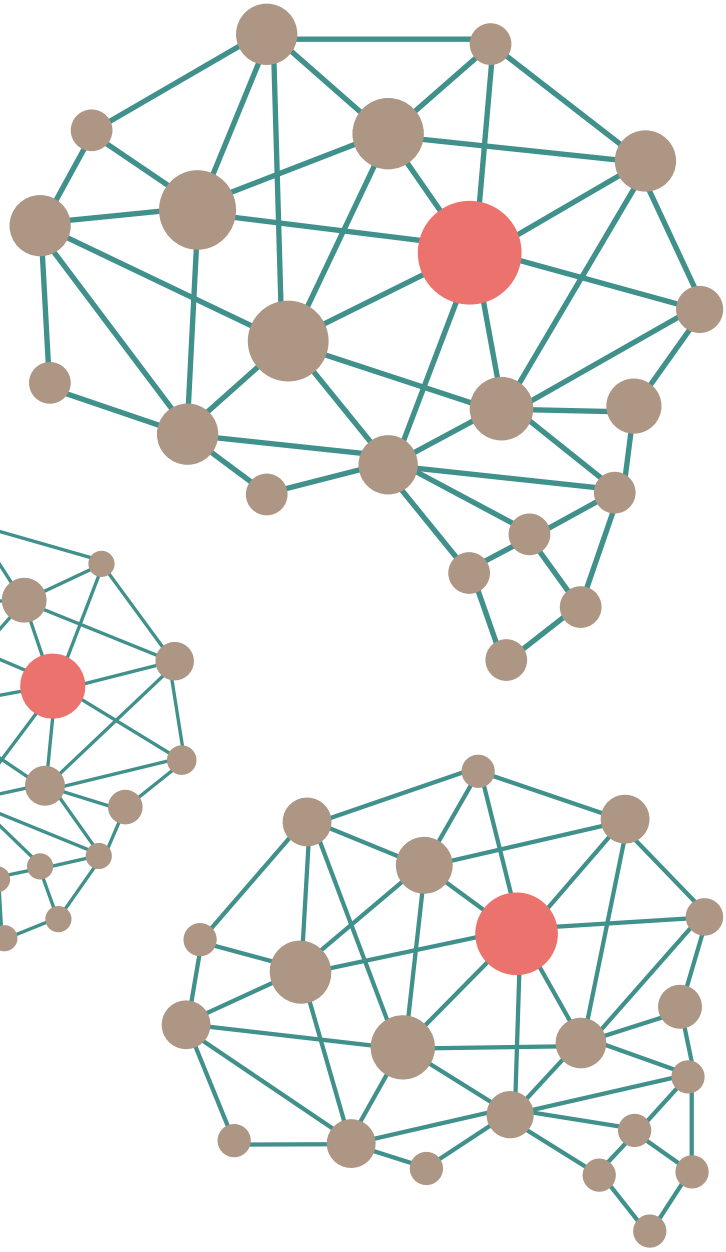
$$\frac{\partial C}{\partial b} = \delta \tag{2.10}$$

The rate of change of the cost with respect to the weight is slightly more complicated, but for a given neuron it is:

$$\frac{\partial C}{\partial w} = a_{in}\delta_{out} \tag{2.11}$$

To be able to properly train and evaluate the results the available data should be split into a training, validation and test set. The training set is used to train the network as described before. For every scan pair in the training set PAM computes the loss and adjusts the weights and biases accordingly. The total training time depends on the number of epochs. An epoch is defined as the moment at which every scan pair in the training set has passed PAM once. After each epoch PAM processes the scan pairs in the validation set to evaluate its current performance. It cannot learn from the scans in the validation set, meaning the internal parameters of the network are not adjusted during validation. The validation set is used to reduce overfitting of the network to the training set. After training for the preset number of epochs is completed the weights and biases with the best performance on the validation set are

selected. This introduces a bias and therefore the performance on the validation set cannot be used as a general performance metric. To be able to give a general performance the test set is used. This is a part of the dataset that is kept separate until the end of the training process. It can then be used for the final evaluation of the model and to test if the model is sufficiently able to generalize.[42]



3

Chapter 3

Extension of PAM to single sequence brain imaging

3.1 Introduction

Brain tumors are among the deadliest cancers. The most aggressive type is glioblastoma as more than two-third of patients will die within two years after diagnosis.[48] Brain tumors can also be metastases from other primary cancers. It is estimated at least 20% of cancer patients develop brain metastases at some point.[49] Most commonly the primary cancers are lung cancer, breast cancer or melanoma. Brain metastases have a poor prognosis with a two year survival rate of less than 8.1%.[50] If treated, therapy is usually multimodal consisting of a combination of surgery, radiotherapy, chemotherapy, immunotherapy and/or targeted therapies.

Both primary brain tumors and brain metastases are difficult to treat. This has multiple reasons. First, tumors infiltrating the brain are not always physically reachable for surgery.[48] On a smaller scale the central nervous system (CNS) is protected by two main barriers: the blood brain barrier (BBB) and the blood-cerebrospinal fluid (CSF) barrier. These barriers impede or filter the transit of molecules from the blood to the brain and may decrease efficacy of systemic treatment. The BBB is formed by endothelial cells, the endothelial and parenchymal basement membrane, pericytes and astrocytes. A schematic overview is provided in figure 3.1. The blood-CSF barrier is formed by choroid plexus epithelial cells. These cells are connected by tight junctions.[49] Another reason systemic treatment proves to be difficult is due to unique features of brain tumors, such as genetic features and the tumor micro-environment (TME). The genetic profile of brain metastases might differ from the primary tumor due to mutations. This introduces an uncertainty in treatment efficacy even if the treatment reaches the brain tumor.[51] The TME is composed of the cancer cells, various stromal cells, lymphatic and blood vessels and

the extracellular matrix. Abnormalities in these components could cause an unfavourable micro-environment allowing for tumor progression and treatment resistance.[52]

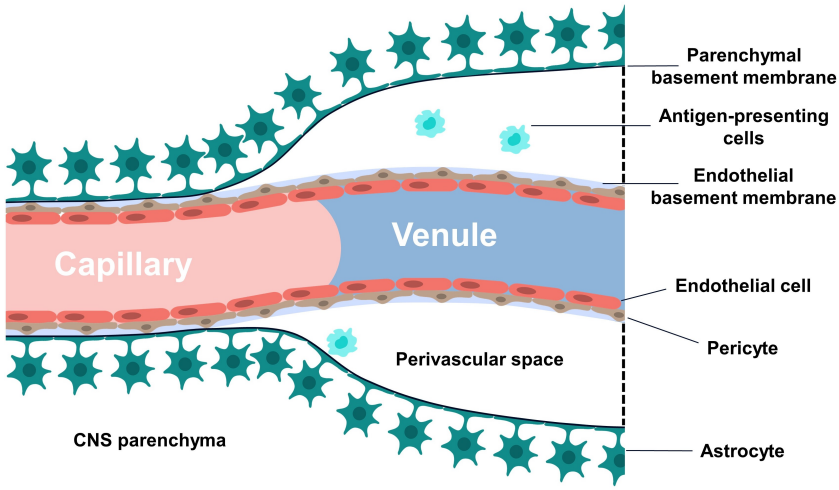


Figure 3.1: The blood brain barrier is formed by endothelial cells with low transcytosis rates and high expression of efflux pumps connected by tight junctions. The endothelial and parenchymal basement membrane, pericytes and astrocytes also play a role in the barrier function. Adapted from Achrol et al.[49]

Treatment of brain tumors with surgery, radiotherapy and/or chemotherapy often has adverse effects.[48] Surgical complications could for example be pain, infection, bowel or bladder complications, thromboembolisms or epilepsy. Radiation may cause encephalopathy, ataxia, diplopia, dysarthria, nystagmus, dementia, incontinence, a decline in cognitive function and cerebral vasculopathy. Adverse effects of chemotherapy may include fatigue, alopecia, constipation, headache, peripheral neuropathy and cardiac side effects.[53]

Treatment status is usually one of the factors included for prognostication in patients with brain metastases. Other factors include patient characteristics, extent of primary disease, performance status and the subtype of brain metastases. Some data driven prognostication tools have been developed based on these factors. Occasionally, histologic information is included as well in prognostication models. The clinical value of current prognostication methods is very limited while neurological disease is the cause of death in 52% of patients with brain metastases.[49] This emphasizes the need for an accurate

prognostication tool.

An important part of prognostication is monitoring treatment response with imaging. In the case of brain tumors MRI is the most sensitive modality. The tumor is monitored by changes in enhancement on specific MRI sequences. However, there are multiple factors that could lead to enhancement not caused by tumor progression. Therefore a distinction must be made between pseudoprogression, radiation necrosis and actual tumor progression or recurrence. Pseudoprogression is when new contrast enhancement is detected in MR images, but biopsy shows no progression of the primary brain tumor. This is often associated with chemotherapy, immunotherapy and radiotherapy. Pseudoprogression usually occurs within 3-4 months of radiation. Incidences of pseudoprogression reported in the literature vary widely, from 9 to 30%.[54] Radionecrosis on the other hand usually occurs 3-12 months after radiotherapy. This is a local tissue reaction to the radiation treatment leading to edema and disruption of the BBB.[55] Radiation necrosis is also reported in patients with brain metastases treated with stereotactic radiosurgery.[54] Tumor progression is quantified by RANO criteria as described in section 1.3.2. However, the RANO criteria define pseudoprogression as contrast enhancement which subsides without any intervention.[56] This means pseudoprogression can only be incorporated retrospectively. This is a challenge that will remain with any newly developed prognostication method.

A more accurate prognostication tool is desirable as it will impact the treatment decision. Treatment decisions in first instance are mostly influenced by the spread of the cancer. Depending on the spread of cancer the treatment will have a curative intent or palliative intent - aiming at curing the disease or extending/improving quality of life, respectively.[57] A more accurate and early prediction of progressive disease would allow for an early switch to the next line of treatment. This spares patients unnecessary treatment related toxicity and psychological burden.[58] As mentioned in section 1.4 prognostic AI monitoring (PAM) is a promising prognostication tool. However, current research with PAM has focused on anatomical CT imaging.[30,43] One of the reasons for this is the technical challenges associated with MR imaging.

3.2 Technological background

3.2.1 Prognostic AI monitoring

As described in section 2.6 PAM is based on the framework of image-to-image registration that uses a fixed and a moving image as input. The moving image is warped to match the landmarks of the fixed image. The architecture of PAM consists of two parts. The first part regresses the 12 parameters of an affine transform, aimed to correct for different head positions. The second part is focused at the elastic deformation, aimed to model longitudinal changes. The same network is used as in the publication by Trebeschi et al.[43] The goal of PAM is not to perform image registration, instead it aims to represent longitudinal changes with the features in the latent space. These features are used for survival prediction and the registration itself is not used.

Network architecture

The part focused at the affine transform is a deep convolutional neural network. Its architecture is a VGG-like CNN. A VGG is a very deep CNN with multiple subsequent convolutional layers, typically with small 3×3 kernels. VGG was developed and first used by Simonyan and Zisserman for large-scale image recognition.[59] In PAM five consecutive convolutional blocks are used. Each convolutional block consists of a convolutional layer, a group normalization layer and a ReLU activation layer. The convolutional layer has a kernel size of 3×3, the number of kernels is 8 and the stride is 2. After the last convolutional block a global average pooling layer is added followed by a fully connected layer and ReLU activation layer. Ultimately leading to the 12 parameters necessary for affine transform, which is applied to the moving image via a spatial transformation layer. The architecture is visualized in figure 3.2.

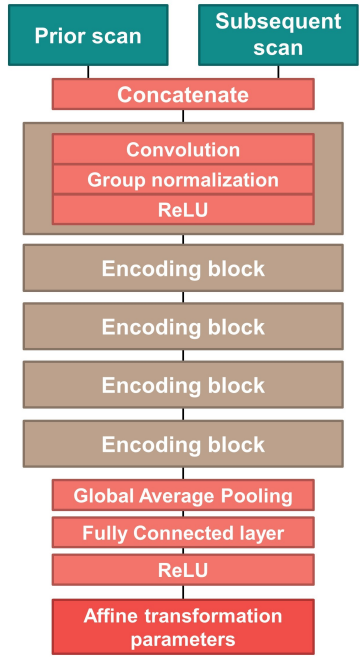


Figure 3.2: Network architecture for affine transformation. Both scans are concatenated and serve as input. There are five consecutive encoding blocks followed by a global average pooling, fully connected and ReLU layer resulting in the parameters for affine transformation.

The second part focuses on the elastic transform. It receives the warped affine image as input, as well as the original fixed image. It follows the U-Net architecture. U-Net is a CNN architecture originally designed for fast and accurate image segmentation.[60] The U-Net architecture can be described in two parts: the contracting or encoder path and the expanding or decoder path. Skip connections are present to connect the feature maps of the encoder with the decoder. These skip connections concatenate the feature maps resulting from the encoder to the corresponding decoder layer.[60] For PAM there are six encoding blocks and six decoding blocks with skip connections between the two. The encoder blocks downsample the image by using a stride of 2. After the fourth block there is a convolutional latent space with a stride of 1. This is followed by the decoder blocks which upsamples the images again via stride. In the encoder layer each convolutional blocks consists of a convolutional layer, group normalization layer and ReLU activation function. In the decoder a deconvolutional block consists of a concatenation (via the skip connection), followed by a deconvolution, group normalization and ReLU layer. A visualization of the network structure is provided in figure 3.3.

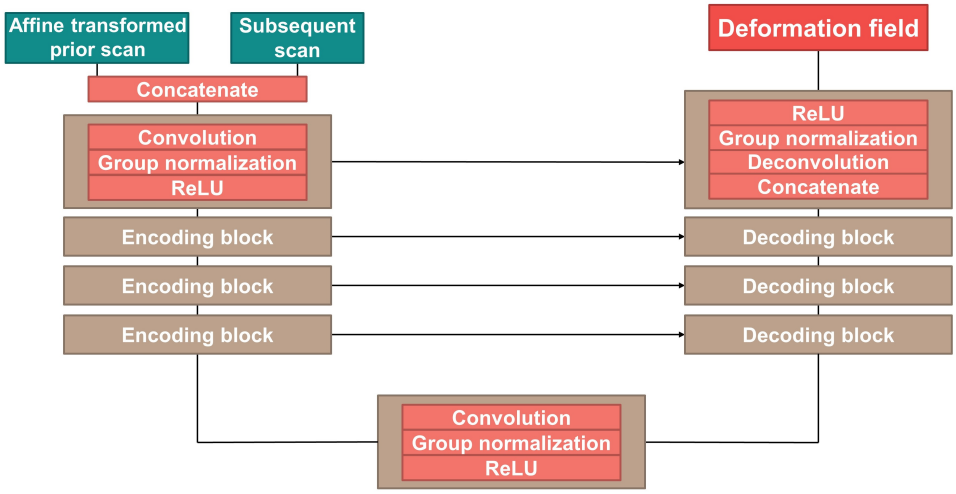


Figure 3.3: Network architecture for elastic transformation. This diagram shows the U-Net architecture with encoding blocks and decoding blocks. Each encoding block consists of a convolution, group normalization and ReLU layer. Each decoding block contains a concatenation with the encoding block via the skip connections, deconvolution, group normalization and ReLU layer.

Loss function

The loss function used in PAM is based on the correlation coefficient loss. The correlation coefficient is calculated by using the covariance between two image volumes (V_1 and V_2). The covariance is calculated as:

$$Cov[V_1, V_2] = \frac{1}{|\Omega|} \sum_{x \in \Omega} V_1(x) V_2(x) - \frac{1}{|\Omega|^2} \sum_{x \in \Omega} V_1(x) \sum_{y \in \Omega} V_2(y) \tag{3.1}$$

In equation 3.1 Ω represents the grid on which the image volumes are defined. The correlation coefficient is defined as:

$$CorrCoef[V_1, V_2] = \frac{Cov[V_1, V_2]}{\sqrt{Cov[V_1, V_1] Cov[V_2, V_2]}} \tag{3.2}$$

The correlation coefficient basically measures the similarity between the two scans. It ranges between -1 and 1. It results in 1 if there is a direct linear correlation between the two scans. The correlation coefficient loss is:

$$L_{Corr}(V_1, V_2) = 1 - CorrCoef[V_1, V_2] \tag{3.3}$$

Next to the correlation loss, a total of three penalties are applied. These are similar to the penalties used by Zhao et al.[46] First, the total variation loss is incorporated. This is a penalty on the elastic deformation. It discourages discontinuity of the deformation field by penalizing large differences, resulting in a smooth deformation:

$$L_{TV} = \frac{1}{3|\Omega|} \sum_{x} \sum_{i=1}^{3} (f(x + e_i) - f(x))^2 \tag{3.4}$$

The orthogonality loss is added as a penalty on the affine transform of the image. This penalizes an excessively non-rigid transform. When I is the identity matrix and A is the affine transform matrix let $\lambda_{1,2,3}$ be the singular values of $I + A$. The orthogonality loss is:

$$L_{ortho} = -6 + \sum_{i=1}^{3} (\lambda_i^2 + \lambda_i^{-2}) \tag{3.5}$$

Lastly, the determinant loss is incorporated as a penalty on the affine transform. This penalizes an affine transform that involves reflection. It is defined as:

$$L_{det} = (-1 + det(A + I))^2 \tag{3.6}$$

The penalties were weighted as 1/10 on the affine loss and 1/1000 on the deformable loss. The lower penalty on deformable loss gave the model more room to adapt to valid anatomical changes.

3.2.2 MRI data harmonization

There are a couple of technical challenges associated with MRI that need to be dealt with to ensure data harmonization. First of all, MRI has no standard image intensity scale like the Hounsfield units in CT. Furthermore, MR imaging usually shows more artifacts such as field inhomogeneity, motion and scanner-specific variations. There are three aspects of harmonization of MR imaging that definitely need to be considered: denoising, bias field correction and standardisation.

Denoising

Denoising is usually incorporated into modern MR scanners. However, noise or artifacts due to respiratory or body motion may also have a significant influence on the appearance of MR images. These issues should be addressed before using MR data for image analysis.[61]

Bias field correction

Bias field is a non-uniformity that is present in MR data caused by inhomogeneity in the magnetic field of the MRI. Bias field is a low frequency signal that reduces the high frequency components of the image. Thus it essentially causes a blur in the MR images.[62] It is possible to remove this bias field with various methods.

An option to remove the bias field is the N3 method. This method is known for its robustness. The N3 method estimates the bias field by sharpening the image histogram using a Guasssian deconvolution. The resulting bias field estimation is smoothed by using a B-spline.[63] An improvement of N3 called N4ITK has been proposed. N4ITK replaces the B-spline smoothing that is used in N3 by a better performing B-spline approximator and it uses a modified optimization scheme.[64] In current practice N4ITK is most commonly used.[61] N4ITK was also used for a deep learning based tumor grading model in soft tissue sarcoma patients[65] and brain tumor patients[66].

There is also the possibility of not performing a bias field correction. A deep learning based algorithm to segment breast MR images shows an U-Net based method was not as much affected by the bias field.[67]

Standardisation

As mentioned, there is no standard image intensity scale in MR images. That means image standardisation must be performed to give the same regions of similar MR images the same intensity. First, the intensity values should be shifted so they are all positive numbers. Subsequently the probability distribution function of the image can be used to truncate the upper and lower narrow tales of the histogram. After these steps the images can be standardised using a two-step Nyul pre-processing method. The Nyul method is focused on mapping the foreground using a certain number of landmarks.[61] The advantage of the Nyul method is the possible application to multimodal data (such as MRI) by using a multidimensional histogram.[63] This method was also used in a deep learning based brain tumor segmentation in MRI images.[66]

3.3 Methods

3.3.1 Dataset

The base registration network for brain MR images is trained with a large dataset from The Cancer Imaging Archive (TCIA). All available datasets that could contain brain MR scans were extracted from the TCIA, resulting in a dataset

with 44903 scans. This dataset was contaminated with for example scans from other organs and diffusion MR images. To clean up the dataset, and ensure that all the scans were brain scans, a neural network (developed at the NKI) predicted the probability of a scan being a precontrast T1, postcontrast T1, T2 or FLAIR sequence. Only scans with a probability of at least 0.99 for the FLAIR sequence were incorporated in this first dataset. To further clean up the dataset scan names containing one of the following keywords were eliminated: *mask*, *prostate*, *breast*, *sarcoma*, *dwi*, *diffusion*, *dit*, *cor*, *sag*, *memprage* and *mprage*. Furthermore all scout scans were removed from the dataset by excluding the scans with only three slices. After the network has been trained with the TCIA images it will be used on a dataset from the NKI. The TCIA dataset eventually consisted of 546 FLAIR scans to train the prognostic AI-monitor on. Further analysis was performed with scans from the NKI which were acquired following the consort diagram in figure 3.4 resulting in 1942 scan pairs of 571 patients. The patient demographic is visualized in table 3.1.

