Signs of Progression

MR image analysis for the management of low-grade glioma

Karin A. van Garderen

Signs of Progression

MR image analysis for the management of adult low-grade glioma

Karin A. van Garderen

Acknowledgements:
This research was funded by the Dutch Cancer Society (KWF project number 11026, GLASS-NL).
This work was performed in the framework of the Medical Delta program Cancer Diagnostics 3.0: Big Data Science of in & ex vivo Imaging. Medical Delta and the department of Radiology and Nuclear Medicine are gratefully acknowledged for financial support for the printing costs of this thesis.

ISBN: 978-94-6483-765-0
Cover: Karin van Garderen
Layout: Karin van Garderen, based on template by Sebastian van der Voort
Printing: Ridderprint, www.ridderprint.nl

Signs of Progression

MR image analysis for the management of adult low-grade glioma

Tekenen van progressie

Beeldanalyse van MRI voor het beleid rond laaggradig glioom bij volwassenen

THESIS

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof. dr. A.L. Bredenoord

and in accordance with the decision of the Doctorate Board.

The public defence shall be held on
Wednesday 13 March 2024 at 13.00 hrs

by

Karin Alida van Garderen
born in Freyung, Germany

Erasmus University Rotterdam

Doctoral Committee

Promotors	Prof. dr. M. Smits
	Dr. ir. S. Klein
Other members	Prof. dr. M.W. Vernooij
	Prof. dr. P.C. Witt Hamer
	Dr. E. Konukoglu

progression /prəˈgrɜʃn/ *noun*
1. the process of developing gradually towards a more advanced state
2. a number of things in a series

Contents

1 Introduction **1**
1.1 Glioma diagnosis and treatment 4
1.2 Interpretation of MRI for glioma 5
1.3 Volume measurement 8
1.4 Computer-aided medical image analysis 8
1.5 Modelling glioma growth 11
1.6 Outline of this thesis 12

I Glioma segmentation and volume measurement **13**

2 Multi-modal segmentation with missing MR sequences **14**
2.1 Introduction 17
2.2 Methodology 18
2.3 Results 21
2.4 Discussion and conclusion 22

3 Pre-radiotherapy tumor burden in glioblastoma **26**
3.1 Introduction 29
3.2 Methods 29
3.3 Results 31
3.4 Discussion 37
3.5 Supplementary materials 40

4 Clinical Implementation of Glioma Segmentation **46**
4.1 Introduction 49
4.2 Materials and equipment 50
4.3 Methods 54
4.4 Results 57
4.5 Discussion 61

II Emerging biomarkers and methods 64

5 Longitudinal characteristics of T2-FLAIR mismatch 66
5.1 Introduction . . . 69
5.2 Methods . . . 70
5.3 Histopathology . . . 73
5.4 Results . . . 74
5.5 Discussion . . . 86
5.6 Supplementary material . . . 89

6 Deep learning-based longitudinal registration 98
6.1 Introduction . . . 100
6.2 Methods . . . 100
6.3 Results . . . 101
6.4 Discussion . . . 103
6.5 Conclusion . . . 103

7 Evaluating glioma growth models for low-grade glioma 106
7.1 Introduction . . . 109
7.2 Methods . . . 111
7.3 Results . . . 118
7.4 Discussion . . . 126
7.5 Conclusion . . . 130

8 Discussion 132
8.1 Glioma segmentation with missing sequences . . . 134
8.2 Volume measurement in clinical practice . . . 137
8.3 Emerging biomarkers . . . 140
8.4 Longitudinal registration . . . 143
8.5 Growth prediction . . . 145
8.6 Deep learning . . . 150
8.7 Conclusion . . . 152

Bibliography 155

Summary 173

Samenvatting 175

Acknowledgements 177

Publications 179

PhD portfolio 181

About the author 185

1
Introduction

She could read anything now, he said, and once you can read anything you can learn everything. It was up to her.
– Delia Owens, Where the Crawdads Sing

Cancer is among the leading causes of death worldwide. We are making progress in the fight against cancer through early detection and increasingly effective treatment. We are increasingly aware of how to reduce our chances of getting cancer by improving our lifestyle and by avoiding exposure to known cancer-causing substances [1]. However, the prognosis for patients diagnosed with diffuse glioma, a type of brain cancer, is still very poor. This type of cancer has no cure and can strike anyone, including young people who are otherwise in good health. What happens when you are suddenly faced with such a diagnosis? Would you opt for a treatment that could give you more time to live, or do you want to spend your remaining months or years outside the hospital? In any case, I imagine you and your loved ones would be eager to know exactly what the prospects are. The diagnostic capabilities available for glioma have improved drastically over the last few years with advances in technology and medical research, but the disease trajectory still comes with a large degree of uncertainty. Patients with low-grade glioma may spend many years in a relatively stable condition never knowing when the inevitable progression of the disease will occur. The treating physician may opt to start treatment if there are signs of progression, with the aim of slowing down the development of the disease, so an early and accurate diagnosis of progression is desired.