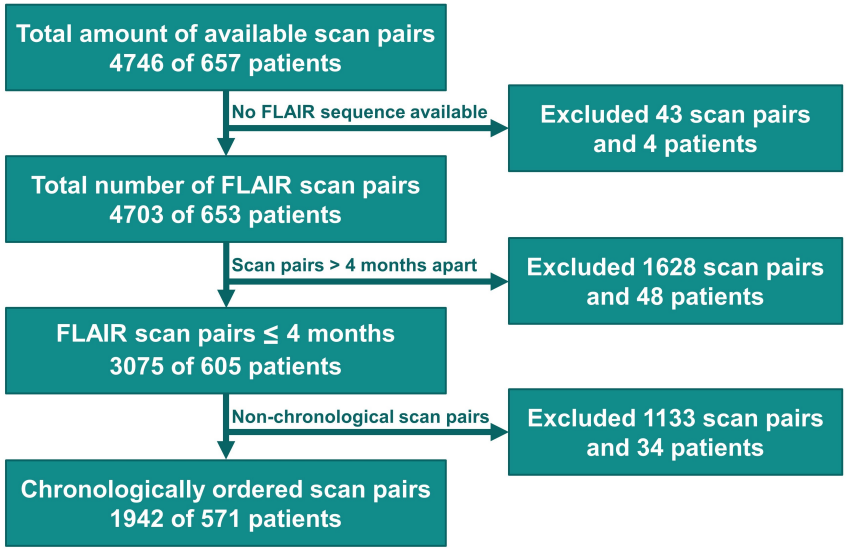


Figure 3.4: Consort diagram of patient and scan selection for analysis of the prognostic AI monitor

Table 3.1: Patient demographic of patients with a FLAIR scan with the following patient characteristics: age, gender, survival and diagnosis.

Characteristic	N = 571	Percentage (%)
Age		
Median	59	
Range	17 - 96	
Gender		
Female	277	49
Male	294	51
1 year survival		
Alive	325	57
Deceased	246	43
Diagnosis		
Bronchus and lung	168	29
Melanoma	151	26
Brain	106	19
Other	146	26

3.3.2 Data harmonization

Data harmonization is essential to ensure inter-scanner and inter-site reliability of the network. The TCIA dataset is a very diverse dataset with MR images acquired at different institutions and with different scanners. To ensure the network is able to generalize over these scans a few pre-processing steps are incorporated: resample and reorient the images, transform images to a uniform scan size, bias field correction, intensity normalization and normalization between 0 and 1.

Resample and reorient scans

The first step is to resample all images to voxels of $1 \times 1 \times 1$ mm^3. Resampling is achieved by linear interpolation between voxels. Reorientation is specifically useful for MRI's, as it is common to do image acquisition and reconstruction in the axial, coronal and sagittal planes. This code reorients all scans to the axial plane.

Uniform scan size

All MR scans are padded and/or cropped to a uniform size. In this research that is $256 \times 256 \times 256$ voxels. This choice was made because this size fits the average brain scan well and 256 allows for easy computing. If the initial scan is

larger than the desired size the scan will be cropped to the desired size. If the initial scan is smaller than the desired size is the scan will be padded with 0's until the desired size is reached.

Bias field correction

The bias field can be removed by applying the N4ITK method. N4ITK assumes an image model similar to N3 in which v is the acquired image, u the uncorrupted image, f the bias field and n the noise. This model is often used and assumes the noise can be approximated by a Gaussian probability function. It is therefore independent of the bias field.[68]

$$v(x) = u(x)f(x) + n(x) \tag{3.7}$$

When assuming a noise-free scenario and using $\hat{u} = log(u)$ the new image model is:

$$\hat{v}(x) = \hat{u}(x) + \hat{f}(x) \tag{3.8}$$

To be able to estimate the bias field an iterative solution for the image model is derived for the corrected image at the n^{th} iteration.[64]

$$\hat{u}^n = \hat{u}^{n-1} - \hat{f}_r^n = \hat{u}^{n-1} - S\left\{\hat{u}^{n-1} - E[\hat{u}|\hat{u}^{n-1}]\right\} \tag{3.9}$$

The initial bias field estimate is usually set to 0. When the bias field estimate is 0 the estimate for the uncorrupted image equals the given image, so $\hat{u}^0 = \hat{v}$. The estimate of the residual bias field at the n^{th} iteration is $\hat{f}_r^n$. The smoothing of the bias field is described by smoothing operator S which is a B-spline approximator. In this smoothing operation $\hat{u}^{n-1}$ resembles the image resulting from the previous iteration. $E[\hat{u}|\hat{u}^{n-1}]$ is the expected value of the true image. The iteratieve scheme as described in equation 3.9 is designed to converge so that $\hat{f}_r^n \to 0$. In other words it converges to having no residual bias field. The total bias field ($\hat{f}_e^n$) is the sum of the residual bias fields[64]:

$$\hat{f}_e^n = \sum_{i=1}^{n} \hat{f}_r^i \tag{3.10}$$

Equation 3.10 can therefore be used to compute the total bias field after following the iterative scheme as described in equation 3.9.[64] The Python implementation was achieved by using the SimpleITK package.[69]

Intensity Normalization

As the Nyul method is one of the few that can be applied to multimodal data it was chosen for the intensity normalization method in this research. The method consists of three steps: choice of histogram landmarks, training and transformation.

Choice of histogram landmarks

In our application of Nyul normalization the histograms are pruned. In this case pruning takes places between the 1st (p_{low}) and 99th (p_{high}) percentile of the intensity range. MR histograms are usually bimodal, meaning that there are two peaks in the histogram. The first peak usually corresponds to the background, while the second peak, or mode, corresponds to the foreground. The foreground is identified using a thresholding approach. The used thresholding maps all voxels with an intensity higher than the mean intensity of the image to the foreground. The histogram landmarks include p_{low} and p_{high} as well as each decile in the intensity distribution of the foreground image.[70]

Training

The choice of landmarks results in an intensity landmark configuration C_L:

$$C_L = [p_{low}, m_{10}, m_{20}, m_{30}, m_{40}, m_{50}, m_{60}, m_{70}, m_{80}, m_{90}, p_{high}]$$

During training images are loaded and their histograms computed. The intensity values corresponding to the elements in C_L are determined. The intensity values belonging to p_{low} and p_{high} are mapped linearly into the minimum and maximum intensity in the standard scale. The outcome is used to find the new mapped locations of $m_{10}, ..., m_{90}$. The rounded means for $m_{10}, ..., m_{90}$ are calculated over all training images. This creates the standard scale.[70]

Transformation

The standard scale landmarks are used to transform a new MR image. The histogram of this new MR image is computed. Subsequently, the intensity values corresponding to p_{low} and p_{high} are determined. The same is done to for each decile. This results in an intensity landmark for the new image. Each element in the intensity landmark is linearly mapped to the corresponding element in the standard scale. This operation results in the transformation of the input image to the standardized image.[70]

The python implementation is based on the code by S. Valverde[71] and the publication of Shah et al.[70]

Scale between 0 and 1

The resulting intensities of the scan are normalized between 0 and 1. The pixel values will be small, which allows for less complicated and faster training of the network. The scans are normalized as described in equation 3.11 in which I represents the MRI scan.

$$I_{norm} = \frac{I - min(I)}{max(I) - min(I)} \tag{3.11}$$

3.3.3 Training of PAM

PAM is trained with the images from the TCIA dataset. 10% of the images from the TCIA dataset were used as a validation set to prevent overfitting of the network. Furthermore, the Adam optimizer with an initial learning rate of 8×10^{-5} was used during training. As described by Trebeschi et al.[43] a curriculum learning scheme was used where the loss was computed on an increasingly smoother version of the image. Kernel sizes for smoothing were $10, 5, 3$ and 1 and each were trained for 20 epochs. That means the network was trained for a total number of 80 epochs. The batch size was set to 2.

3.3.4 Prediction

The trained network is used to extract quantitative features from the scan pair that served as input for prognostication. These features are based on the feature maps in the deepest layer of the PAM architecture as visible in figure 3.5. Global average pooling is applied to these features. The dataset is split patient-wise in a train and test set. Both the training and test set contain 50% of the patients. The features of the scan pairs of the patients in the training set are fed to a random forest classifier (RFC) which is used to predict probabilities for the one-year survival after the follow-up scan of a patient. The RFC outputs a score between 0 and 1 serving as a probability of the one-year survival after the later scan in the scan pair. A random search for hyperparameters with grouped 5 fold cross validation is used to tune the hyperparameters within the training set. Feature selection was performed by removing features with a Pearson correlation > 0.6 and subsequently dropping features with a variance inflation factor > 5. The same procedure will be repeated while training the RFC exclusively with the scan pairs of primary brain tumors since this is the only cancer type in which survival is not affected by changes elsewhere in the body.

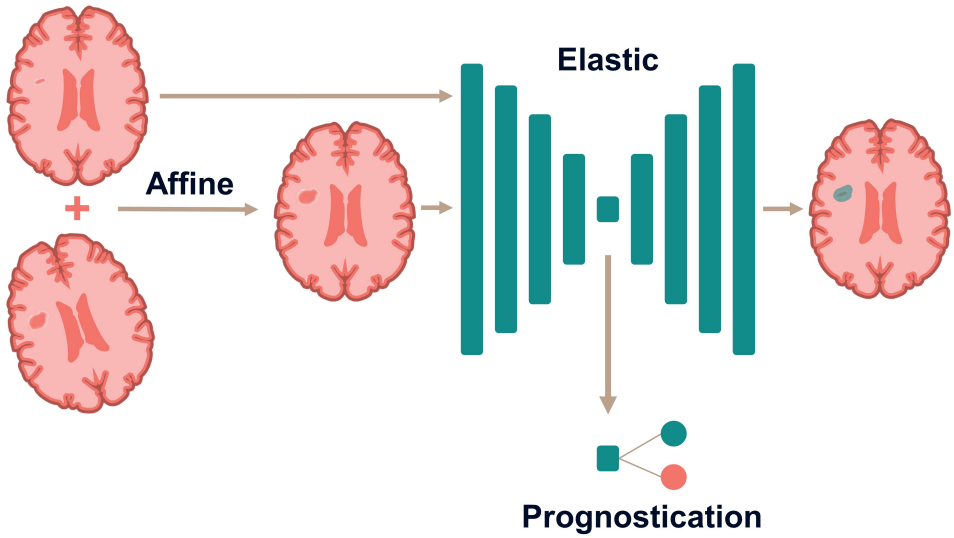


Figure 3.5: Overview of the prognostic AI monitor. Two MRI scans serve as input. The first step is affine transformation which aligns the two scans. The second step is the quantification of the elastic deformation field. The deepest layer in the PAM architecture is used for prognostication using a random forest classifier.

Due to the S-shape of the resulting ROC-AUC curve after training the RFC in some cases median centering of features was also tried.[72] Experiments will also be conducted to include temporal information in training the classifier. Adding the time between scans as well as relative timing of scans in the follow-up of a patient will be investigated. This is illustrated in figure 3.6.

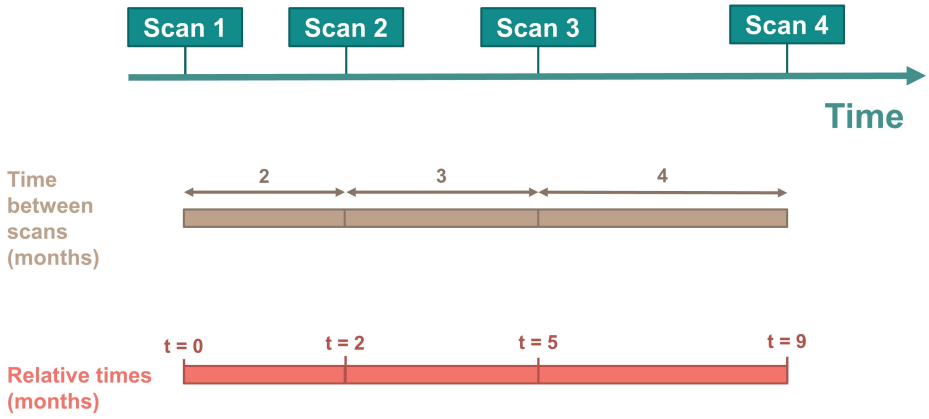


Figure 3.6: Illustration demonstrating the two ways of incorporating temporal information as a feature.

3.3.5 Statistical analysis

The risk score resulting from PAM will be analysed in two ways. Associative analysis will be performed using Cox time-varying regression analysis. This will analyse the relation between the PAM risk score and patient survival as well as the relation between possible confounders and patient survival. A predictive analysis will be based on the concordance index (C-index) and Kaplan-Meier survival curves. The C-index is similar to the normally used area under the curve (AUC) of the receiver operating characteristic (ROC) curve. However, the C-index is capable of taking into account censored data, making it the method of choice for survival analysis. Kaplan-Meier survival curves for different values of the PAM risk score will also be provided. The three risk categories will be created using a quantile-based discretization function. To compare the three risk categories a logrank test will be performed in which the null hypothesis is no difference in survival between both groups. Subanalyses will be done for various cancer types and methods.

3.4 Results

3.4.1 Data harmonization

A random subset of 10 FLAIR images was taken from the TCIA dataset to show the effect of data harmonization on the image histograms. The histograms before and after data harmonization are visible in figure 3.7.

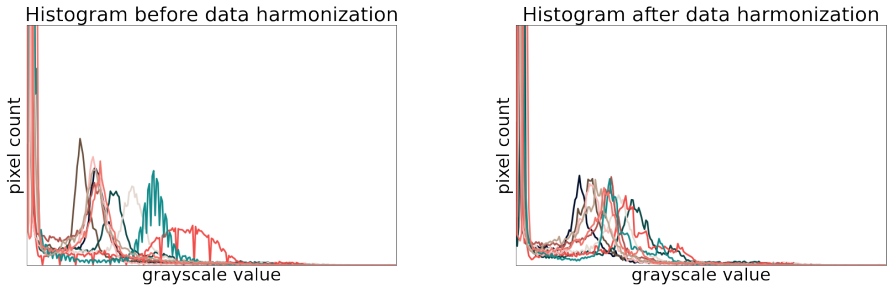


Figure 3.7: Histogram plots showing the effect of data harmonization. The second peaks in the histograms before data harmonization (a) are shifted towards one another after data harmonization (b).

The general effect of all data harmonization steps for a sample of three random images from the TCIA database is visualized in figure 3.8.

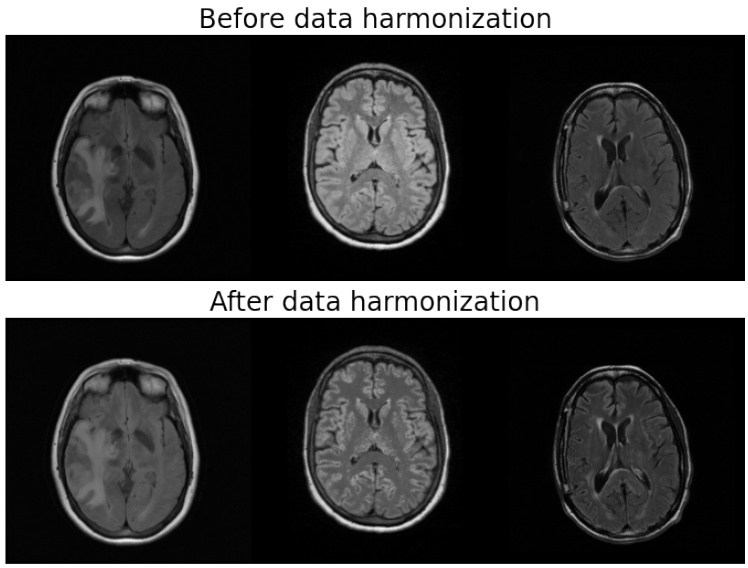


Figure 3.8: The effect of all described data harmonization steps on a random sample of three MRI scans from the TCIA dataset.

The effect of data harmonization on training of PAM has also been investigated. This is shown in figure 3.9.

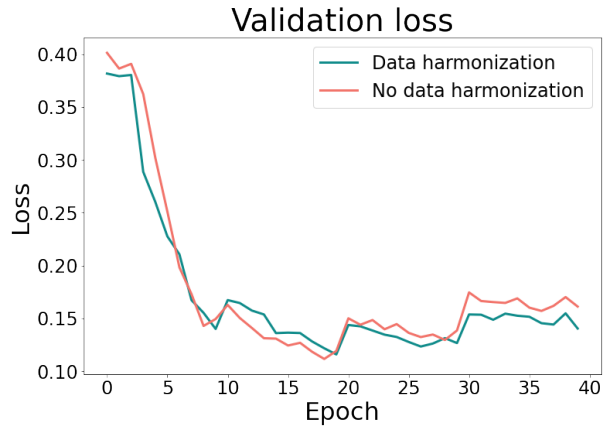


Figure 3.9: Effect of data harmonization on validation loss of PAM.

3.4.2 Visual representation of image registration

An example of image registration using PAM is provided for two randomly selected images. Figure 3.10 shows the original fixed and moving image. Figure 3.11 shows the transformed image.

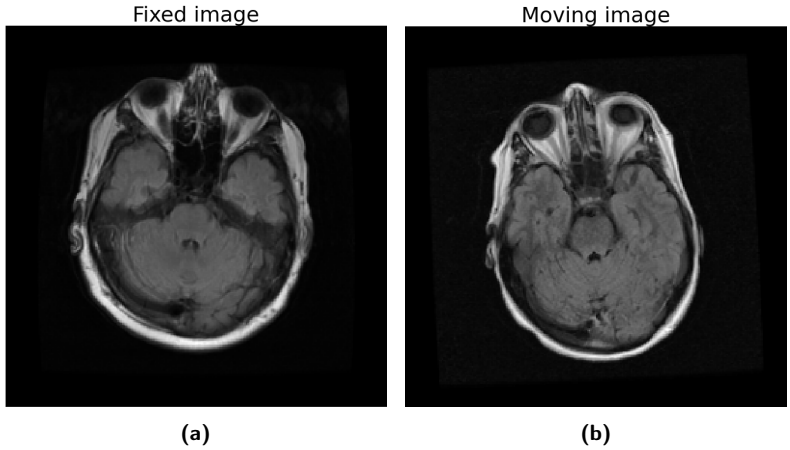


Figure 3.10: Original fixed (a) and moving image (b) before image registration by PAM.

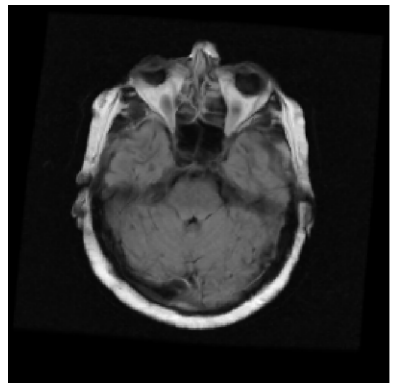
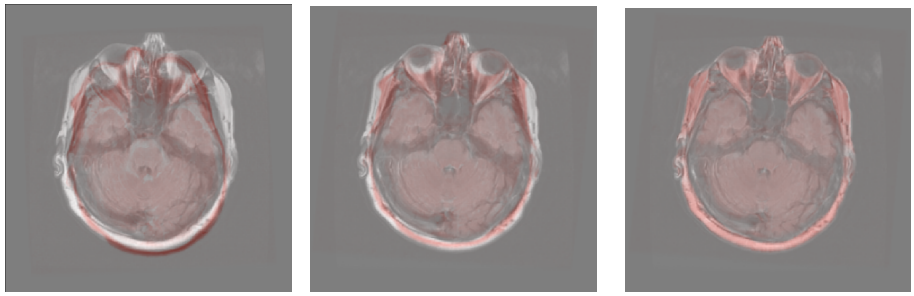


Figure 3.11: Transformed moving image after image registration by the prognostic AI monitor algorithm

Figure 3.12 shows the overlays of the fixed and (transformed) moving image.



(a) (b) (c)

Figure 3.12: Overlay of (a) fixed and original moving image, (b) fixed image and affine transform of the moving image and (c) fixed image and deformable transform of the moving image

Lastly, a checkerboard combining the fixed and transformed moving image is provided in figure 3.13 to show the effect of the transformation.

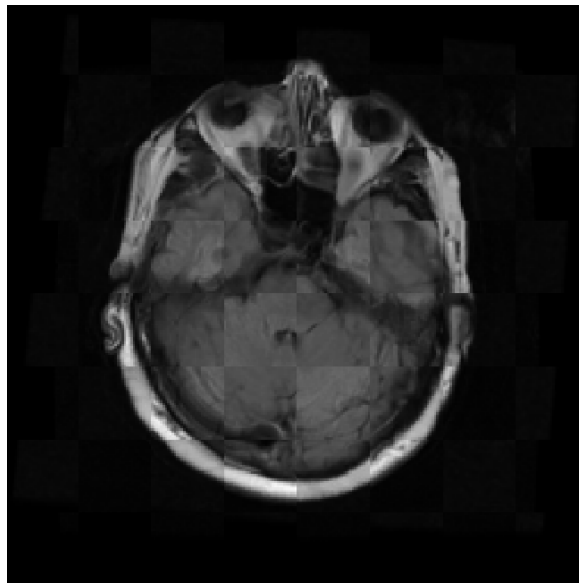


Figure 3.13: Checkerboard image of the fixed image and the final transform of the moving image.

3.4.3 Statistical analysis

The statistical analysis is divided in two parts: the associative and predictive analysis.

Associative analysis

The associative analysis was further subdivided in two parts. In the first part data from the complete dataset is used which includes various cancer types. In the second part only the subset of patients with a primary malignancy of the brain is used.