Even if we do not find a cure, improving our understanding of the disease can increase the quality of life for patients by decreasing uncertainty and informing treatment decisions. The aim of the research in this thesis is to improve the accuracy of diagnosis throughout the course of the disease by means of magnetic resonance (MR) image analysis. Specifically, I explore the role of quantitative measurements, emerging imaging markers and predictive modelling in the management of glioma. These methods can aid the radiologist to predict the timing, location and severity of tumor progression, to ultimately improve the quality of life for glioma patients.

1.1 Glioma diagnosis and treatment

Every year approximately one thousand people in the Netherlands are diagnosed with diffuse glioma, a type of infiltrative brain tumor that originates from the glial cells. The first signs of glioma can be headaches, seizures or neurological deficits. In some cases the diagnosis comes after an incidental finding, when the patient has not noticed any symptoms yet [2]. The initial diagnosis is made through magnetic resonance imaging (MRI), but tissue is needed for the final diagnosis of glioma and its type. These types are defined in the 2021 WHO classification [3] according to their molecular characteristics. In addition to the molecular type, diffuse glioma are assigned a grade. A higher grade corresponds to a worse prognosis and it is also defined by both molecular and

histological characteristics. There are three main categories for adult-type diffuse glioma, listed here in order of increasing severity:

- **Oligodendroglioma, IDH-mutant and 1p/19q-codeleted**: grade 2 or 3.
- **Astrocytoma, IDH-mutant**: grade 2, 3 or 4.
- **Glioblastoma, IDH-wildtype**: grade 4.

Unlike the main glioma types, the tumor grade can change as the disease develops. Although astrocytoma may initially present as a grade 2 tumor and grow relatively slowly, they are known to develop more malignant behavior and decreased therapy response over time. This change is called a malignant transformation [4]. The GLASS (Glioma Longitudinal AnalySiS) consortium was founded to study the development of glioma over time, and the GLASS-NL cohort is one of the initiatives to study specifically IDH-mutant astrocytoma through a multi-center study in the Netherlands.

The recommended treatment for diffuse glioma is maximal safe resection through craniotomy, which may be followed by chemo- and radiotherapy depending on the severity of the disease [5]. Treatment can prolong life and reduce symptoms, but can also cause burden to the patient. The choice of treatment depends on the condition of the patient and the development of the disease, which is monitored through regular MRI examinations. Patients may receive multiple cycles of treatment interleaved with periods of watchful waiting, where a new treatment might be started if there are signs of progression. Progression is defined as a worsening of the disease and can occur in the form of increasing symptoms or changes visible on MRI. The current guidelines for the diagnosis of progression are based on changes in T2-weighted hyperintensities and contrast enhancement, as observed on MRI, or changes in medication or symptoms [6].

1.2 Interpretation of MRI for glioma

MRI provides a non-invasive way to image the brain with high resolution and excellent image contrast, which makes it an ideal modality for the imaging of glioma. The MR scanner is capable of generating images with different contrasts depending on the scan protocol, which is a sequence of instructions for the generation of radiofrequency pulses and measurements. For the imaging of glioma, both the T1-weighted and T2-weighted sequences provide important information. One indicator of a highly malignant tumor is that it compromises the blood-brain-barrier locally. The T1-weighted sequence is used in combination with a contrast agent, based on Gadolinium, to elicit a high signal intensity in the blood. This contrast agent does not normally pass through the

blood-brain barrier, so by comparing the T1-weighted scan before and after infusing contrast agent the radiologist can detect areas where the blood-brain barrier has been compromised. High-grade glioma often present with a ring of contrast enhancement and a center of necrotic tissue, usually accompanied by vasogenic edema that is hyperintense on T2-weighted imaging. This region of edema may also contain infiltrating tumor cells, but the extent of infiltration in the edema is generally unknown.

Low-grade glioma will often present only as a hyperintense region on T2-weighted scans and without contrast enhancement. The interpretation of T2-weighted hyperintensities depends on multiple factors, such as recent treatment, and the diffuse nature of glioma makes it impossible to determine an exact boundary of the tumor. T2-weighted Fluid-Attenuated Inversion Recovery (T2w-FLAIR) can be used in addition to the T2-weighted scan to assess the non-enhancing region of the tumor, as it provides a better contrast between edematous regions and cerebrospinal fluid (CSF). Figure 1.1 contains an example of both a high and low-grade glioma as it presents on T1w and T2w MRI.