All cancer types

First the association of the plain PAM score with survival was assessed. Our findings show that this results in a high hazard ratio of 13.05, but the confidence interval is rather broad. This results in the high hazard ratio not being significant. It was observed that for some cancer types the AUC curve had an S-shaped distribution. To correct for this median centering was applied. This resulted in no improvement in the predictive performance as can be observed in table 3.2. The next step was to add the time between scans in the classifier. This increased the predictive performance of the model to a hazard ratio of 35.77 with a p-value of 0.01. In an attempt to further improve the performance the relative timing of each scan was incorporated. This decreased the performance in the associative analysis to a hazard ratio of 9.7 with a p-value of <0.005. To minimize overfitting feature selection was performed which resulted in a hazard ratio of 34.54 with a p-value of <0.005. The resulting logarithms of the hazard ratios are visualised in figure 3.14 and the corresponding hazard ratios with 95% confidence interval and p-value are shown in table 3.2.

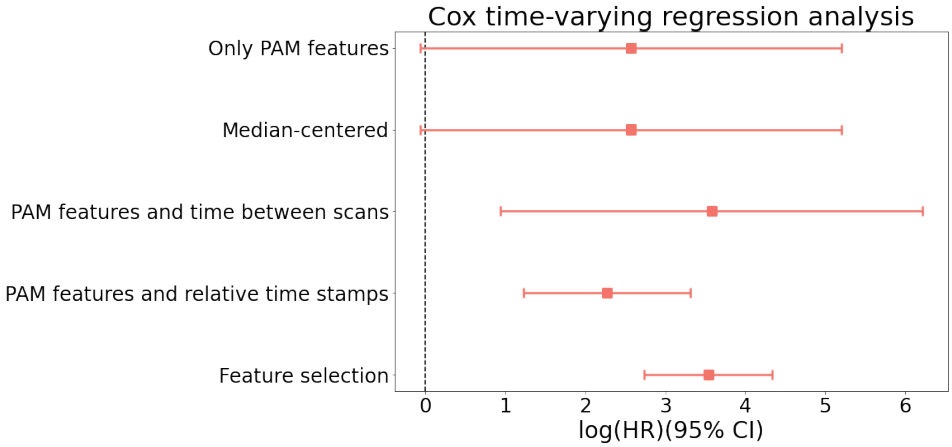


Figure 3.14: Cox time-varying regression analysis for the prognostic AI monitor risk score for various methods of computing the risk score.

Table 3.2: Hazard ratios for the prognostic AI monitor risk score computed in various ways.

Method	Hazard ratio	95% CI	p-value
Only PAM features	13.05	[0.93 - 183.46]	0.06
Center PAM score around median	13.05	[0.93 - 183.46]	0.06
PAM features and time between scans	35.77	[2.54 - 503.63]	0.01
PAM features and relative time stamps	9.7	[3.41 - 27.58]	<0.005
Feature selection	34.54	[15.52 - 76.87]	<0.005

Further subanalyses and analysis of confounders were done for the risk score as predicted in the last method. The RFC in this method was trained on a total 1037 scan pairs of 285 patients in the and the test set consisted of 937 scan pairs from 286 patients. The outcome of the Cox time-varying regression model on the test set is presented in figure 3.15. The corresponding hazard ratios with 95% confidence and p-value are visible in table 3.3. Only the PAM risk score and whole brain radiation therapy show a significant association with survival ($p < 0.05$).

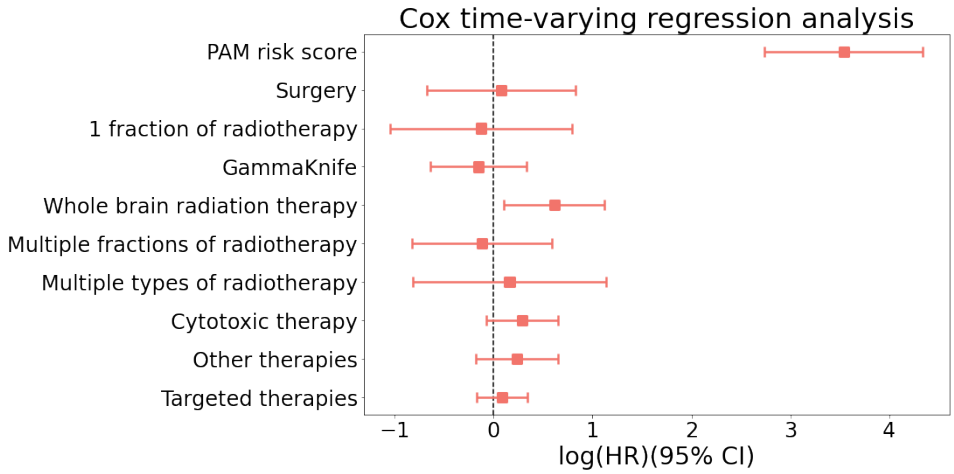


Figure 3.15: Cox time-varying regression analysis for the prognostic AI monitor risk score. Different treatment options were used as confounders in this analysis.

Table 3.3: Overview of the hazard ratios, 95% confidence interval and p-value resulting from Cox time varying analysis for the risk score as predicted by the prognostic AI monitor and various treatment options as possible confounders.

Variable	Hazard ratio	95% CI	p-value
PAM risk score	34.54	[15.52 - 76.87]	<0.005
Surgery	1.21	[0.55 - 2.64]	0.63
1 fraction of radiotherapy	0.87	[0.35 - 2.17]	0.76
GammaKnife	0.78	[0.47 - 1.27]	0.32
Whole brain radiation therapy	1.86	[1.13 - 3.06]	0.01
Multiple fractions of radiotherapy	0.79	[0.39 - 1.63]	0.53
Multiple types of radiotherapy	0.95	[0.32 - 2.86]	0.93
Cytotoxic therapy	0.20	[0.85 - 1.76]	0.28
Other therapies	1.10	[0.73 - 1.65]	0.65
Targeted therapies	1.00	[0.77 - 1.30]	1.00

Primary malignancy of the brain

When training the RFC solely on patients with primary brain tumors the results differ from above. The training set consists of 194 scan pairs from 53 patients and the test set contains 238 scan pairs of 54 patients. The Cox-time varying regression analysis for these patients can be found in figure 3.16 with the corresponding hazard ratios in table 3.4. Only the PAM risk score shows a significant association with survival.

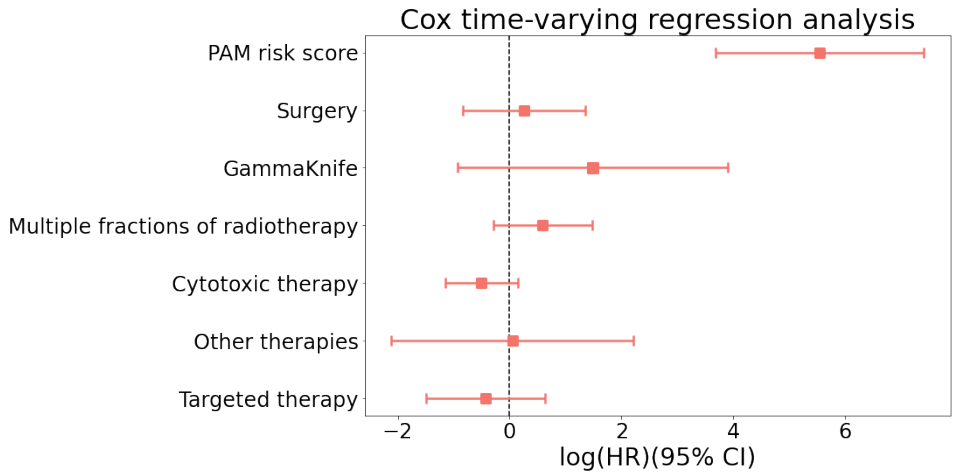


Figure 3.16: Cox time-varying regression analysis for the prognostic AI monitor risk score and confounders for patients with a primary brain tumor. The confounders are the various treatment options of the patient.

Table 3.4: Overview of the hazard ratios, 95% confidence interval and p-value resulting from Cox time varying analysis for patients with a primary malignancy of the brain. The risk score as predicted by the prognostic AI monitor and various treatment options as possible confounders are used in this analysis.

Variable	Hazard ratio	95% CI	p-value
PAM risk score	254.95	[39.51 - 1645.15]	<0.005
Surgery	1.29	[0.43 - 3.88]	0.63
GammaKnife	4.41	[0.39 - 49.70]	0.23
Multiple fractions of radiotherapy	1.81	[0.75 - 4.38]	0.19
Cytotoxic therapy	0.61	[0.32 - 1.16]	0.13
Other therapies	1.05	[0.12 - 9.15	0.96
Targeted therapies	0.65	[0.22 - 1.88]	0.42

Predictive analysis

All cancer types

The five different methods to compute a risk score were also evaluated on their predictive quality. The train set for the first three methods consists of 1588 scan pairs of 302 patients, the test set consists of 1531 scan pairs of 303 patients. The method using relative timing of scans with and without feature selection has 1037 scan pairs of 285 patients in the training set and 937 scan pairs of 286 patients in the test set. The resulting C-indices can be found in table 3.5. The C-index for overall survival, so as long as we had data for a given patient, is provided. Moreover, the C-index for a 1 year follow up of the patient is provided taking into account the survival time within that year. The C-index for the best performing RFC trained on the complete dataset with all cancer types is 0.59. A further sub-analysis on the best performing method for different cancer types is provided in 3.6.

Table 3.5: C-indices for the prognostic AI monitor risk score computed in various ways.

Method	C-index (overall survival)	C-index (1 year surival)
Only PAM features	0.54	0.54
Center PAM score around median	0.54	0.54
PAM features and time between scans	0.55	0.55
PAM features and relative time stamps	0.59	0.60
Feature selection	0.59	0.61

Table 3.6: C-indices for the predicted risk score for the three most common cancer types in the dataset.

Cancer type	C-index (overall survival)	C-index (1 year surival)
All cancer types	0.59	0.61
Brain, lung and melanoma	0.61	0.64
Brain	0.64	0.70
Bronchus and lung	0.60	0.61
Melanoma	0.60	0.62

The Kaplan Meier survival scores for the three PAM risk categories are provided in figure 3.17. High risk is defined as a PAM risk score between 0.394 and 0.648, medium risk between 0.289 and 0.394 and low risk between 0.145 and 0.289. Table 3.7 shows all three risk categories significantly differ from each other ($p < 0.05$).

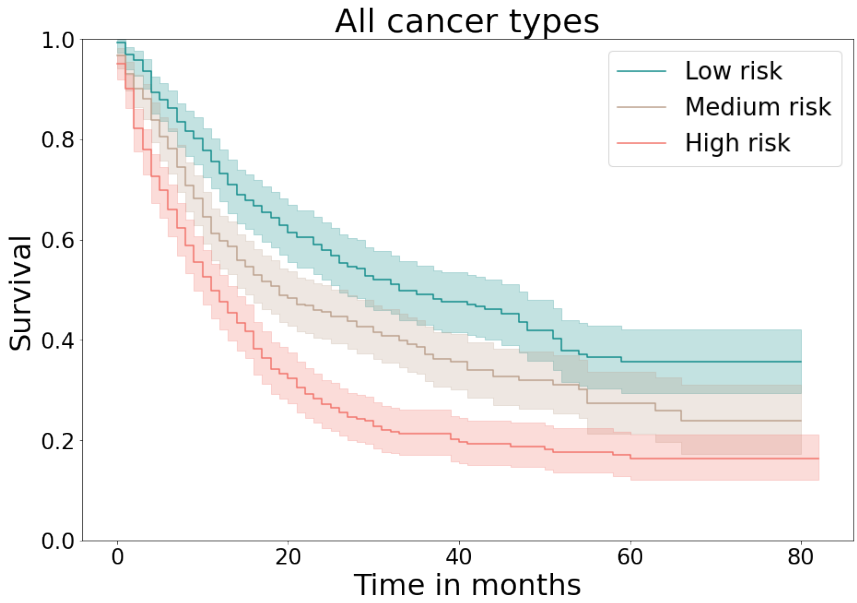


Figure 3.17: Kaplan Meier survival curves for three risk categories based on the risk score provided by the prognostic AI monitor for the complete dataset with all cancer types.

Table 3.7: Results of the logrank test for all cancer types based on the predicted risk scores with the FLAIR sequence.

Groups	Test statistic	p
Low risk - medium risk	11.40	<0.005
Medium risk - high risk	10.07	<0.005
Low risk - high risk	37.13	<0.005

Primary malignancy of the brain

The C-index corresponding to the PAM risk score when the RFC is trained on patients with a primary malignancy of the brain is 0.67 when looking at total survival. When using the cut-off of 1 year survival the C-index equals 0.73 . The Kaplan Meier survival scores for the three PAM risk categories are provided in figure 3.18. High risk is defined as a PAM risk score between 0.353 and 0.55, medium risk between 0.274 and 0.353 and low risk between 0.202 and 0.274. These cut-offs were determined based on the complete dataset of patients with primary malignant neoplasms of the brain, so the test set does not necessarily have an even distribution. This difference in group sizes explains the different sizes of confidence intervals as visible in figure 3.18. Table 3.8 shows all three risk categories significantly differ from each other ($p < 0.05$).

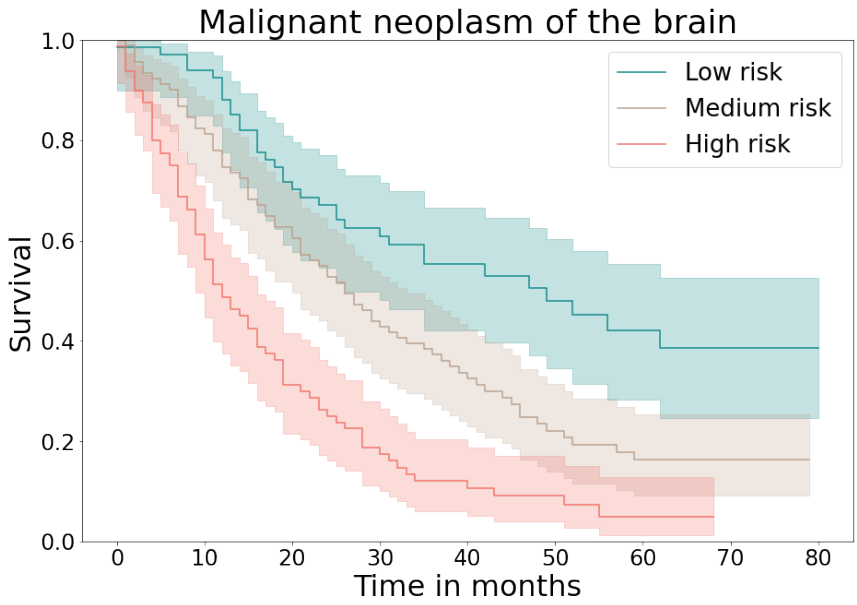


Figure 3.18: Kaplan Meier survival curves for three risk categories based on the risk score provided by the prognostic AI monitor.

Table 3.8: Results of the logrank test for patients with primary brain cancer based on the predicted risk scores with FLAIR sequence.

Groups	Test statistic	p
Low risk - medium risk	9.30	<0.005
Medium risk - high risk	17.50	<0.005
Low risk - high risk	41.70	<0.005

3.5 Discussion

In this chapter we investigated the conversion of PAM to single sequence MRI including a preprocessing method and analyzed its performance for survival prediction.

The data harmonization algorithm showed a better alignment of image histograms. However, it remains questionable if this affects the performance of PAM. The effect of data harmonization becomes more apparent in the last epochs. Due to the curriculum learning scheme these are the epochs with smaller smoothing kernels hence focusing on the details. Since the algorithm is easy to implement and fast it could easily be implemented in the downloading pipeline of MRI scans from the server.

The associative analysis shows the risk score as predicted by PAM has a significant association with survival in both the model trained on all cancer types and the model trained specifically for primary malignancies of the brain. The broad confidence interval of the hazard ratio of the predicted scores shows there is still an uncertainty in the predictions by PAM. The fact that most treatments show no significant association with survival can be explained by the way the dataset was build and missing information. The Cox time-varying regression model does not allow for categorical variables, which means all treatment options were set to dummy variables with value 0 or 1. This for examples lacks information about medication dosages, therapies before the start of follow-up for this study and the difference between an emergency and planned surgery.

The hazard ratio for primary brain cancer only is higher than of the pancancer data suggesting the survival of patients with primary malignancies of the brain is better predictable than metastatic disease. This is supported by the higher C-indices of the classifier trained only on patients with primary malignancy of the brain. This difference can be explained by the fact that for primary brain malignancy the total tumor burden is visible on the brain MRI, while this is

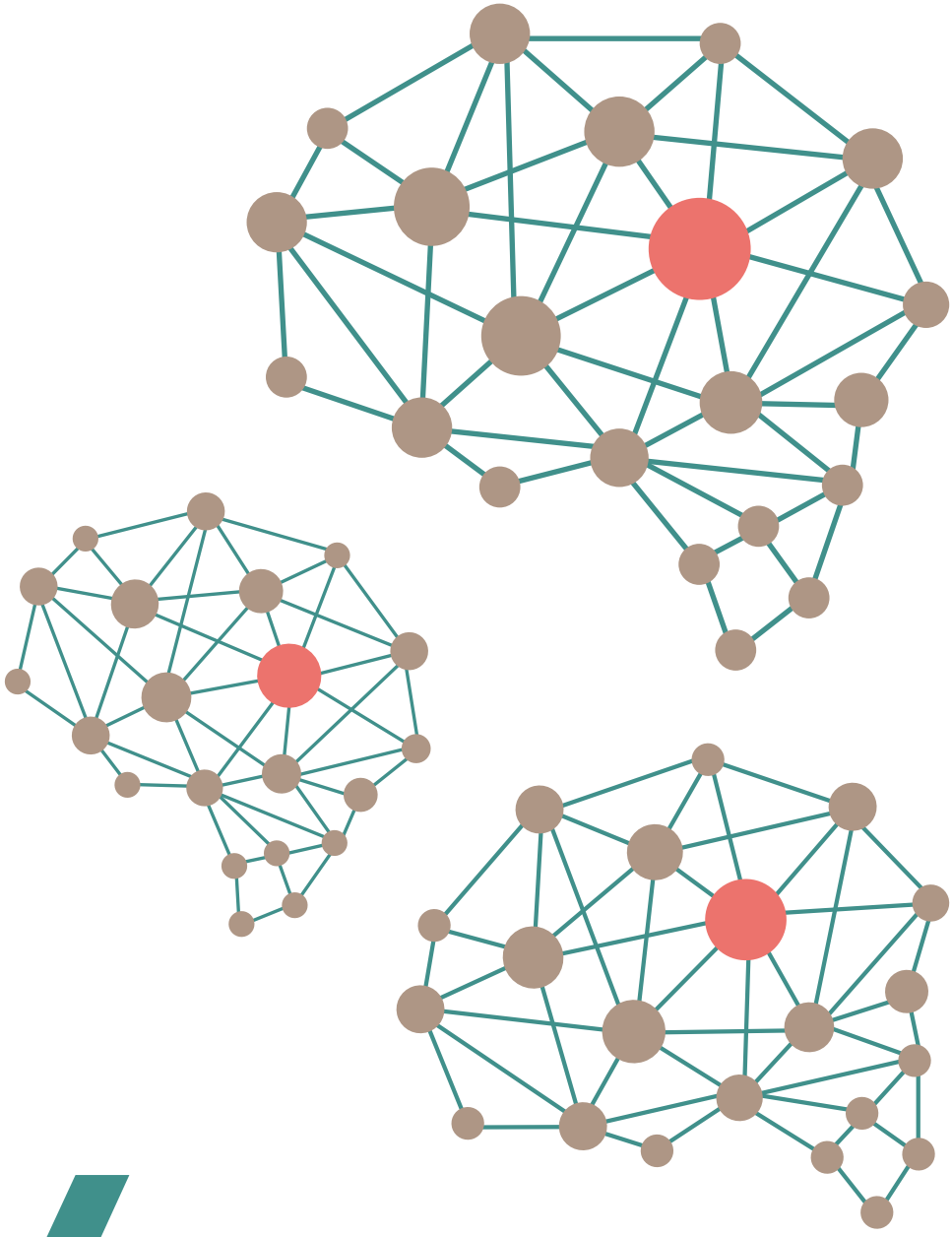
often not the case in metastatic disease. Thawani et al. found that for patients with brain metastases due to lung cancer the brain metastases were the cause of death in 33% of patients, while 67% of deaths was caused by progression of systemic disease.[73] The Kaplan Meier curves in the predictive analysis show that when using PAM for all cancer types the algorithm performs particularly well in isolating the low risk patients. The difficulty in distinguishing between tumor progression and treatment side effects is a possible explanation for the results. Low risk patients do not show an increase in hyperintense signal on the FLAIR images. The high and medium risks are more difficult to differentiate as an increase in signal intensity may be due to progression or side effects. A combination of sequences should provide more information and will be discussed in chapter 4. When PAM is trained only on patients with a primary malignancy of the brain PAM is able to isolate the low, medium and high risk patients. These could clinically be translated to patients with response to treatment, stable disease or progressive disease, respectively.