1.2.1 Treatment effect

The assessment of MRI becomes more challenging when a patient has received treatment, due to the radiological changes that are attributed to treatment effect. The definition of treatment effect, as used in this thesis, is any abnormality in the MRI of the brain that is a direct result of treatment and is not induced by the tumor. Immediate post-surgery imaging can show contrast leakage, hemorrhage, ischaemia and swelling. This makes immediate estimation of the residual tumor difficult, but the effects typically subside within a few months. Radiation therapy can lead to acute edema, an increase of the enhancing lesion, or appearance of new areas of enhancement and white matter changes, especially if combined with chemotherapy. The increase of the enhancing lesion can mimic tumor progression [7, 8] and is therefore referred to as pseudoprogression. Left untreated, pseudoprogression will stabilize or subside, as opposed to progressive glioma, but it can appear months to even years after the end of radiotherapy [9]. White matter changes (leukoencephalopathy) can appear as a result of both radio- and chemotherapy in the form of T2-weighed hyperintensities [10]. This presents a challenge when estimating the extent of infiltrating non-enhancing tumor, especially when the hyperintensities overlap with the infiltrative tumor. Conventional MRI has limited utility to distinguish true progression from pseudoprogression intially [11]. This presents a dilemma for the treatment management, as waiting to follow further development of the lesion may have consequences for the efficacy of follow-up treatment.

Treatment effect is not only a challenge for clinical decision-making, but also for research. If, in clinical trials, treatment response and progression

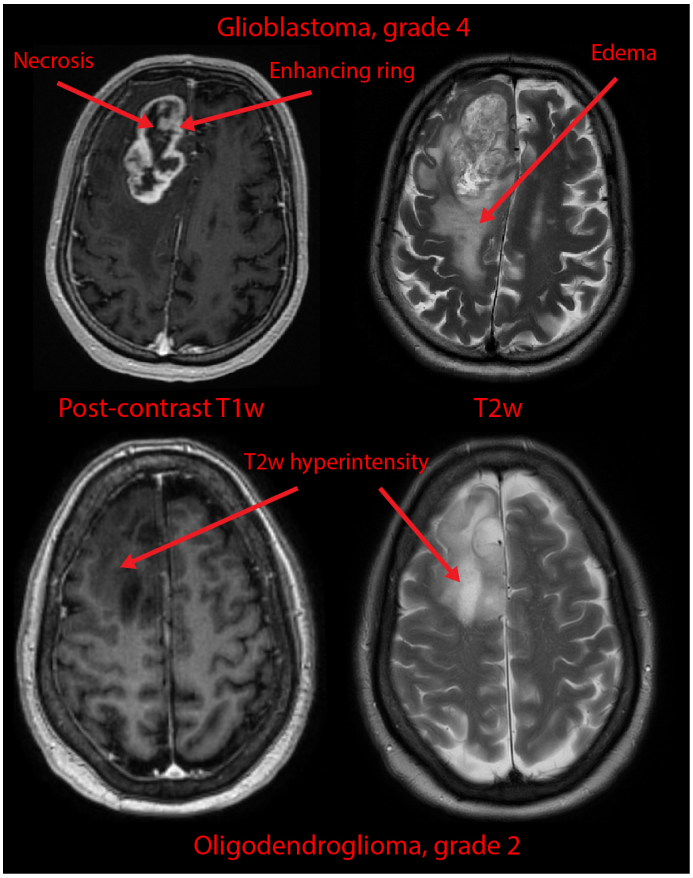


Figure 1.1: Examples of MRI depicting glioma, with notable structures indicated.

are used as outcome measures, and those are at least partially defined by radiological findings, then the distinction between pseudoprogression and true progression is essential. Furthermore, tumor progression and treatment effect are not mutually exclusive and an imaging abnormality can be a mixture of treatment effect and tumor growth. If we want to quantify the extent of tumor growth, then such heterogeneous lesions present a major challenge. This is part of the reason that the inter-rater agreement in the delineation of recurrent non-enhancing glioma is poor [12].

There is also a phenomenon called pseudoresponse, where treatment induces an effect that mimics treatment response even though it is not directly related to a clinically relevant reduction in tumor activity. This is mainly found in anti-angiogenic treatment that affects the tumor vasculature, thereby reducing the uptake of contrast agent in the lesion and alleviating edema, causing a drastic reduction in the visible lesion size. The overall survival may not improve despite this initially promising response, suggesting that the effect on tumor activity is limited [13, 14].