Although these results show a promising new application of the prognostic AI-monitor there is also still room for improvement. First of all, the RFC does not take into account the fact that multiple scan pairs are from the same patient. Of course the training and testing set were split patient-wise to avoid bias by training and testing on the same patient. However, it looks at each scan pair as an isolate entity. To correct for this as much as possible in the RFC the scans were ordered chronologically with relative time stamps. The first scan pair started at time 0 and so on for each following scan pair. These time stamps were also fed to the RFC as extra features as well as the duration between the scans. This gives some indication of time to the RFC, but is not the same as seeing one patient as an isolate entity with various scores predicted through the course of time. The same holds for the 1 year survival cut-off that was used. The RFC needs a given cut off with a binary classification of survival to make its prediction. In reality it matters of course if a patient died within 1 or 11 months within that year. One way to overcome this problem would be to use a random survival forest.[74] The random survival forest is able to deal with censored data. It uses features, the survival time and the possible death of a patient as input and can be used to predict survival as well as a risk score. Another option would be to further investigate DeepSurv. DeepSurv is a Cox proportional hazards deep neural network. It is build to model treatment effect based on patient variables.[75] It was decided it is beyond the scope of this research to further investigate the random survival forest or DeepSurv as a substitute of the RFC. Further research could point out the possible advantages and disadvantages of the random survival forest for PAM and further investigate other options such as DeepSurv.

Another option to possibly boost the performance of PAM is to include more data. Of the original 4703 scan pairs only 1942 could be used in the final analysis of the FLAIR scans. This is due to the chronological ordering of the scans. Initially there were scan pairs between the first baseline scan and the third follow up scan (if these were four months apart). These were excluded from the dataset. Since the performance is better when using the relative time stamps re-including these scans would decrease performance. An option worth investigating is extending the time between scans from four to six months. Especially in patients with long radiologic and clinical stable disease the interval of follow-up scans may well be extended beyond four months.[76]

3.6 Conclusion

The first step of translating PAM to MRI shows PAM is able to extract the low risk patients in a pancancer dataset of patients with brain tumors. When trained and evaluated on primary brain tumors where the total patient tumor burden is visible on the brain scan PAM is even able to categorize low, medium and high risk patients. In both cases the risk score as computed by PAM shows a significant association with survival. Further research must be conducted in incorporating the timing of various scans of the same patient in the computation of the risk score and to increase the time between scans beyond four months. Additionally, only using one sequence is limited use of all available data, so the method should be extend to multisequence imaging of the brain.



4

Chapter 4

Extension of PAM to multisequence imaging of the brain

4.1 Introduction

The role of MRI in brain tumor evaluation is determining lesion location, extent of tissue involvement and mass effect upon the brain, ventricular system and/or vasculature. In general the two most common types of MRI images are the T1 and T2 weighted images. The foundation of MRI evaluation of brain tumors is provided by the structural sequences: T2, fluid-attenuated inversion recovery (FLAIR) and T1 pre- and postcontrast.[77]

4.1.1 MRI sequences

T1-weighted
T1-weighted MR images are acquired by using a short echo time (TE) and repetition time (TR) and rely on the longitudinal relaxation of the net magnetization vector. T1 weighted images are used to evaluate the tissue architecture. The T1 time of fat is shorter than water. That means the fat vector realigns with B_0 faster than the water vector. After a given TR the next $90°$ RF pulse is applied. This flips the longitudinal components into the transverse plane. That means the transverse magnetization in fat is also larger than in water. The fast recovery of the longitudinal magnetization and subsequent $90°$ RF pulse make fat appear bright and water dark on T1-weighted images when the TR is short. When using a long TR both fat and water have returned to equilibrium and no contrast is present.[78] Figure 4.1 shows the T1 relaxation and corresponding MR image.

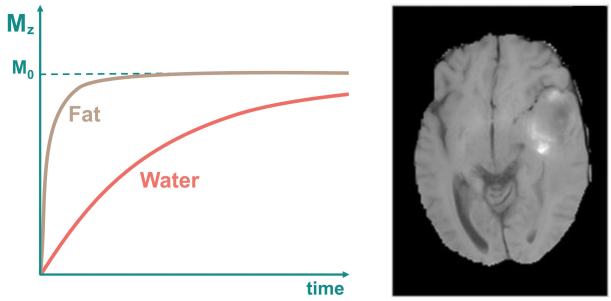


Figure 4.1: T1-weighted MR imaging. a) shows the faster recovery of the longitudinal magnetization of fat compared to water. b) shows the effect of these differences in T1 in an image of the brain. Adapted from Usman and Rajpoot.[79]

T2-weighted

T2-weighted MR images are acquired by using long TE and TR times. In contrast to T1 images, T2 images rely on the transverse magnetization component of the net magnetization vector. T2-weighted MR images are structural images and therefore used to evaluate the tissue architecture. Water appears bright on T2-weighted images as the transverse magnetization component takes relatively long to decay. Fat, on the other hand, has a short T2 time meaning the transverse magnetization decays faster. This leads to fat appearing darker on T2-weighted images.[78] With regards to brain tumors high intensity may be seen if there is peritumoral edema present, but also in nonenhancing tumor, white matter injury and gliosis. The peritumoral edema could be vasogenic or infiltrative in nature. Sometimes this hyperintense area cannot be differentiated from the primary tumor.[77] Figure 4.2 shows the T2 relaxation and corresponding MR image.

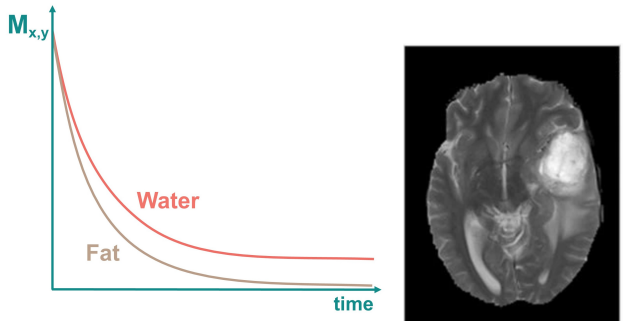


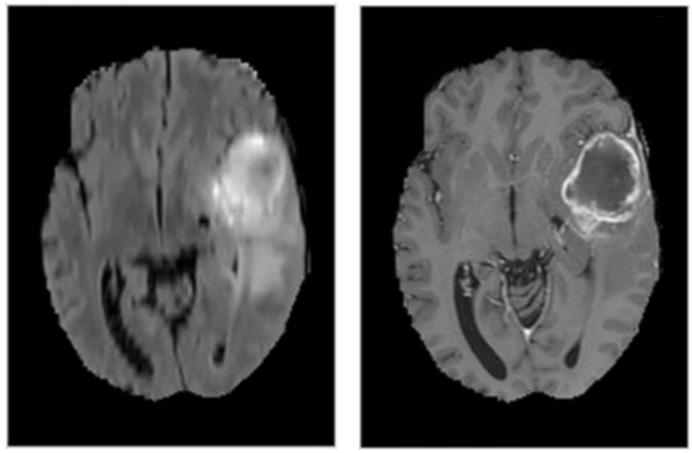
Figure 4.2: T2-weighted MR imaging. a) shows the decay of the transverse magnetization of both water and fat b) shows the effect of these differences in T2 in an image of the brain. Adapted from Usman and Rajpoot.[79]

FLAIR

FLAIR is a variation of an inversion recovery sequence. In the FLAIR sequence the signal from CSF is nulled by selecting an inversion time (TI) corresponding to the time of recovery of CSF from 180 degrees to the transverse plane. So, there is no longitudinal magnetisation present in the CSF. This sequence cancels out the signal from the CSF and can be used to inspect for pathology adjacent to the CSF.[78] FLAIR MR images are acquired by using very long TE and TR times. High intensity is seen in peritumoral edema (vasogenic and infiltrative), nonenhancing tumor, white matter injury and gliosis.[77] Figure 4.3a shows the FLAIR image.

Postcontrast T1

The most used contrast agent in MRI is gadolinium. Gadolinium is a non-toxic paramagnetic contrast agent with a short T1 time. This means gadolinium appears bright on T1 weighted images. Postcontrast enhancement reflects nonspecific breakdown of the blood–brain barrier. Gadolinium accentuates areas in the brain where the BBB is leaking. When the contrast agent leaks through the BBB it could be a feature of tumors, but it is also seen in some non-neoplastic conditions.[77] Figure 4.3b shows the postcontrast T1 image.



(a) **(b)**

Figure 4.3: a) Fluid attenuated inversion recovery image of the brain. There is no signal originating from the cerebral spinal fluid. b) T1 image with gadolinium contrast. Adapted from Usman and Rajpoot.[79]

4.1.2 Clinical significance of various sequences

Multisequence MRI is generally used in the evaluation of brain tumors. The different sequences provide complementary information. The exact image characteristics depend on the type of tumor. High grade glioma usually shows enhancement in the tumor on the T1 contrast enhanced image. The tumor causes disruptions in the blood brain barrier leading to uptake of contrast in the tumor. However, breakdown of the BBB is not unique to neoplasms and not all gliomas cause a disruption in the BBB. Hence there might be non-enhancing components of the tumor on the T1 contrast enhanced image. Moreover hyperintense regions on FLAIR and T2 may be caused by peritumoral edema. Peritumoral edema can be vasogenic which is a reaction of the tissue to the tumor in which it leaks plasma fluid, but there are no tumor cells present. This is usually seen in brain metastases or extra axial tumors. On the other hand, infiltrative edema is a mixture of vasogenic edema and infiltrating tumor cells more often seen in gliomas.[77] It might be difficult to define the exact location and extent of the tumor. In radiotherapy the Gross Tumor Volume is used. This is defined as the hyperintense region on FLAIR and T2-weighted images combined with the region showing contrast enhancement in the T1-weighted image. Sometimes a small margin is added before radiotherapy to account for microscopic infiltration of tumor cells (Clinicial Target Volume). Especially in high grade glioma there can be a lot of heterogeneity within the lesion. It may consist of enhancing tumor, non-enhancing tumor, (post-radiation) necrosis and peritumoral edema.[80]

Having multiple tumor lesions in the brain is often associated with metastases. In general metastases appear hypointense to isointense to the surrounding brain tissue on T1-weighted images. Surrounding hypointense edema may be observed. Metastases predominantly appear hyperintense on T2-weighted and FLAIR images. The peritumoral edema may even appear more hyperintense.[81] Brain metastases of melanoma do not always fit this general concept. Appearance of these metastases is usually though one of two patterns: the melanotic pattern or the amelanotic pattern. The melanotic pattern is characterized by hyperintense signal on T1-weighted images and hypointense signal on T2-weighted images. The amelanotic pattern matches the before mentioned general concept.[82] Commonly, brain metastases densely enhance in postcontrast images due to disruption of the BBB and angioneogenesis. The common enhancement patterns are solid enhancement or ring-like enhacement. The ring is usually thick and irregular. It is often possible to differentiate from non-neoplastic lesions based on the enhancement pattern, but a distinction between metastases and a primary brain tumor is usually not possible.[81]

Standardised assessment of brain tumors by a radiologist usually follows the RANO criteria. The current RANO criteria as described in section 1.3.2 take into account some of the information of different sequences. The MRI characteristics for each possible response category in the RANO criteria are provided in table 4.1. This table does not contain all RANO criteria such as clinical status and use of corticosteroids. The RANO criteria provide a decent foundation for the assessment of brain tumors, but do not take into account all possible prognostic features of the brain and brain disease. There are numerous possible variations in the brain, especially after treatment. For example, a pseudo-response might exhibit itself on the MRI represented by a decrease in contrast enhancement in the tumor without the tumor actually responding to the treatment. This principle can be observed in patients receiving treatment with bevacizumab. Pseudo-progression might also occur, especially after chemoradiation with temozolomide. In 20% of the patients the tumor seems to initially qualify for progressive disease while after some time response is visible.[83]

Table 4.1: Overview of MRI image characteristics associated with the various response categories in the Response Assessment in Neuro-Oncology (RANO) criteria. Adapted from Chukwueke and Wen.[22]

Response category	MRI characteristics
Complete response	No new lesions Complete resolution of tumor enhancement on the T1 contrast enhanced image Stable or decreased T2/FLAIR signal abnormality
Partial response	No new lesions $\geq$50% reduction of tumor enhancement on the T1 contrast enhanced image Stable or decreased T2/FLAIR signal abnormality
Stable disease	No new lesions <50% decrease but <25 % increase in tumor enhancement on the T1 contrast enhanced image Stable or decreased T2/FLAIR signal abnormality
Progressive disease	Any new lesions >25% increase in tumor enhancement on the T1 contrast enhanced image Increase in FLAIR/T2 signal abnormality

4.2 Methods

Scans were selected from the TCIA dataset as described in section 3.3.1. The resulting distribution among sequences can be found in table 4.2. These scans are used to train PAM on.

Table 4.2: Available scans per sequences after selecting scans from the TCIA dataset.

Sequence	Number of scans
FLAIR	546
Postcontrast	287
Precontrast	387
T2	742

Further analysis and survival prediction was done on the NKI dataset. The consort diagram in figure 4.4 demonstrates the steps taken to get to the final dataset. Patient characteristics of the patients included in the final dataset can be found in 4.3.

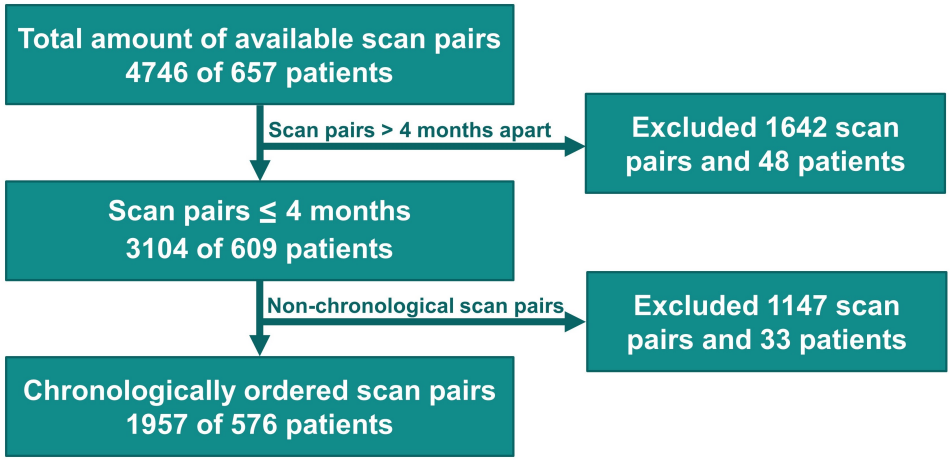


Figure 4.4: Consort diagram of patient and scan selection for analysis of the prognostic AI monitor

Table 4.3: Patient demographic of patients with multisequence scans showing the following patient characteristics: age, gender, survival and diagnosis.

Characteristic	N = 576	Percentage (%)
Age		
Median	60	
Range	17 - 96	
Gender		
Female	278	48
Male	298	52
1 year survival		
Alive	327	57
Deceased	249	43
Diagnosis		
Bronchus and lung	171	30
Melanoma	152	26
Brain	107	19
Other	146	25

An external validation will be performed on the dataset from Deventer Ziekenhuis. A total of 119 patients were included The patient demographic of this subset is shown in figure 4.4.

Table 4.4: Patient demographic of patients from Deventer Ziekenhuis with multisequence scans showing the following patient characteristics: age, gender, survival and diagnosis.

Characteristic	N = 119	Percentage (%)
Age		
Median	61	
Range	17 - 85	
Gender		
Female	63	53
Male	56	47
1 year survival		
Alive	45	38
Deceased	74	62
Diagnosis		
Brain	67	56
Bronchus and lung	30	25
Breast	15	13
Other	7	6

4.2.1 Creating multisequence MR data

At first instance all sequences are saved as separate files by the MR scanner. However, PAM must receive the data as one grid. In order to achieve this the sequences are first concatenated. Not all four sequences are available for each patient. This leads to various shapes for the multisequence MR image data. For example, $256 \times 256 \times 256 \times 2$ for scans with two available sequences and $256 \times 256 \times 256 \times 4$ for scans with four available sequences. This makes computation very difficult as the network is not designed to differentiate between sequences and therefore cannot identify which sequence is missing. To overcome this problem all multisequence data is saved in size $256 \times 256 \times 256 \times 4$ where the missing sequences are replaced by zeros. This is illustrated in figure 4.5. This way the network can compare the two MRI scans and exclude sequences from its evaluation that are not present in both baseline and follow-up scan. It remains important to only feed the network scan pairs that have at least one overlapping sequence.

Figure 4.5: Dealing with missing sequences in multisequence data. Missing sequences are replaced by zeros to ensure a uniform scan size.

4.2.2 Adaptation of the network to multisequence images

The original PAM is designed for single sequence data. Its design must therefore be altered to be able to deal with multisequence MR data. Since the same layers and architecture still apply it suffices to change the input shape to the shape of the multisequence data. That means the network receives an input with shape $256 \times 256 \times 256 \times 4$ instead of $256 \times 256 \times 256 \times 1$. The loss function must also be adapted to deal with multisequence data and possible missing sequences. Each sequence is saved in one channel of the input grid. The goal is to do

a channel wise correlation between baseline and follow-up scan. Subsequently, one loss resulting from these channel wise correlations should be calculated. For each channel or sequence the loss is calculated as described in section 3.2.1. Shen et al.[84] described a method to use a multiclass loss based on various individual losses. This method was adapted to PAM to combine the losses of the individual sequences to one multisequence loss as shown in equation 4.1 in which $\mathcal{L}$ is the multisequence loss, n the number of sequences, w_s the weight for a given sequence and L_s the loss for a given sequence.

$$\mathcal{L} = \frac{\sum_{s=1}^{n} w_s L_s}{\sum_{s=1}^{n} w_s + \epsilon} \tag{4.1}$$

The weight (w_s) equals zero when a sequence is not present in either the baseline or follow-up scan. When a sequence is present in both baseline and follow-up scans $w_s = 1$. ϵ is set to $1 \cdot 10^{-6}$ to avoid division by zero.

4.2.3 Training of PAM

As in the previous chapter PAM is trained with the images from the TCIA dataset. 10% of the images from the TCIA dataset were used as a validation set to prevent overfitting of the network. Furthermore, the Adam optimizer with an initial learning rate of 8×10^{-5} was used during training. A curriculum learning scheme was used with kernel sizes for smoothing $10, 5, 3$ and 1 trained for 50 epochs each. That means the network was trained for a total number of 200 epochs. The batch size was set to 2.

4.2.4 Prediction

The trained network is used to extract quantitative features from the scan pair that served as input in a similar manner as the previous chapter. First, global average pooling is applied to the features. Then the features are fed to a random forest classifier (RFC) which is used to predict the one-year survival after the follow-up scan of a patient. The RFC outputs a score between 0 and 1 serving as a probability prediction of the one-year survival. As in chapter 3 experiments will be conducted with incorporating a time component and feature selection.

4.2.5 Statistical analysis

The risk score resulting from PAM will be analysed in a similar way as is described in section 3.3.5. That means there will be an associative analysis using Cox time varying regression and a predictive analysis using the C-index and Kaplan-Meier survival curves. To compare the three risk categories a logrank test will be

performed in which the null hypothesis is no difference in survival between both groups. Both analyses will be done for all cancer types in the dataset as well as for patients with a primary malignancy of the brain. For the external validation not all treatment confounders have dates, so a Cox time varying analysis cannot be performed. Instead the normal Cox regression analysis is used. Subanalyses will again be done for different cancer types and the different methods of training the classifier.

4.3 Results

4.3.1 Statistical analysis

The results of the statistical analyses will be presented for training the RFC on all cancer types in the dataset and for training the RFC only on primary brain tumors.

All cancer types

The statistical analysis is divided in two parts: the associative and predictive analysis.

Associative analysis

First the association of the plain PAM score with survival was assessed. Our findings show that this results in a high hazard ratio of 84.34 with a p-value of <0.005. It was observed that for some cancer types the ROC-AUC curve had an S-shaped distribution. To correct for this median centering was applied. This resulted in no improvement in the predictive performance as can be observed in table 4.5. The next step was to add the time between scans in the classifier. This increased the predictive performance of the model to a hazard ratio of 141.45 with a p-value of <0.005. In an attempt to further improve the performance the relative timing of each scan was incorporated. This decreased the performance in the associative analysis to a hazard ratio of 32.10 with a p-value of <0.005. Lastly, to reduce overfitting feature selection was performed. This resulted in a drop op performance with a hazard ratio of 2.49 ($p = 0.03$). The results are visualised in figure 4.6 and the corresponding hazard ratios with 95% confidence interval and p-value are shown in table 4.5.

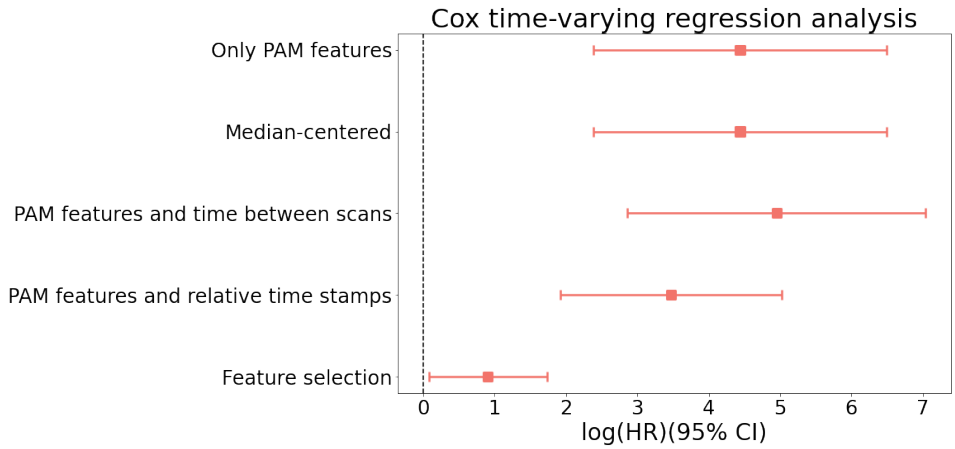


Figure 4.6: Cox time-varying regression analysis for the prognostic AI monitor risk score for various methods of computing the risk score in multisequence scans.

Table 4.5: Hazard ratios for the prognostic AI monitor risk score computed in various ways in multisequence scans.

Method	Hazard ratio	95% CI	p-value
Only PAM features	84.34	[10.76 - 661.13]	<0.005
Center PAM score around median	84.34	[10.76 - 661.13]	<0.005
PAM features and time between scans	141.45	[17.55 - 1140.18]	<0.005
PAM features and relative time stamps	32.10	[6.82 - 151.16]	<0.005
Feature selection	2.49	[1.09 - 5.70]	0.03

Further subanalyses and analysis of confounders were done for the risk score as predicted with the PAM score and relative time stamps. The RFC in this method was trained on a total of 1029 scan pairs of 288 patients and the test set consisted of 940 scan pairs from 289 patients. The outcome of the Cox time-varying regression model on the test set is presented in figure 4.7. The corresponding hazard ratios with 95% confidence interval and p-value are visible in table 4.6. Only the PAM score and whole brain radiation therapy show a significant association with survival ($p < 0.05$).

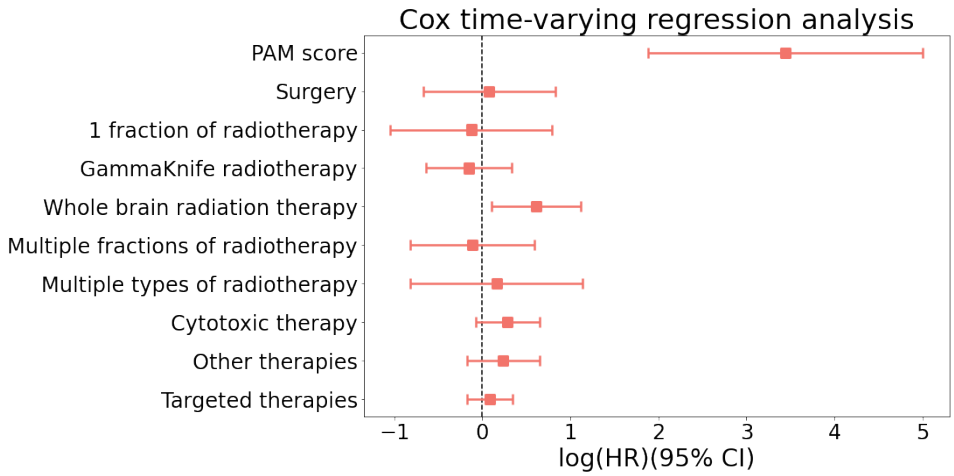


Figure 4.7: Cox time-varying regression analysis for the prognostic AI monitor risk score. Different treatment options were used as confounders in this analysis.

Table 4.6: Overview of the hazard ratios, 95% confidence interval and p-value resulting from Cox time varying analysis for the risk score as predicted by the prognostic AI monitor and various treatment options as possible confounders.

Method	Hazard ratio	95% CI	p-value
PAM score	31.27	[6.57 - 148.87]	<0.005
Surgery	1.08	[0.51 - 2.28]	0.84
One fraction of radiotherapy	0.88	[0.5 - 2.21]	0.79
GammaKnife radiotherapy	0.85	[0.53 - 1.39]	0.53
Whole brain radiation therapy	1.84	[1.11 - 3.07]	0.02
Multiple fractions of radiotherapy	0.89	[0.44 - 1.81]	0.75
Multiple types of radiotherapy	1.17	[0.44 - 3.11]	0.75
Cytotoxic therapy	1.33	[0.93 - 1.91]	0.12
Other therapies	1.26	[0.84 - 1.91]	0.27
Targeted therapies	1.09	[0.84 - 1.41]	0.51

Predictive analysis

The four different methods to compute a risk score were also evaluated on their predictive quality. The train set consists of 1478 scan pairs of 304 patients, the test set consists of 1626 scan pairs of 305 patients. When only using chronological scans with relative time stamps there were 1010 scan pairs of 288 patients in the training set and 959 scan pairs of 289 patients in the test set. The resulting C-indices can be found in table 4.7. The C-index for the best performing RFC trained on the complete dataset with all cancer types is 0.62. A further sub-analysis on the best performing method for different cancer types is provided in table 4.8.

Table 4.7: C-indices for trained random forest classifiers with median-centered risk scores and with and various ways of feeding time information to the random forest classifier.

Method	C-index (overall survival)	C-index (1-year survival)
Only PAM features	0.56	0.58
Center PAM score around median	0.56	0.58
PAM features and time between scans	0.57	0.58
PAM features and relative time stamps	0.60	0.62
Feature selection	0.54	0.55

Table 4.8: C-indices for the predicted risk score for the three most common cancer types in the dataset. These cancer types are brain, bronchus and lung and melanoma.

Cancer type	C-index (overall survival)	C-index (1 year surival)
All cancer types	0.60	0.62
Brain, lung and melanoma	0.59	0.62
Brain	0.60	0.67
Bronchus and lung	0.60	0.63
Melanoma	0.58	0.61

The Kaplan Meier survival scores for the three PAM risk categories are provided in figure 4.8. High risk is defined as a PAM risk score between 0.371 and 0.635, medium risk between 0.291 and 0.371 and low risk between 0.166 and 0.291.

Table 4.11 shows medium and high risk and low and high risk significantly differ from each other ($p < 0.05$).

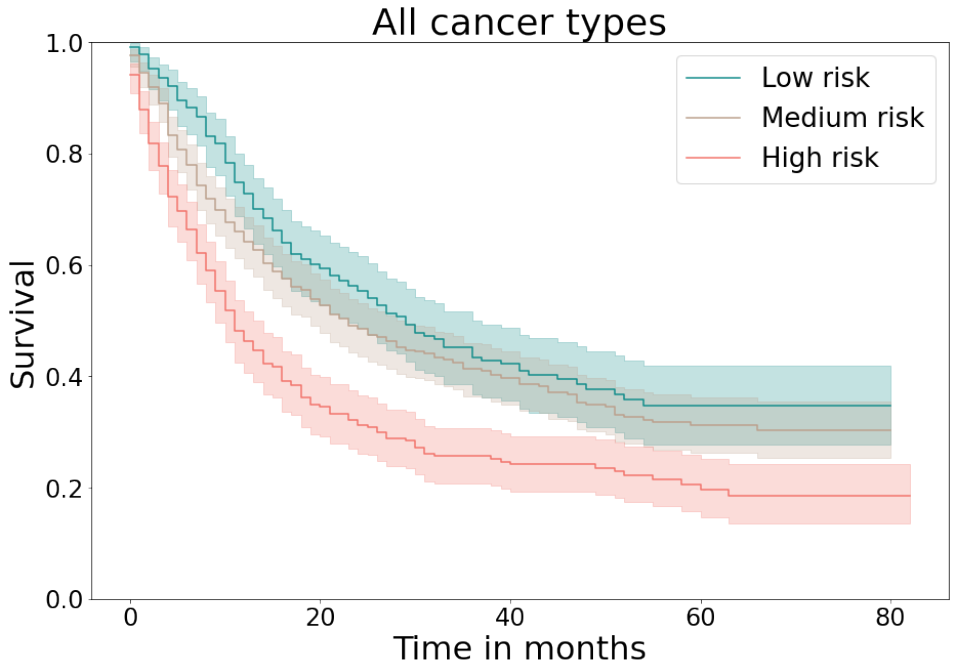


Figure 4.8: Kaplan Meier survival curves for three risk categories based on the risk score provided by the prognostic AI monitor for the complete dataset with all cancer types.

Table 4.9: Results of the logrank test for patients with different cancer types based on the predicted risk scores on multisequence MRI.

Groups	Test statistic	p
Low risk - medium risk	2.28	0.13
Medium risk - high risk	24.14	<0.005
Low risk - high risk	32.84	<0.005

Primary malignancy of the brain

Associative analysis

When training the RFC with the PAM score and relative timing of scans solely on patients with primary brain tumors the results differ from above. The training set consists of 204 scan pairs from 54 patients and the test set contains 218 scan pairs of 54 patients. The Cox-time varying regression analysis for

these patients can be found in figure 4.9 with the corresponding hazard ratios in table 4.10. Only the PAM risk score shows a significant association with survival.

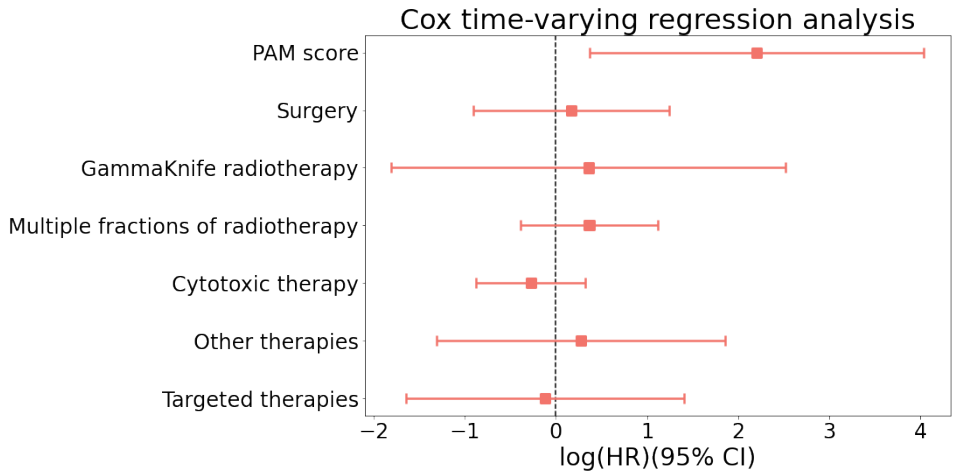


Figure 4.9: Cox time-varying regression analysis for the prognostic AI monitor risk score and confounders for patients with a primary brain tumor. The confounders are the various treatment options of the patient.

Table 4.10: Overview of the hazard ratios, 95% confidence interval and p-value resulting from Cox time varying analysis for the risk score as predicted by the prognostic AI monitor and various treatment options as possible confounders.

Method	Hazard ratio	95% CI	p-value
PAM score	9.08	[1.46 - 58.70]	0.02
Surgery	1.19	[0.40 - 3.48]	0.76
GammaKnife radiotherapy	1.43	[0.16 - 12.44]	0.74
Multiple fractions of radiotherapy	1.44	[0.68 - 3.07]	0.34
Cytotoxic therapy	0.76	[0.42 - 1.38]	0.37
Other therapies	1.32	[0.27 - 6.43]	0.73
Targeted therapies	0.89	[0.19 - 4.09]	0.88

Predictive analysis

The C-index corresponding to the PAM risk score when the RFC is trained on multisequence images of patients with a primary malignancy of the brain is 0.62 when looking at total survival. When using the cut-off of 1 year survival

the C-index equals 0.63. This RFC was trained on 204 scan pairs from 54 patients and the test set consists of 218 scan pairs of 54 patients. The Kaplan Meier survival scores for the three PAM risk categories are provided in figure 4.10. High risk is defined as a PAM risk score between 0.476 and 0.894, medium risk between 0.192 and 0.476 and low risk between 0.00621 and 0.192. These cut-offs were determined based on the complete dataset of patients with primary malignancy of the brain, so the test set does not have an even distribution. There are 23 scan pairs in the low risk group, 121 in the medium risk group and 74 in the low risk group. This difference in group sizes explains the different sizes of confidence intervals. Table 4.11 shows medium and high risk and low and high risk groups significantly differ from each other ($p < 0.05$).

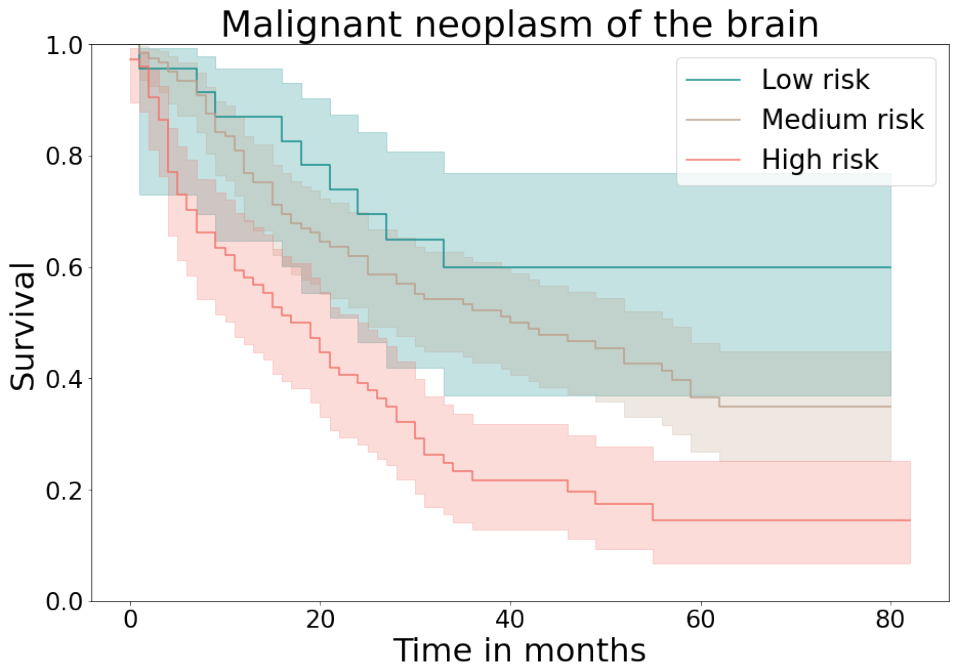


Figure 4.10: Kaplan Meier survival curves for three risk categories based on the risk score provided by the prognostic AI monitor. The prognostic AI monitor was trained on multisequence images of patients with primary brain cancer.

Table 4.11: Results of the logrank test for patients with primary brain cancer based on the predicted risk scores on multisequence MRI.

Groups	Test statistic	p
Low risk - medium risk	2.06	0.15
Medium risk - high risk	18.36	<0.005
Low risk - high risk	12.10	<0.005

4.3.2 External validation with data from Deventer Ziekenhuis

An external validation was performed on 248 scan pairs from 114 patients from Deventer Ziekenhuis. The RFC trained on the PAM score and relative timing of scans of the NKI data was applied to the external dataset from Deventer Ziekenhuis.

Associative analysis

Cox regression was used to explore the association between the PAM score and possible treatments and patient survival. The outcome of this model on the external validation set from Deventer Ziekenhuis is presented in figure 4.11. The corresponding hazard ratios, 95% confidence interval (CI) and p-values can be found in table 4.12. Both the PAM score and chemotherapy show a significant association with survival.

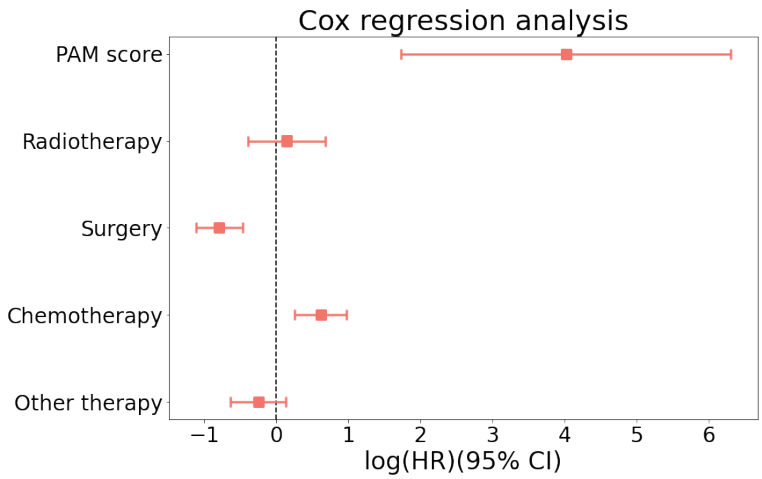


Figure 4.11: Cox regression analysis for the prognostic AI monitor risk score. Different treatments were used as confounders in this analysis.

Table 4.12: Overview of the hazard ratios, 95% confidence interval and p-value resulting from Cox regression analysis for the risk score as predicted by the prognostic AI monitor and various treatment options as possible confounders.

Method	Hazard ratio	95% CI	p-value
PAM score	55.54	[5.61 - 549.54]	<0.005
Radiotherapy	1.15	[0.67 - 1.98]	0.61
Surgery	0.45	[0.33 - 12.44]	<0.005
Chemotherapy	1.84	[1.29 - 2.65]	<0.005
Other therapies	0.78	[0.53 - 1.14]	0.20

Predictive analysis

Predictive analysis was done for the same dataset as the associative analysis. C-indices were computed for the complete dataset and various subsets of cancer types. The results can be found in 4.13. The C-index is highest (0.58) in brain cancer patients when looking at the end-point of 1 year survival.

Table 4.13: C-indices for the predicted risk score for the three most common cancer types in the dataset. These cancer types are brain, bronchus and lung and breast.

Cancer type	C-index (overall survival)	C-index (1 year surival)
All cancer types	0.55	0.55
Brain, lung and breast	0.55	0.54
Brain	0.57	0.58
Bronchus and lung	0.48	0.48
Breast	0.51	0.48

The Kaplan Meier survival curves for three PAM risk categories are provided in figure 4.12. The same cut-offs for high, medium and low risk were used as for the NKI dataset. This resulted in 163 patients in the high risk group, 79 patients in the medium risk group and only 6 patients in the low risk group. The small group sizes lead to larger confidence intervals as visible in figure 4.12. The results of the logrank test in table 4.14 show only the difference between the medium and high risk group is statistically significant.

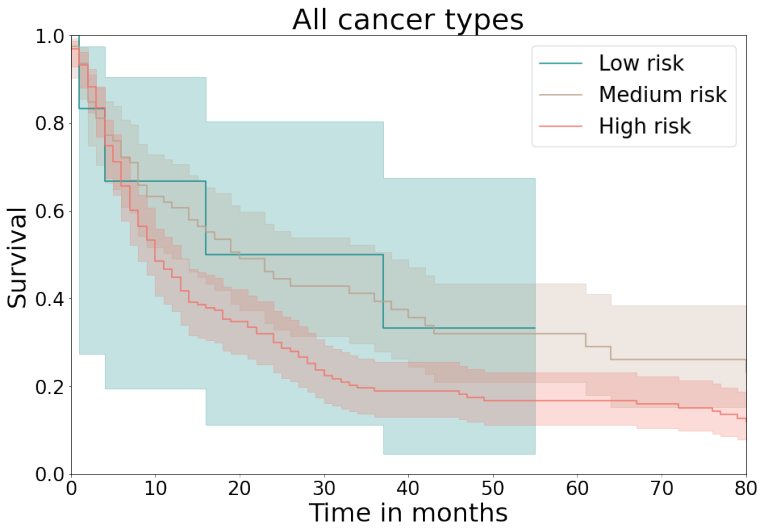


Figure 4.12: Kaplan Meier survival curves for three risk categories based on the risk score provided by the prognostic AI monitor.

Table 4.14: Results of the logrank test for patients from Deventer Ziekenhuis based on the predicted risk scores on multisequence MRI.

Groups	Test statistic	p
Low risk - medium risk	0	0.99
Medium risk - high risk	6.17	0.01
Low risk - high risk	0.85	0.36

4.4 Discussion

In this chapter we investigated the extension of PAM to multisequence MRI and analyzed its performance for survival prediction. The method is also validated with an external dataset. We see quite promising first results in both associative and predictive statistical analysis.

The associative analysis shows the risk score as predicted by PAM has a significant association with survival in both the model trained on all cancer types and the model trained specifically for primary malignancies of the brain. The broad confidence interval of the hazard ratio of the predicted scores shows there is still an uncertainty in the predictions by PAM. The hazard ratio of the classifier trained on patients with primary malignancy of the brain is lower than the classifier trained on all cancer types. This could possibly be caused by

different overlapping sequences and a smaller dataset. The only requirement for PAM is that a scan pair has at least 1 overlapping sequence, but not what sequence this is. Additionally, all sequences share the same weight in computing the loss. However, the T1 postcontrast and the FLAIR sequence might be most clinically relevant and are also used in the RANO criteria.[22] Building upon this theory it might also be worth considering including functional rather than only structural imaging of the brain. Research has shown diffusion weighted imaging, magnetic resonance spectroscopy and perfusion MR imaging can aid in predicting the differential diagnosis of space-occupying lesions within the brain.[85] These sequences will be technically more difficult to incorporate, but should definitely be considered in future improvement of PAM.