1.3 Volume measurement

The change in enhancing and non-enhancing tumor volume is an important marker for treatment response and disease progression. In clinical practice, a full volume measurement is rarely performed as it costs too much time. Instead, a measurement of two perpendicular diameters is recommended [15] to estimate the size of the lesion. In non-enhancing glioma this method of measurement is often not applicable due to the shape of the lesion, even though the slow but gradual growth of non-enhancing glioma is an important marker for the risk of malignant progression [6]. If a reliable automated method of volume measurement would be available, this would potentially increase the accuracy of the measurement and provide a quantification of growth for lesions that are currently considered unmeasurable.

The automated delineation of structures, also called semantic segmentation, is a long-standing area of research in computer vision and medical image analysis. In recent years, machine learning and particularly convolutional neural networks (CNN's) have gained traction as the state-of-the-art method to solve the semantic segmentation problem in medical imaging [16]. During the years in which I have been working on this thesis, I have seen that medical image segmentation has matured from a range of methodologies that require problem-specific tuning to a problem where plug-and-play methods [17] are readily available. For the specific case of glioma segmentation, the recurring BraTS challenge has been a driver for innovation in this field [18]. Pre-trained and extensively validated methods have become available [19, 20], removing even the need to collect a training dataset before applying these methods.

1.4 Computer-aided medical image analysis

The automated segmentation of glioma is only one example of the wide range of methods available for medical image analysis. First I will describe the most relevant types of problems - image registration, semantic segmentation and computer-aided diagnosis - before discussing the concept of machine learning which is the methodology often used to solve these problems. Note that many

more problems and solutions exist in the field, but a complete discussion of computer-aided medical image analysis is beyond the scope of this thesis.

1.4.1 Image registration

Medical image registration is the process of transforming multiple medical images into a common space so that corresponding locations are aligned. This could involve multiple images of the same subject taken at a different time or with a different image contrast (intra-subject), or it could involve images of different subjects (inter-subject). Registration can be used to align the different MRI sequences in a single scan session so that the different contrasts can be evaluated together. This is an important step before automated segmentation of the glioma can take place, because the segmentation assumes an image where the different contrasts can be viewed like color channels in a photograph. The subject may move slightly during the scan session and each contrast may be acquired with a different resolution, field of view and direction. The registration serves to correct for slight displacements and bring the images to a common space.

Intra-subject registration of multiple scan sessions can be required in order to assess the changes over time. Additionally, a registration to an atlas can be performed. An atlas is a reference image, derived from one or multiple images, that may contain additional information such as the location of relevant regions of interest [21, 22]. By registering an image to the atlas we can enrich the data using information contained in the atlas [23, 24]. For example, an atlas of healthy subjects can be used to identify the location of the glioma and estimate the underlying tissue structure.

Performing a registration is solving an optimization problem, which is something that computers can do very effectively. Typically, the registration optimizes an error term that is defined by the intensities in the image, assuming that the intensities in corresponding locations are highly correlated. The registration algorithm finds a deformation that minimizes this error term by changing the parameters of a deformation model. The deformation model defines what types of deformation are possible, such as translation, rotation, shearing or even more complex warping of the image. Depending on the application, different error terms and deformation models may be used [25].

1.4.2 Semantic segmentation

Segmentation is the process of delineating structures or regions of interest. An example of this is the delineation of tumor-induced abnormalities on an MR image. A segmentation can be performed manually, automatically or semi-automatically. Semi-automatic segmentation involves tools that use image information to generate the most likely segmentation based on user input,

e.g. by automatically filling regions of similar intensities. This is especially relevant in the segmentation of 3D images, which would otherwise require the user to perform a multitude of 2D segmentations. Ideally, however, user interaction would not be required at all. Automatic segmentation methods require no user interaction, and therefore enable the processing of images at large scale. The methods for automatic segmentation are numerous, but as opposed to semi-automatic segmentation they typically require a method to be tailored specifically to each application, and data has to be used to develop this method. One example of an approach to automatic segmentation is atlas registration, where an initial segmentation can be transformed from the atlas to a specific patient [23, 24]. This can be used to segment the healthy brain, or its regions, but is not applicable to pathology that presents differently in each patient. In recent years, the dominant approach to semantic segmentation is machine learning, and specifically convolutional neural networks [16], which are discussed in section 1.4.4.