The predictive analysis shows the highest C-index for training the random forest classifier with the PAM features and relative time stamps of the scans. When looking at the various cancer types the C-index is highest for patients with primary malignancy of the brain when looking at the 1 year survival. This can be explained by the total patient tumor burden being present in the brain MRI. The performance is worst for melanoma patients. The Kaplan Meier curves in the predictive analysis show that for all cancer types PAM performs particularly well in extracting the high risk patients. In the previous chapter using only the FLAIR sequence showed better performance in differentiation between medium and low risk patients. This finding suggest the different sequences provide complementary information. If a change is visible across all sequences this is usually a worse prognosis. For example an increase in contrast enhancement and increased cerebral edema with a mass effect on the brain.[86] When a tumor is responding to treatment these changes are often more controlled in the sense they are limited to the tumor itself. The ability of the network to distinguish between medium and low risk patients still needs improvement. When looking at PAM trained and evaluated on primary brain cancer only it is visible that it is mainly able to isolate the high risk patients as well. It can be argued that extracting the high risk patients is the most clinically relevant group. When there is disease progression it means the tumor is not responding to the current line of treatment. This would be a reason for a switch of treatment. However, when a patient is stable or responding the clinical consequence would be the same. In both cases the current line of treatment will be continued following the initially planned scheme.

The results of the statistical analysis of the external dataset are quite promising. Given there was no time varying information about the treatment confounders only a normal Cox regression could be used. The normal Cox regression analysis has less power than the time-varying analysis used for the data from AvL/NKI.

However, it does show a significant association between the PAM risk score and patient survival. The C-indices are lower than they were for the NKI dataset. There could be various reasons for this including a shift in MRI intensities, an overfitting of the RFC for the AvL/NKI data or a difference in patient outcomes between the AvL/NKI and Deventer Ziekenhuis. Structural MRI is known for its challenges in inter-site comparability. Scanner characteristics such as the magnetic field strength, design of sequences and the RF-receiving coil may all introduce variability in the resulting MR images.[87] However, the preprocessing algorithm as designed in chapter 3 was built to ensure inter-scanner and inter-site reproducability. The features extracted directly from the MR images are comparable for both the AvL/NKI data and the data from Deventer Ziekenhuis. This supports the assumption that the preprocessing algorithm substantially boosts the inter-site comparability. A more likely reason for the lower C-index is the differences in patient characteristics between Deventer Hospital and the AvL/NKI. In the AvL/NKI dataset the three most common cancer types were primary brain, melanoma and bronchus and lung while in the dataset from Deventer Ziekenhuis the three most common cancer types were primary brain, bronchus and lung and breast. This difference in cancer type is also reflected in patient outcome. The percentage of patients alive after one-year is 57% for the AvL/NKI data while it is only 38% for the Deventer Ziekenhuis data. The Kaplan Meier curves further supports this as there are only six scan pairs classified in the low risk category based on the PAM score. This evidence suggest larger and more heterogeneous dataset could further boost the performance of PAM. Including cancer type, either as an extra feature or by training specific random forest classifiers per cancer type is also worth investigating in future research.

One of the potential pitfalls of the prognostic AI monitor is the fact that it only looks at change between scans. Based on the input dataset it should learn what changes are positive and what changes are negative for the prognosis of the patient. This concept works well when a patient shows progressive disease in a scan pair. However, when the scan remains more or less stable afterwards the risk score computed by PAM will be low, because there are no major changes on the scan. This same concept applies to the RANO criteria where such scan would be classified as stable disease. The radiologist would be able to deduce the scan may be stable, but the prognosis of the patient is still very poor. However, if progression stops and the next scan remains stable this does generally mean a better prognosis than with continued progression. It may be the first indicator of a new line of treatment working. This is also why a radiologist always compares a scan to the previous scan. The before mentioned considerations are all nuances a radiologist is able to notice, but PAM cannot

at this point in time.

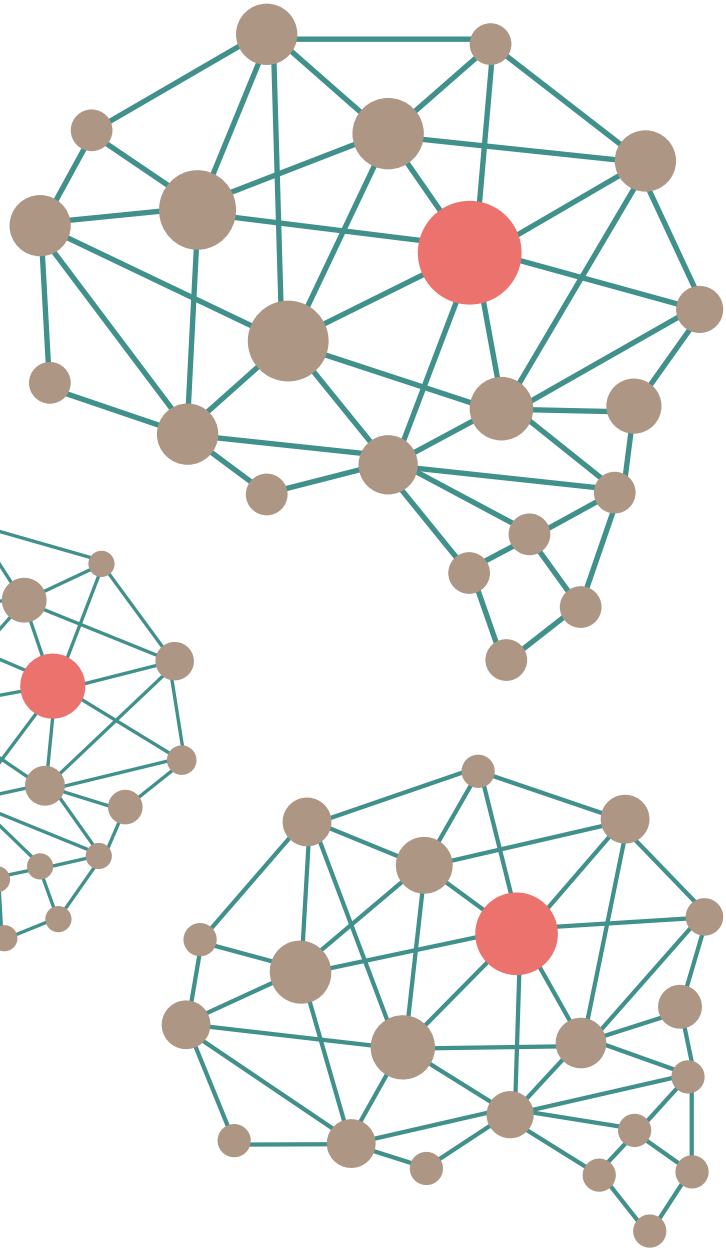
Another point worth considering is if a prognostication tool is the most optimal application of PAM. Another option is to use PAM to build a staging tool. Prognostication is more difficult as it is influenced by many more factors than what can be seen on imaging. Staging is usually done according to the tumor-node-metastasis (TNM) classification. However, for the brain this is not the case. Reasons for using a different staging system are: size does not matter as much as location and tumor type (T), there are no lymphatics in the brain (N) and brain cancer patients do not usually develop metastases (M).[88] Instead the classification by the world health organisation (WHO) is used. In the WHO grading system a grade between 1 and 4 is assigned based on tumor histology and molecular information.[89] Due to the nature of this classification it is difficult to model this stage using PAM. Deep learning can be used to classify brain tumors to one of the histologic subtypes.[90] However, since PAM is specifically aimed at modelling longitudinal changes in scan pairs it is more suitable to predict progression, stable disease or response to treatment.

Additionally, PAM lacks explainability. It is a black-box concept which outputs a risk score for 1-year survival based on features. These features are not traceable to specific changes in the scan. Therefore it is difficult to validate and explain PAM to clinicians. This is caused by the entangled feature representation that is inherent to PAM as it is build now. Feature disentanglement was tried to make the model more explainable. For this a Hessian penalty was used. It succeeded at making PAM more explainable as deformations were better traceable to a specific anatomic location. However, this coincided with a drop in prognostic performance of PAM.[91] It is worth considering to further improve the explainability of the model, especially before application in the clinics. A more detailed overview of the clinical implementation of PAM in the future will be discussed in chapter 6.

4.5 Conclusion

This work provides the first proof-of-concept of translating the prognostic AI monitor to multisequence images. The risk scores as computed by PAM show a significant association with survival. PAM performs particularly well at extracting the high risk patients in a pancancer dataset as well as PAM trained on primary brain cancer. There are also still opportunities for further enhancing the performance of PAM. These mainly lie in adjusting the weights of the various sequences and making the model more explainable. More research

should be conducted to explore further possibilities of PAM in multisequence imaging.



5

Chapter 5

Comparison with current methods for response monitoring

5.1 Introduction

To be able to frame the potential use of PAM in the clinics, we compare it to the current gold standard. Currently the Response Assessment in Neuro-Oncology criteria (RANO) are most commonly used in research and clinical trials. The RANO criteria aim to objectify response assessment based on patient imaging, but also include some other factors. These include patient performance and treatment. The RANO criteria for high grade glioma are most often used in the clinical research setting and are based on the postcontrast T1 scan, the T2/FLAIR scan, patient status and use of corticosteroids.[22] Table 5.1 shows the requirements for the different response categories in high grade glioma. Noteworthy is the fact that an increase in corticosteroids alone while the patient remains clinically stable is not a reason to classify as progressive disease. This is described as *NA* in the table.

Table 5.1: Response Assessment in Neuro-Oncology criteria for high grade glioma. Based on image characteristics on T1 postcontrast MRI, T2/FLAIR MRI a patient can be classified to one of the four response categories. Adapted from Wen et al.

Criterion	Complete Response	Partial Response	Stable Disease	Progressive Disease
T1 postcontrast MRI	None	$\geq$ 50% $\downarrow$	$<$50% $\downarrow$ to $<$25% $\uparrow$	$\geq$ 25% $\uparrow$
T2/FLAIR MRI	Stable or $\downarrow$	Stable or $\downarrow$	Stable or $\downarrow$	$\uparrow$
New lesion	None	None	None	Present
Corticosteroids	None	Stable or $\downarrow$	Stable or $\downarrow$	NA
Clinical status	Stable or $\uparrow$	Stable or $\uparrow$	Stable or $\uparrow$	$\downarrow$
Required criteria for response category	All	All	All	Any

The RANO criteria, just like RECIST, make a distinction between measurable and non-measurable disease. Measurable disease is defined as contrast-enhancing lesions in two dimensions having clear margins. The two perpendicular diameters should be at least 10 mm each. It must be visible on two or more axial slices with a slice thickness of maximal 5 mm and no interslice gaps. Tumors around a cyst or surgical cavity present a challenge for accurate measurement, so these should be classified as non-measurable disease unless there is a node of more than 10 mm in diameter. Non-measurable lesions are the lesions that do not qualify for measurable disease, because they can only be measured in one direction, have no clear margins or have perpendicular diameters of less than 10 mm. In case there are multiple lesions a minimum of two and a maximum of five lesions should be selected for measurement. Generally the largest lesions are chosen for this given that they allow for reproducible repeated measurements.[92] The increase or decrease as mentioned in the criterion *T1 postcontrast MRI* corresponds to the sum of the products of perpendicular diameters of measurable contrast enhancing lesions in the brain. An example of the RANO measurement for a patient is provided in figure 5.1. In this example there is a 38% decrease in enhancing disease. Assuming all other factors mentioned in table 5.1 remain stable this patient would be assigned to the stable disease response category.

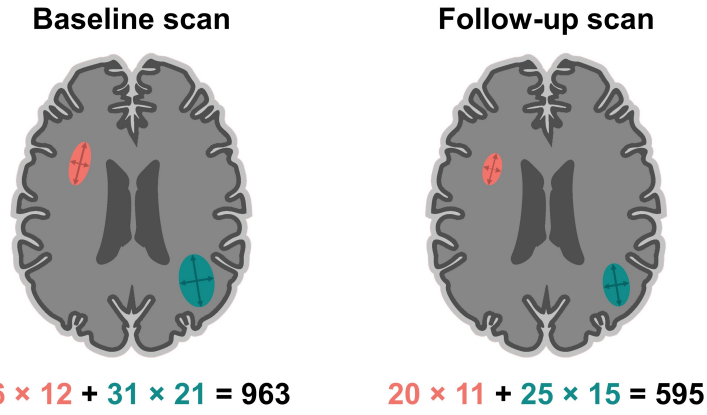


Figure 5.1: Examples of measurements as defined by the Response Criteria in Neuro-Oncology

Worth mentioning is how the RANO criteria try to incorporate pseudoprogression. The proposed criteria from the RANO working group give a few recommendations considering pseudoprogression in patients with chemoradiotherapy. The most import one being that within 12 weeks after completion of chemoradiotherapy a patient can only show progression if there is new or expansion of contrast enhancing lesion beyond the 80% isodose line

used in radiotherapy.[56,92]

Next to the RANO criteria, segmentations and thereby 3D volumetric assessments are an upcoming topic for response monitoring. The RANO criteria are also based on tumor volume, but only measured in 2D. Three dimensional measurements provide a more accurate metric for actual tumor volume. It has been shown that tumor volume in glioblastoma patients is a significant prognostic factor.[93] However, 3D brain tumor segmentation is a very labour and time consuming task. Several methods have been developed for automatic brain tumor segmentation of which nnU-Net is the most promising. This is a deep learning-based segmentation method.[94] There is a version of nnU-Net trained on primary brain tumors in the Brain Tumor Segmentation (BraTS) challenge. This took first place in the competition with dice scores (a measure of overlap) of 88.95, 85.06 and 82.03 for whole tumor, tumor core and enhancing tumor, respectively.[95] This is not applied in a clinical setting yet, but given its promising results and rather user friendly implementation nnU-Net will be included in this research.

5.2 Method

5.2.1 Dataset

It was decided to only include patients with primary brain malignancy in this sub-analysis as these are the most common patients to perform a RANO assessment on and a pretrained nnU-Net exists for this patient group. The RANO assessements and volumetric assessments will eventually be compared with the PAM network trained and evaluated on primary brain images only. The scans in the test set of PAM were manually evaluated and assigned RANO scores by the researcher with support of an experienced radiologist at the NKI. A total of 218 scan pairs of 54 patients were included for RANO and volumetric assessments of the NKI scans. RANO assessments were also performed on 139 scan pairs of 62 patients from Deventer Ziekenhuis.

5.2.2 RANO assessments

RANO assessments were performed using the MRI scans and reports of the scans by radiologists. All scans were systemically approached following these steps:

1. Scrolling trough the FLAIR and T1 postcontrast sequence of the scan and reading the radiologist's report to get a general overview of the scan

2. Check for any new lesions that were not present on the previous scan

3. Define each lesion as measurable or non-measurable
4. Pick the target lesions by choosing the ones that allow for repeated reproducible measurements on the T1 postcontrast scan with a maximum of 5
5. Measure the largest diameter and the largest perpendicular diameter for each target lesion
6. Calculate the sum of the products of the perpendicular diameters
7. Check if there is any significant increase in hyperintensity on the FLAIR scan with help of the radiologist's report
8. Assign scan pair to complete response, partial response, stable disease or progressive disease

Since this dataset was created manually solely based on the MRI scan pairs there is no information about steroid use and clinical status of the patient. The RANO assessment as made in this research therefore only incorporates information that can be directly extracted from the scan. The same applies for the 12-week window of pseudoprogression as discussed before. This is not incorporated in the RANO assessment in this study. The criteria are therefore a simplified version of table 5.1 where only the first three rows are taken into account.

5.2.3 Volumetric assessments

Volumetric assessments will be acquired by using nnU-Net. The MR images were skull stripped using an automated brain extraction tool (BET).[96] In this research inference was ran using the pretrained nnU-Net which placed first in BraTS.[95] No modifications were made to the existing nnU-Net. A segmentation of enhancing tumor, necrosis and peritumoral edema was acquired. The size of these three segmented volumes for baseline and follow-up scan were fed to a random forest classifier to compute a risk score for one year survival between 0 and 1.

5.2.4 Training PAM

The risk scores were computed by the prognostic AI monitor as described earlier. The exact training process, feature extraction and prediction of the risk score can be found in chapter 4.

5.2.5 Statistical analysis

For the dataset from NKI, Kaplan Meier survival curves were plotted for each RANO response category and volumetric based risk score and compared to those resulting from the PAM risk score. C-indices were computed for the volumetric assessments and the PAM risk score. For the external validation set from Deventer Ziekenhuis the PAM score will be compared with the RANO category based on Kaplan Meier survival curves. A logrank test is used to compare the different response categories and risk groups.

5.3 Results

The Kaplan Meier curves for RANO assessments on the dataset from NKI are visible in figure 5.2. The corresponding logrank test is presented in table 5.2. There were 7 scan pairs labeled as partial response, 142 as stable disease and 68 as progressive disease.

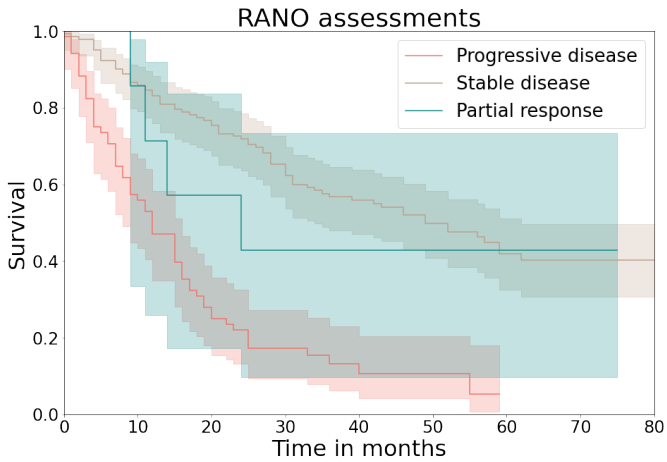


Figure 5.2: Kaplan Meier curves for manual RANO assessments on scan pairs from NKI.

Table 5.2: Results of the logrank test for the response categories as determined by RANO assessments.

Groups	Test statistic	p
Progressive disease - stable disease	61.11	<0.005
Stable disease - Partial response	0.07	0.80
Partial response - Progressive disease	3.92	0.05

The volumetric assessments could only be performed for patients with all four sequences present. This resulted in a training set of 79 scan pairs (26 patients) and a test set of 110 scan pairs (27 patients). 22 scan pairs were categorized as low risk, 49 as medium risk and 39 as high risk. Cut-offs are between 0 and 0.0969 for low risk, between 0.0969 and 0.386 for medium risk and between 0.386 and 1.0 for high risk.

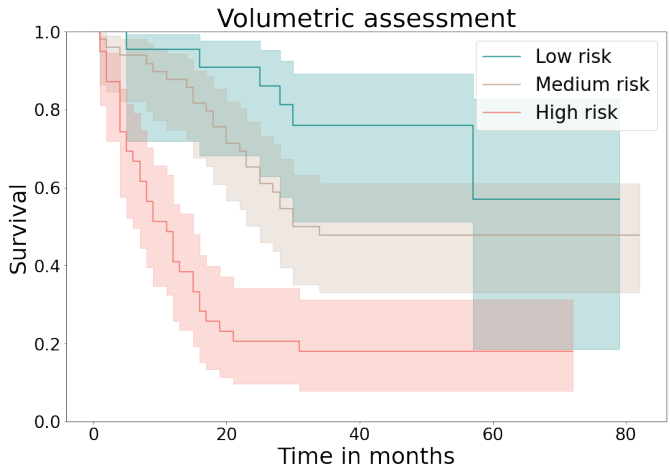


Figure 5.3: Kaplan Meier curves for volumetric assessments on scan pairs from the NKI dataset.

Table 5.3: Results of the logrank test for the risk groups as determined by volumetric assessment.

Groups	Test statistic	p
Low risk - medium risk	3.50	0.06
Medium risk - high risk	18.90	<0.005
Low risk - high risk	19.83	<0.005

The cut-offs used for the risk groups based on the PAM scores are between 0.476 and 0.894 for high risk, between 0.192 and 0.476 for medium risk and between 0.00621 and 0.192 for low risk. There are 23 scan pairs in the low risk group, 121 in the medium risk group and 74 in the low risk group.

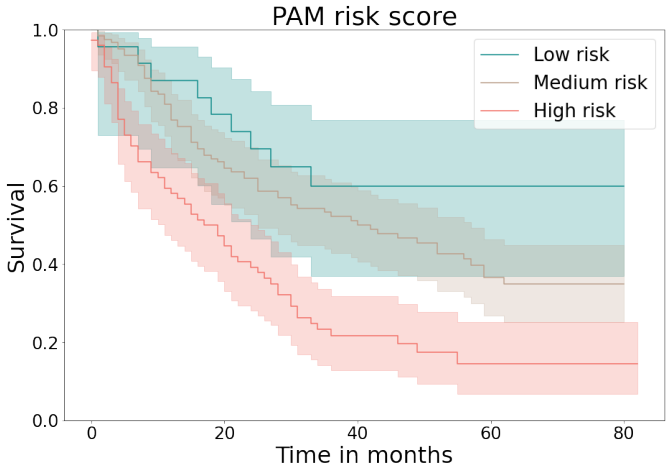


Figure 5.4: Kaplan Meier curves for PAM risk scores on scan pairs from NKI.

Table 5.4: Results of the logrank test for the risk groups as determined by multisequence PAM

Groups	Test statistic	p
Low risk - medium risk	2.06	0.15
Medium risk - high risk	18.36	<0.005
Low risk - high risk	12.10	<0.005

Table 5.5 shows the C-indices resulting from the risk scores based on PAM and on the volumetric assessments. The Pearson correlation between the scores predicted by the volumetric assessments and the scores predicted by PAM equals 0.15.

Table 5.5: C-indices resulting from predictions based on the prognostic AI monitor and on volumetric asssessments

Method	Overall survival	1 year survival
Volumetric assessment	0.77	0.87
Prognostic AI monitoring	0.62	0.63

The Kaplan Meier curves for the RANO assessment on the external validation dataset from Deventer Ziekenhuis are visible in figure 5.5. There is only 1 scan pair classified in the partial response category, so this was left out of the figure. The logrank test returned a test statistic of 37.91 ($p < 0.005$). There are 95 scan pairs labeled stable disease and 45 progressive disease.

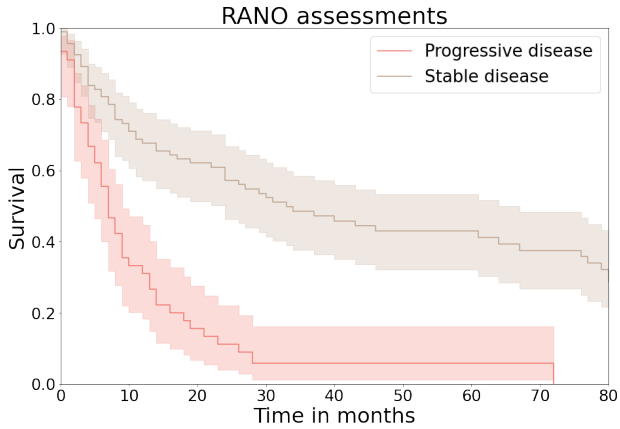


Figure 5.5: Kaplan Meier curves for manual RANO assessments on scan pairs from the Deventer Ziekenhuis.

The PAM scores use cut-offs between 0.248 and 0.483 for low risk, between 0.483 and 0.539 for medium risk and between 0.539 and 0.664 for high risk. This resulted in 48 patients in the high risk group, 47 patients in the medium risk group and 48 patients in the low risk group. Figure 5.6 shows the Kaplan Meier curves. The results of the logrank test can be found in table 5.6

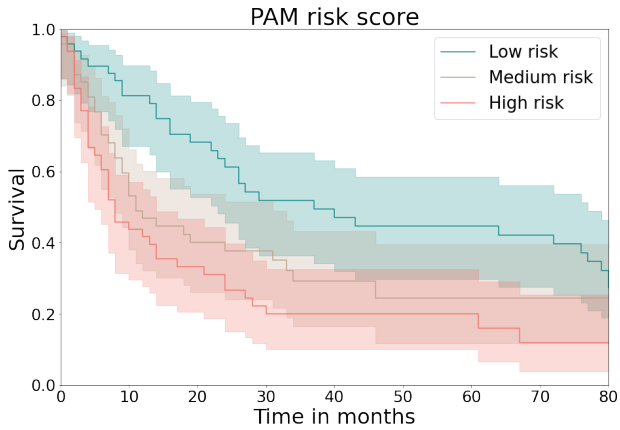


Figure 5.6: Kaplan Meier curves for PAM risk scores on scan pairs from the Deventer Ziekenhuis.

Table 5.6: Results of the logrank test for the risk groups as determined by PAM for the data from Deventer Ziekenhuis.

Groups	Test statistic	p
Low risk - medium risk	2.59	0.11
Medium risk - high risk	2.01	0.16
Low risk - high risk	9.11	<0.005

5.4 Discussion

In this chapter the prognostic AI monitor risk score was compared to the RANO criteria and automated volumetric assessments based on segmentations by nnU-Net. The results demonstrated the volumetric assesment has the best prognostic performance. In the external dataset from Deventer ziekenhuis RANO performed better than PAM. On the dataset from NKI performance of RANO, volumetric assessments and PAM score are all showing significant differences between high and medium risk (for RANO progressive disease and stable disease). Relatively simple improvements of PAM could be incorporated to further improve prognostic performance.

For both datasets the RANO progressive disease groups show a worse prognosis than the PAM high risk group. This finding favours the prognostic performance of RANO over PAM. In the dataset from NKI the stable disease and partial response overlap in the RANO assessments as well as the medium and low risk groups in the PAM risk scores. PAM does have a better performance when looking at the low risk patients which have a better survival than the partial response group in the RANO criteria. In the dataset from Deventer Ziekenhuis there was only one scan pair of one patient showing partial response making it difficult to compare RANO and PAM.

PAM scores are classified into low, medium and high risk groups using the 33rd and 67th percentile. This does not take into account the actual uneven distribution of these risk groups in the dataset. The RANO measurements showed the vast majority of patients is within the stable disease or progressive disease category. Rarely, a patient shows partial response on the brain MRIs in the dataset. Finding other cut off values for the PAM risk scores will therefore further boost prognostic performance.

PAM could also be further improved by incorporating diagnosis or at least by separating high grade from low grade glioma as is done in the RANO criteria. High and low grade glioma have very different prognoses. The median survival

of high grade glioma is 18 months[97] while the median surival of low grade glioma is 13 years[98]. Imaging characteristics also differ between high and low grade glioma. The RANO criteria for low grade glioma are solely based on the T2/FLAIR sequence while the RANO criteria for high grade glioma mainly use the postcontrast T1 sequence. Distinguishing between these two types of primary brain cancer will likely further boost prognostic performance.

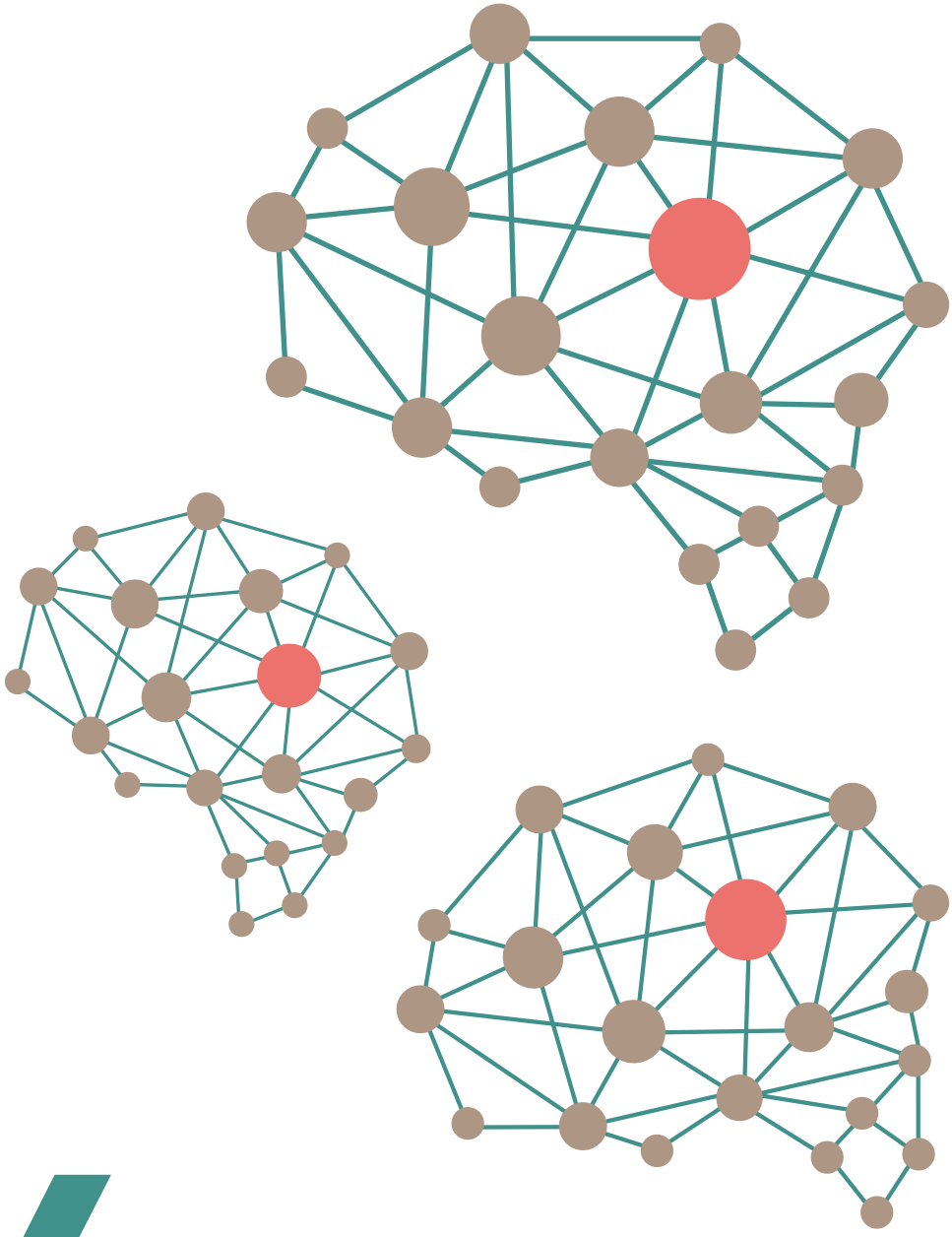
The RANO criteria as used in this chapter are a simplified version of the actual RANO criteria. The corticosteroid use and clinical status of the patient are not always available in the dataset. The way the RANO criteria try to incorporate pseudoprogression was not a workable situation so this was not applied in this research. Both the exact dates of chemotherapy cycles and the 80% isodose line were not available for all patients. This missing data has influenced the RANO assessments, but most likely would not have caused any major differences.

During this research all RANO assessments were done by a single non-expert reader. RANO criteria are known to have a large inter-user variability[28] which is not present in the current research. This might lead to the RANO measurements as presented in this chapter to give an overestimation of prognostic performance. Furthermore, RANO assessments are labour intensive and time consuming. RANO criteria are the current research standard to quantify longitudinal changes in the scans. However, reading through the radiologist's reports rarely showed RANO measurements. This illustrates the limited clinical use of the RANO criteria. In comparison to the RANO criteria PAM is easier to implement clinically and can therefore be used on more patients.

The volumetric assessments, like the RANO assessments, show a good cut-off between medium and high risk scan pairs. Since nnU-Net only works when a FLAIR, T2, T1 precontrast and T1 postcontrast sequence are present there are less patients included in this study. The group size is roughly twice as small as with the other methods. This inevitably leads to broader confidence intervals which makes it more difficult to draw conclusions from this data. This method shows to be very promising. The need for all four sequences is a disadvantage when comparing the volumetric assessments to PAM, but its prognostic performance is promising reaching an higher C-index than PAM. The correlation between the scores predicted by the volumetric assessments and by PAM is low, meaning they are likely complementary to each other. The prognostic value of the volumetric assessments should definitely be considered and might even be combined with PAM in future research.

5.5 Conclusion

The overall prognostic performance of PAM, volumetric assessments and the RANO criteria is similar in this research as they all show significant separation of high and medium risk patients. The external validation set showed a superior performance of RANO when compared to PAM. Further improvements of PAM are needed to reach an equal prognostic performance across datasets. This would lead the way for a more effortless and less time consuming clinical implementation of longitudinal response monitoring.



6

Chapter 6

General conclusion and future outlook

6.1 General conclusion

This thesis has provided the evidence for an alternative method of longitudinal monitoring of cancer patients based on brain MRI. The prognostic AI monitor, a deep learning based image registration algorithm, is able to extract prognostic features from MRI.

Chapter 3 showed the necessary preprocessing steps to reduce inter-site and inter-scanner variability. It showed a successful conversion of PAM from CT to single sequence MRI while maintaining its prognostic value. Especially in patients with primary brain tumours PAM is able to adequately categorize patients in low, medium and high risk groups. Future research should focus at improving the implementation of a patient's timeline in the algorithm instead of using each scan pair as a separate entity. Chapter 4 extended the research to multisequence imaging of the brain. This provided the first proof of concept of translating PAM to multisequence MRI. The risk scores as computed by the prognostic AI monitor show a significant association with patient survival. These scores showed to be particularly useful in extracting high risk patients. Further opportunities for improvement lie in optimizing the weights of the various sequences, improving prognostic performance by experimenting with other methods to predict survival based on features and making the model more explainable. Chapter 5 compared PAM to the currently used RANO criteria and automated volumetric assessment using nnU-Net. All three methods show significant differences between patients with progressive and stable disease. Adequate extractions of patients with response remains a challenge. On the external validation set the RANO criteria performed superior to PAM. The mentioned improvements throughout this thesis should be researched to further boost performance and pave the way for a new clinical implementation of

response monitoring.

6.2 Future outlook

There are various technical improvements already elaborately discussed in this thesis, but this future outlook will focus on what happens after prognostic performance reaches a similar or higher level than RECIST/RANO.

The first steps are to expand MRI PAM to other organs. This is a relatively simple step, but will increase involvement and support from other specialties. When developing the preprocessing method for brain MRI this was already kept in mind. It was designed as a generic algorithm that is able to preprocess all structural MRI sequences. The advantage of the brain is that it is generally very static. A change in the scan is often caused by pathology. This does not necessarily apply to other organs as well. The conversion to MRI of other organs with more complex anatomy will therefore be done in steps. The first step will be to extend PAM to MRI of the liver. The next step will be to extend PAM to MRI of the lower pelvis. These steps show increasingly more non-pathological changes in scans over time, for example slight displacements of the small intestine between scans. It will require extra attention to adjust PAM to make a distinction between physiological and pathological changes in a follow-up scan.

One of the advantages of PAM is the fact that it is a personalized approach to response monitoring. It makes a survival prediction based on the imaging characteristics of an individual patient. Currently survival is used as a substitute for treatment response, because it can be easily obtained and allows for comparison with other methods. However, this is not the most optimal metric for clinical use as it does not account for deaths unrelated to cancer or treatment. The outcome of PAM should be converted to a more clinically relevant metric. In a clinical setting a patient diagnosed with cancer follows a treatment plan consisting of chemotherapy, immunotherapy, targeted therapy, radiation therapy and/or surgery. Imaging is used to determine the response to treatment and to find possible treatment related side effects. Based on the imaging and the condition of the patient the treating physician can decide to follow initial treatment plan, adjust the treatment plan (e.g. lower dosage due to adverse effects) or stop treatment. There would be increased clinical value if PAM is able to aid in this decision instead of predicting one year survival. This will require a different study set-up with close cooperation of clinicians.

When PAM reaches adequate prognostic performance, meaning equal or better than the current RANO/RECIST criteria, a clinical study could provide meaningful insights into PAM's place in the clinic. The care of a cancer patient is already multidisciplinary with various specialists involved. PAM could be one of the tools they use to gain adequate insights in the treatment response of a patient. To be able for clinicians to use PAM it should be more illustrative and explainable. It should provide information about how a change in the scan contributes to the final output. Additionally, the usability of PAM should be increased. At this point in time PAM is still only available in a research setting, meaning it can only be used by people with programming knowledge. A graphical user interface should be developed to allow users to interact with PAM in a more simple manner. Ideally, the outcome of PAM could then be illustrated next to the scan on tumor board meetings. That way all clinicians involved in the treatment of the patients can take into account the additional information provided by PAM. Its additional value in such setting should be determined in a clinical prospective multi-center study.

All in all, this thesis has shown that AI based image registration can be used to extract prognostic features from brain MRI. It described the first steps in possibly improving care of individual cancer patients and provides a foundation for further studies.

Bibliography

[1] F. Michor, Y. Iwasa, and M.A. Nowak. Dynamics of cancer progression. *Nature Reviews Cancer*, 4(3):197–205, 2004. doi: 10.1038/nrc1295.

[2] D. Hanahan and R.A. Weinberg. The Hallmarks of Cancer. *Cell*, 100(1):57–70, 2000. doi: 10.1016/S0092-8674(00)81683-9.

[3] D. Hanahan and R.A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, 2011. doi: 10.1016/J.CELL.2011.02.013/ATTACHMENT/3F528E16-8B3C-4D8D-8DE5-43E0C98D8475/MMC1.PDF.

[4] C.A. Klein. Cancer progression and the invisible phase of metastatic colonization. *Nature Reviews Cancer*, 20(11):681–694, 2020. doi: 10.1038/s41568-020-00300-6.

[5] V.T. DeVita, T.S. Lawrence, and S.A. Rosenberg. *Cancer: Principles & Practice of Oncology*. Wolters Kluwer Health, Philadelphia, 10 edition, 2015.

[6] D. Abshire and M.K. Lang. The Evolution of Radiation Therapy in Treating Cancer. *Seminars in Oncology Nursing*, 34(2):151–157, 2018. doi: 10.1016/J.SONCN.2018.03.006.

[7] M. Shields. Chemotherapeutics. In Simone Badal and Rupika Delgoda, editors, *Pharmacognosy*, pages 295–313. Academic Press, Boston, 2017. doi: 10.1016/B978-0-12-802104-0.00014-7.

[8] P. Nygren. What is cancer chemotherapy? *Acta Oncologica*, 40(2-3):166–174, 2001. doi: 10.1080/02841860151116204.

[9] A.D. Waldman, J.M. Fritz, and M.J. Lenardo. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nature Reviews Immunology*, 20(11):651–668, 2020. doi: 10.1038/s41577-020-0306-5.

[10] C. Sawyers. Targeted cancer therapy. *Nature*, 432(7015):294–297, 2004. doi: 10.1038/nature03095.

[11] K. I. Pritchard. Endocrine therapy: is the first generation of targeted drugs the last? *Journal of Internal Medicine*, 274(2):144–152, 2013. doi: 10.1111/JOIM.12065.

[12] N. A. Seebacher, A. E. Stacy, G. M. Porter, and A. M. Merlot. Clinical development of targeted and immune based anti-cancer therapies. *Journal of Experimental & Clinical Cancer Research*, 38(1):1–39, 2019. doi: 10.1186/S13046-019-1094-2.

[13] L. Fass. Imaging and cancer: A review. *Molecular Oncology*, 2(2):115–152, 2008. doi: 10.1016/J.MOLONC.2008.04.001.

[14] S.G. Tandel, M. Biswas, O.G. Kakde, A. Tiwari, H.S. Suri, M. Turk, J.R. Laird, C.K. Asare, A.A. Ankrah, N.N. Khanna, B.K. Madhusudhan, L. Saba, and J.S. Suri. A review on a deep learning perspective in brain cancer classification. *Cancers*, 11(1):111, 2019. doi: 10.3390/cancers11010111.

[15] M.L. Specchia, E.M. Frisicale, E. Carini, A. Di Pilla, D. Cappa, A. Barbara, W. Ricciardi, and G. Damiani. The impact of tumor board on cancer care: Evidence from an umbrella review. *BMC Health Services Research*, 20(1):1–14, 2020. doi: 10.1186/S12913-020-4930-3/TABLES/3.

[16] B.W. Lamb, K.F. Brown, K. Nagpal, C. Vincent, J.S.A. Green, and N. Sevdalis. Quality of care management decisions by multidisciplinary cancer teams: A systematic review. *Annals of Surgical Oncology*, 18(8):2116–2125, 2011. doi: 10.1245/S10434-011-1675-6/TABLES/5.

[17] J. Prades, E. Remue, E. van Hoof, and J.M. Borras. Is it worth reorganising cancer services on the basis of multidisciplinary teams (MDTs)? A systematic review of the objectives and organisation of MDTs and their impact on patient outcomes. *Health Policy*, 119(4):464–474, 2015. doi: 10.1016/J.HEALTHPOL.2014.09.006.

[18] T. Soukup, B.W. Lamb, S. Sarkar, S. Arora, S. Shah, A. Darzi, J.S.A. Green, and N. Sevdalis. Predictors of Treatment Decisions in Multidisciplinary Oncology Meetings: A Quantitative Observational Study. *Annals of Surgical Oncology*, 23(13):4410–4417, 2016. doi: 10.1245/S10434-016-5347-4/TABLES/3.

[19] P. Therasse, S.G. Arbuck, E.A. Eisenhauer, J. Wanders, R.S. Kaplan, L. Rubinstein, J. Verweij, M. Van Glabbeke, A.T. van Oosterom, M.C. Christian, and S.G. Gwyther. New guidelines to evaluate the response to treatment in solid tumors. *Journal of the National Cancer Institute*, 92(3):205–216, 2000. doi: 10.1093/jnci/92.3.205.

[20] E.A. Eisenhauer, P. Therasse, J. Bogaerts, L.H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *European Journal of Cancer*, 45(2):228–247, 2009. doi: 10.1016/j.ejca.2008.10.026.

[21] L. Seymour, J. Bogaerts, A. Perrone, R. Ford, L.H. Schwartz, S. Mandrekar, N.U. Lin, S. Litière, J. Dancey, A. Chen, F.S. Hodi, P. Therasse, O.S. Hoekstra, L.K. Shankar, J.D. Wolchok, M. Ballinger, C. Caramella, and E.G.E. de Vries. iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. *The Lancet: Oncology*, 18 (3):e143–e152, 2017. doi: 10.1016/S1470-2045(17)30074-8.

[22] U.N. Chukwueke and P.Y. Wen. Use of the Response Assessment in Neuro-Oncology (RANO) criteria in clinical trials and clinical practice. *CNS Oncology*, 8(1), 2019. doi: 10.2217/cns-2018-0007.

[23] P.Y. Wen, T.F. Cloughesy, B.M. Ellingson, D.A. Reardon, H.A. Fine, L. Abrey, K. Ballman, M. Bendszuz, J. Buckner, S.M. Chang, M.D. Prados, W.B. Pope, A. Gregory Sorensen, M. van den Bent, and W.A. Yung. Report of the Jumpstarting Brain Tumor Drug Development Coalition and FDA clinical trials neuroimaging endpoint workshop (January 30, 2014, Bethesda MD). *Neuro-Oncology*, 16(suppl. 7):vii36–vii47, 2014. doi: 10.1093/neuonc/nou226.

[24] R.L. Morgan and D.R. Camidge. Reviewing RECIST in the Era of Prolonged and Targeted Therapy. *Journal of Thoracic Oncology*, 13(2):154–164, 2018. doi: 10.1016/j.jtho.2017.10.015.

[25] S. Peungjesada, H.H. Chuang, S.R. Prasad, H. Choi, E.M. Loyer, and Y. Bronstein. Evaluation of cancer treatment in the abdomen: Trends and advances. *World Journal of Radiology*, 5(3):126, 2013. doi: 10.4329/wjr.v5.i3.126.

[26] Y.S. Chun, J.N. Vauthey, P. Boonsirikamchai, D.M. Maru, S. Kopetz, M. Palavecino, S.A. Curley, E.K. Abdalla, H. Kaur, C. Charnsangavej, and E.M. Loyer. Association of computed tomography morphologic criteria with pathologic response and survival in patients treated with bevacizumab for colorectal liver metastases. *JAMA*, 302(21):2338–2344, 2009. doi: 10.1001/jama.2009.1755.

[27] S. Grimaldi, M. Terroir, and C. Caramella. Advances in oncological treatment: limitations of RECIST 1.1 criteria. *The Quarterly Journal of Nuclear Medicine and Molecular Imaging*, 62(2):129–139, 2018. doi: 10.23736/S1824-4785.17.03038-2.

[28] D. Muenzel, H.P. Engels, M. Bruegel, V. Kehl, E.J. Rummeny, and S. Metz. Intra- and inter-observer variability in measurement of target lesions: implication on response evaluation according to RECIST 1.1. *Radiology and Oncology*, 46(1):8–18, 2012. doi: 10.2478/v10019-012-0009-z.

[29] S.H. Yoon, K.W. Kim, J.M. Goo, D.W. Kim, and S. Hahn. Observer variability in RECIST-based tumour burden measurements: a meta-analysis. *European Journal of Cancer*, 53:5–15, 2016. doi: 10.1016/j.ejca.2015.10.014.

[30] S. Trebeschi, Z. Bodalal, N. van Dijk, T.N. Boellaard, P. Apfaltrer, T.M. Tareco Bucho, T.D.L. Nguyen-Kim, M.S. van der Heijden, H.J.W.L. Aerts, and R.G.H. Beets-Tan. Development of a Prognostic AI-Monitor for Metastatic Urothelial Cancer Patients Receiving Immunotherapy. *Frontiers in Oncology*, 11, 2021. doi: 10.3389/fonc.2021.637804.

[31] D. Hui, C.E. Paiva, E.G. Del Fabbro, C. Steer, J. Naberhuis, M. van de Wetering, P. Fernández-Ortega, T. Morita, S.Y. Suh, E. Bruera, and M. Mori. Prognostication in advanced cancer: update and directions for future research. *Supportive Care in Cancer*, 27(6):1973–1984, 2019. doi: 10.1007/s00520-019-04727-y.

[32] V.P.B. Grover, J.M. Tognarelli, M.M.E. Crossey, I.J. Cox, S.D. Taylor-Robinson, and M.J.W. McPhail. Magnetic Resonance Imaging: Principles and Techniques: Lessons for Clinicians. *Journal of Clinical and Experimental Hepatology*, 5(3):246, 2015. doi: 10.1016/J.JCEH.2015.08.001.

[33] R. Bitar, G. Leung, R. Perng, S. Tadros, A.R. Moody, J. Sarrazin, C. McGregor, M. Christakis, S. Symons, A. Nelson, and T.P. Roberts. MR pulse sequences: What every radiologist wants to know but is afraid to ask. *Radiographics*, 26(2):513–537, 2006. doi: 10.1148/RG.262055063/ASSET/IMAGES/LARGE/G06MR09G20A.JPEG.

[34] J.P. Ridgway. Cardiovascular magnetic resonance physics for clinicians: part I. *Journal of Cardiovascular Magnetic Resonance*, 12(1):71, 2010. doi: 10.1186/1532-429X-12-71.

[35] R. J.M. Van Geuns, P. A. Wielopolski, H. G. De Bruin, B. J. Rensing, P. M.A. Van Ooijen, M. Hulshoff, M. Oudkerk, and P. J. De Feyter. Basic principles of magnetic resonance imaging. *Progress in Cardiovascular Diseases*, 42(2):149–156, 1999. doi: 10.1016/S0033-0620(99)70014-9.

[36] P. Ongsulee. Artificial intelligence, machine learning and deep learning. *International Conference on ICT and Knowledge Engineering*, pages 1–6, 2018. doi: 10.1109/ICTKE.2017.8259629.

[37] Y. Lecun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539.

[38] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61: 85–117, 2015. doi: 10.1016/J.NEUNET.2014.09.003.

[39] I.H. Sarker. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, 2(6):1–20, 2021. doi: 10.1007/S42979-021-00815-1/FIGURES/13.

[40] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. URL `https://www.deeplearningbook.org/`.

[41] K. O'shea and R. Nash. An Introduction to Convolutional Neural Networks. URL `https://arxiv.org/pdf/1511.08458.pdf`.

[42] R. Yamashita, M. Nishio, R.K.G. Do, and K. Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9(4):611–629, 2018. doi: 10.1007/S13244-018-0639-9/FIGURES/15.

[43] S. Trebeschi, Z. Bodalal, T.N. Boellaard, T. M. Tareco Bucho, S.G. Drago, I. Kurilova, A.M. Calin-Vainak, A. Delli Pizzi, M. Muller, K. Hummelink, K.J. Hartemink, T.D.L. Nguyen-Kim, E.F. Smit, H.J.W.L. Aerts, and R.G.H. Beets-Tan. Prognostic Value of Deep Learning-Mediated Treatment Monitoring in Lung Cancer Patients Receiving Immunotherapy. *Frontiers in Oncology*, 11, 2021. doi: 10.3389/fonc.2021.609054.

[44] S. Albawi, T.A. Mohammed, and S. Al-Zawi. Understanding of a convolutional neural network. *Proceedings of 2017 International Conference on Engineering and Technology*, pages 1–6, 2018. doi: 10.1109/ICENGTECHNOL.2017.8308186.

[45] Y. Wu and K. He. Group Normalization. 2018. doi: 10.48550/arxiv.1803.08494.

[46] S. Zhao, T. Lau, J. Luo, and Y. Xu. Unsupervised 3D End-to-End Medical Image Registration with Volume Tweening Network. *IEEE Journal of Biomedical and Health Informatics*, 24(5):1394–1404, 2020. doi: 10.1109/JBHI.2019.2951024.

[47] M.A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015. URL `http://neuralnetworksanddeeplearning.com`.

[48] K. Aldape, K.M. Brindle, L. Chesler, R. Chopra, A. Gajjar, M.R. Gilbert, N. Gottardo, D.H. Gutmann, D. Hargrave, E.C. Holland, D.T.W. Jones, J.A. Joyce, P. Kearns, M.W. Kieran, I.K. Mellinghoff, M. Merchant, S.M. Pfister, S.M. Pollard, V. Ramaswamy, J.N. Rich, G.W. Robinson, D.H. Rowitch, J.H. Sampson, M.D. Taylor, P. Workman, and R.J. Gilbertson. Challenges to curing primary brain tumours. *Nature Reviews Clinical Oncology*, 16(8):509–520, 2019. doi: 10.1038/s41571-019-0177-5.

[49] A.S. Achrol, R.C. Rennert, C. Anders, R. Soffietti, M.S. Ahluwalia, L. Nayak, S. Peters, N.D. Arvold, G.R. Harsh, P.S. Steeg, and S.D. Chang. Brain metastases. *Nature Reviews Disease Primers*, 5(1), 2019. doi: 10.1038/S41572-018-0055-Y.

[50] W. A. Hall, H. R. Djalilian, E. S. Nussbaum, and K. H. Cho. Long-term survival with metastatic cancer to the brain. *Medical oncology*, 17(4):279–286, 2000. doi: 10.1007/BF02782192.

[51] V.A. Venur, V. Karivedu, and M.S. Ahluwalia. Systemic therapy for brain metastases. *Handbook of Clinical Neurology*, 149:137–153, 2018. doi: 10.1016/B978-0-12-811161-1.00011-6.

[52] R.K. Jain. Normalizing Tumor Microenvironment to Treat Cancer: Bench to Bedside to Biomarkers. *Journal of Clinical Oncology*, 31(17):2218, 2013. doi: 10.1200/JCO.2012.46.3653.

[53] M. Vargo. Brain tumor rehabilitation. *American Journal of Physical Medicine and Rehabilitation*, 90(5):50–62, 2011. doi: 10.1097/PHM.0B013E31820BE31F.

[54] S.C. Thust, M. J. van den Bent, and M. Smits. Pseudoprogression of brain tumors. *Journal of Magnetic Resonance Imaging*, 48(3):571–589, 2018. doi: 10.1002/JMRI.26171.

[55] A.I. Mehta, C.W. Kanaly, A.H. Friedman, D.D. Bigner, and J.H. Sampson. Monitoring Radiographic Brain Tumor Progression. *Toxins*, 3(3):191, 2011. doi: 10.3390/TOXINS3030191.

[56] D. J. Leao, P. G. Craig, L. F. Godoy, C. C. Leite, and B. Policeni. Response Assessment in Neuro-Oncology Criteria for Gliomas: Practical Approach Using Conventional and Advanced Techniques. *American Journal of Neuroradiology*, 41(1):10–20, 2019. doi: 10.3174/AJNR.A6358.

[57] K.S. Saini and C. Twelves. Determining lines of therapy in patients with solid cancers: a proposed new systematic and comprehensive framework. *British Journal of Cancer*, 125 (2):155–163, 2021. doi: 10.1038/s41416-021-01319-8.

[58] C. D. Marcus, V. Ladam-Marcus, C. Cucu, O. Bouché, L. Lucas, and C. Hoeffel. Imaging techniques to evaluate the response to treatment in oncology: Current standards and perspectives. *Critical Reviews in Oncology/Hematology*, 72(3):217–238, 2009. doi: 10.1016/J.CRITREVONC.2008.07.012.

[59] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2015. URL `https://arxiv.org/pdf/1409.1556.pdf`.

[60] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv*, 2015. doi: 10.48550/arxiv.1505.04597.

[61] S. Masoudi, S.A. Harmon, S. Mehralivand, S.M. Walker, H. Raviprakash, U. Bagci, P.L. Choyke, and B. Turkbey. Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research. *Journal of Medical Imaging*, 8(1), 2021. doi: 10.1117/1.JMI.8.1.010901.

[62] J. Juntu, J. Sijbers, D. Dyck, and J. Gielen. Bias field correction for MRI images. In *Advances in Soft Computing*, Advances in soft computing, pages 543–551. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. doi: 10.1007/3-540-32390-2{_}64.

[63] J.V. Manjón. MRI Preprocessing. In *Imaging Biomarkers*, pages 53–63. Springer, Cham, 2017. doi: 10.1007/978-3-319-43504-6{_}5.

[64] N.J. Tustison, B.B. Avants, P.A. Cook, Y. Zheng, A. Egan, P.A. Yushkevich, and J.C. Gee. N4ITK: improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29 (6):1310–1320, 2010. doi: 10.1109/TMI.2010.2046908.

[65] F. Navarro, H. Dapper, R. Asadpour, C. Knebel, M.B. Spraker, V. Schwarze, S.K. Schaub, N.A. Mayr, K. Specht, H.C. Woodruff, P. Lambin, A.S. Gersing, M.J. Nyflot, B.H. Menze, S.E. Combs, and J.C. Peeken. Development and External Validation of Deep-Learning-Based Tumor Grading Models in Soft-Tissue Sarcoma Patients Using MR Imaging. *Cancers*, 13(12):2866, 2021. doi: 10.3390/cancers13122866.

[66] V. Shreyas and V. Pankajakshan. A deep learning architecture for brain tumor segmentation in MRI images. *2017 IEEE 19th International Workshop on Multimedia Signal Processing*, pages 1–6, 2017. doi: 10.1109/MMSP.2017.8122291.

[67] M.U. Dalmış, G. Litjens, K. Holland, A. Setio, R. Mann, N. Karssemeijer, and A. Gubern-Mérida. Using deep learning to segment breast and fibroglandular tissue in MRI volumes. *Medical Physics*, 44(2):533–546, 2017. doi: 10.1002/mp.12079.

[68] U. Vovk, F. Pernuš, and B. Likar. A review of methods for correction of intensity inhomogeneity in MRI. *IEEE Transactions on Medical Imaging*, 26(3):405–421, 2007. doi: 10.1109/TMI.2006.891486.

[69] B.C. Lowekamp, D.T. Chen, L. Ibáñez, and D. Blezek. The design of simpleITK. *Frontiers in Neuroinformatics*, 7:45, 2013. doi: 10.3389/FNINF.2013.00045/BIBTEX.

[70] M. Shah, Y. Xiao, N. Subbanna, S. Francis, D.L. Arnold, D.L. Collins, and T. Arbel. Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. *Medical Image Analysis*, 15(2):267–282, 2011. doi: 10.1016/J.MEDIA.2010.12.003.

[71] S. Valverde. MRI Intensity Normalization, 2019. URL `https://github.com/sergivalverde/MRI_intensity_normalization`.

[72] K.M. Ho. Effect of non-linearity of a predictor on the shape and magnitude of its receiver-operating-characteristic curve in predicting a binary outcome. *Scientific Reports*, 7(1): 1–7, 2017. doi: 10.1038/s41598-017-10408-9.

[73] R. Thawani, K. Fakhoury, and K. Becker. Cause of mortality in patients with lung cancer and brain metastasis. *Journal of Clincal Oncology*, 38(15):e21743–e21743, 2020. doi: 10.1200/JCO.2020.38.15{_}SUPPL.E21743.

[74] H. Ishwaran, U.B. Kogalur, E.H. Blackstone, and M.S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008. doi: 10.1214/08-AOAS169.

[75] J.L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018. doi: 10.1186/s12874-018-0482-1.

[76] M. Geurts, W. Taal, and M. van den Bent. Richtlijn Gliomen. Technical report, Erasmus MC, Rotterdam, 2020. URL `https://www.erasmusmc.nl/-/media/erasmusmc/pdf/1-themaspecifiek/neurologie-richtlijnen/gliomen.pdf`.

[77] J.E. Villanueva-Meyer, M.C. Mabray, and S. Cha. Current Clinical Brain Tumor Imaging. *Neurosurgery*, 81(3):397–415, 2017. doi: 10.1093/neuros/nyx103.

[78] C. Westbrook and C. Kaut. *Image Weighting and Contrast*. Blackwell Science Ltd, Oxford, 2 edition, 2000. ISBN 0632042052.

[79] K. Usman and K. Rajpoot. Brain tumor classification from multi-modality MRI using wavelets and machine learning. *Pattern Analysis and Applications*, 20(3):871–881, 8 2017. doi: 10.1007/S10044-017-0597-8/FIGURES/8.

[80] A. Sengupta, S. Agarwal, P.K. Gupta, S. Ahlawat, R. Patir, R.K. Gupta, and A. Singh. On differentiation between vasogenic edema and non-enhancing tumor in high-grade glioma patients using a support vector machine classifier based upon pre and post-surgery MRI images. *European Journal of Radiology*, 106:199–208, 2018. doi: 10.1016/J.EJRAD.2018.07.018.

[81] P. W. Schaefer, R.F. Budzik, and R.G. Gonzalez. Imaging of Cerebral Metastases. *Neurosurgery Clinics of North America*, 7(3):393–423, 1996. doi: 10.1016/S1042-3680(18)30369-3.

[82] E.J. Escott. A variety of appearances of malignant melanoma in the head: A review. *Radiographics*, 21(3):625–639, 2001. doi: 10.1148/radiographics.21.3.g01ma19625.

[83] M.C. Mabray, R.F. Barajas, and S. Cha. Modern Brain Tumor Imaging. *Brain Tumor Research and Treatment*, 3(1):8, 2015. doi: 10.14791/BTRT.2015.3.1.8.

[84] C. Shen, H.R. Roth, H. Oda, M. Oda, Y. Hayashi, K. Misawa, and K. Mori. On the influence of Dice loss function in multi-class organ segmentation of abdominal CT using 3D fully convolutional networks. 2018. doi: 10.48550/arxiv.1801.05912.

[85] B. Aydın, H. Aydın, E. Birgi, and B. Hekimoğlu. Diagnostic Value of Diffusion-weighted Magnetic Resonance (MR) Imaging, MR Perfusion, and MR Spectroscopy in Addition to Conventional MR Imaging in Intracranial Space-occupying Lesions. *Cureus*, 11(12), 2019. doi: 10.7759/CUREUS.6409.

[86] N.A. Mohile. Medical Complications of Brain Tumors. *CONTINUUM Lifelong Learning in Neurology*, 23(6):1635–1652, 2017. doi: 10.1212/CON.0000000000000540.

[87] Q. Chen, S. Hu, P. Long, F. Lu, Y. Shi, and Y. Li. A Transfer Learning Approach for Malignant Prostate Lesion Detection on Multiparametric MRI. *Technology in Cancer Research & Treatment*, 18, 2019. doi: 10.1177/1533033819858363.

[88] M.B. Amin, S.B. Edge, F. L. Greene, D.R. Byrd, R.K. Brookland, M.K. Washington, J.E. Gershenwald, C.C. Compton, K.R. Hess, D.C. Sullivan, J.M. Jessup, J.D. Brierley, L.E. Gaspar, R.L. Schilsky, C. M. Balch, D.P. Winchester, E.A. Asare, M. Madera, D.M. Gress, and L.R. Meyer, editors. *AJCC Cancer Staging Manual*. Springer International Publishing, Cham, 2017. doi: 10.1007/978-3-319-40618-3.

[89] D.N. Louis, A. Perry, P. Wesseling, D.J. Brat, I.A. Cree, D. Figarella-Branger, C. Hawkins, H.K. Ng, S.M. Pfister, G. Reifenberger, R. Soffietti, A. Von Deimling, and D.W. Ellison. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology*, 23(8):1231–1251, 2021. doi: 10.1093/NEUONC/NOAB106.

[90] J.S. Paul, A.J. Plassard, B.A. Landman, and D. Fabbri. Deep learning for brain tumor classification. *Proceedings SPIE, Medical Imaging 2017: Biomedical Applications in Molecular, Structural and Functional Imaging*, 10137(13):253–268, 2017. doi: 10.1117/12.2254195.

[91] I.P. Loohuis. Exploring the Prognostic Value of Deep Learning Image-to-Image Registration for Immunotherapy Patient Monitoring. Technical report, University of Twente, Enschede, 2022. URL `https://essay.utwente.nl/90489/1/Loohuis_MA_TNW.pdf`.

[92] P.Y. Wen, D.R. Macdonald, D.A. Reardon, T.F. Cloughesy, A. G. Sorensen, E. Galanis, J. DeGroot, W. Wick, M.R. Gilbert, A.B. Lassman, C. Tsien, T. Mikkelsen, E.T. Wong, M.C. Chamberlain, R. Stupp, K.R. Lamborn, M.A. Vogelbaum, M.J. Van Den Bent, and S.M. Chang. Updated response assessment criteria for high-grade gliomas: Response assessment in neuro-oncology working group. *Journal of Clinical Oncology*, 28(11): 1963–1972, 2010. doi: 10.1200/JCO.2009.26.3541.

[93] S. Bette, M. Barz, B. Wiestler, T. Huber, J. Gerhardt, N. Buchmann, S. E. Combs, F. Schmidt-Graf, C. Delbridge, C. Zimmer, J. S. Kirschke, B. Meyer, Y.M. Ryang, F. Ringel, and J. Gempt. Prognostic Value of Tumor Volume in Glioblastoma Patients: Size Also Matters for Patients with Incomplete Resection. *Annals of surgical oncology*, 25(2):558–564, 2018. doi: 10.1245/S10434-017-6253-0.

[94] F. Isensee, P.F. Jäger, P.M. Full, P. Vollmuth, and K.H. Maier-Hein. nnU-Net for Brain Tumor Segmentation. URL `https://arxiv.org/pdf/2011.00848.pdf`.

[95] F. Isensee, P.F. Jaeger, S.A.A. Kohl, J. Petersen, and K.H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2020. doi: 10.1038/s41592-020-01008-z.

[96] S.M. Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3): 143–155, 2002. doi: 10.1002/hbm.10062.

[97] R. Noiphithak and K. Veerasarn. Clinical predictors for survival and treatment outcome of high-grade glioma in Prasat Neurological Institute. *Asian Journal of Neurosurgery*, 12 (1):28, 2017. doi: 10.4103/1793-5482.148791.

[98] N.A.O. Bush and S. Chang. Treatment strategies for low-grade glioma in adults. *Journal of Oncology Practice*, 12(12):1235–1241, 2016. doi: 10.1200/JOP.2016.018622.