1.4.3 Computer-aided diagnosis

So far the methodologies discussed in this section were used to enhance images or provide measurements, which may aid the diagnosis, but it is also possible to design methods that provide a diagnosis directly. In the field of machine learning this problem is called classification. In the context of glioma, the prediction of the glioma type is an important diagnostic problem where machine learning methods can contribute, potentially being able to predict a molecular diagnosis from MRI without the need for tumor tissue [26]. This thesis, however, is concerned explicitly with the management of glioma after initial diagnosis. There are plenty of classification problems to be solved here as well, such as the diagnosis of tumor progression versus treatment response [27], or malignant versus non-malignant progression, but they are not tackled in this thesis. Instead, the methods presented here aim to increase understanding and offer tools to aid the assessment by a clinical expert, rather than provide a diagnosis directly.

1.4.4 Machine learning

The terms machine learning, artificial intelligence and deep learning are often used interchangeably, but they refer to slightly different groups of techniques. Artificial intelligence is a term that can be used for any method that in some way mimics intelligence. The definition includes machine learning and deep learning, but also algorithms that are based on logic rather than data, such as reasoning systems and conventional image processing techniques. More specifically, machine learning describes algorithms that use data, or some other form of input, to distill some general patterns and relationships. These

methods require a dataset of relevant examples, called a training set, from which some patterns may be distilled that generalize to new examples. The term 'deep learning' is used for specific methods of machine learning where a neural network of multiple layers is optimized using gradient descent [28]. A detailed description of the methodology is beyond the scope of this thesis, though the CNN's I use for image segmentation in this thesis are an example of deep learning.

In order to automatically segment glioma on an MRI, we can train a CNN using examples [19, 29]. These examples have to be created e.g. by manually segmenting a number of cases. The network is then optimized by adjusting millions of parameters to recreate these examples to the best of its ability. Machine learning and specifically CNNs can also be applied to the classification of images, which works mostly in the same way as semantic segmentation. In this case, the algorithm is trained using examples of images and their diagnosis in order to predict the correct diagnosis for each image.

However, the goal is not to recreate the examples but to be able to perform new segmentations or diagnoses for unseen cases. This is called generalization. In general, machine learning methods do not come with any guarantees on their degree of generalization. Unlike conventional statistics, it is difficult to find out what logic and assumptions the CNN uses to perform its task, which makes it difficult to reason about the limits of its generalization. This is why a test set is needed to evaluate the performance with cases that have not been used to optimize the network. The performance on the test set is a reliable estimate of the performance of the method in practice, but it is important to consider whether the test set is representative for the expected population and conditions [30]. The performance may degrade considerably if we apply the method in different conditions or on a different population.

1.5 Modelling glioma growth

What if we could not only delineate the tumor now, and in the past, but also look into the future? In clinical practice, the radiologist and other treating physicians already perform an estimation of what the future will bring. If they see the tumor increasing in size, they assume that it will continue to grow and therefore may decide to start treatment. If abnormalities start to appear in the corpus callosum, they may assume that the actual extent of tumor infiltration has extended into the contralateral side of the brain. This is based on the knowledge that the true extent of tumor infiltration is largely unseen, and glioma cells tend to infiltrate faster through white matter tracts. It makes sense to turn this knowledge into a quantitative prediction and try to simulate tumor growth based on our knowledge and observations. Not only would this be good method to test our assumptions, it may prove to be useful

for treatment decisions and localized treatment. This is a long-standing area of research where machine learning is a relative newcomer [31, 32], and explicit biophysical models make up the state-of-the-art solutions [33, 34].

In a biophysical model of glioma, the diffuse nature of the tumor is modelled through its local cell density. The areas where tumor infiltration is visible, through contrast-enhancement on T1-weighted or hyperintensity on T2-weighted imaging, are assumed to have at least a certain density of tumor cells. The healthy-appearing brain may still contain infiltrating tumor cells, but in a lower density. To formalize this assumption, the visible outlines of the tumor on MRI are assumed to be an isodensity contour of tumor cells $c(x)$. To simulate tumor growth, a model of diffusion and proliferation is assumed according to the Fisher-Kolmogorov equation:

$$\frac{dc(x)}{dt} = \nabla D \nabla c(x) + \rho c(x)(1 - c(x)),$$

where the parameters D and ρ define the rate of diffusion and proliferation respectively. When adding the boundary conditions imposed by the brain anatomy, and an initial condition that is estimated for each specific patient, this system of partial differential equations provides an estimate of the tumor cell density over time. Research has shown that this method can approximate the current and future tumor shape for individual patients [35, 36].

1.6 Outline of this thesis

In this thesis, I aim to explore methods to improve the MR image-based analysis of glioma after initial treatment. The term 'method' is interpreted broadly, because the methodologies I use range from deep learning and biophysical modelling to visual assessment. Although the application of deep learning for the purpose of volume measurement is explored in more detail and broader scope, the main focus of this thesis is on the assessment of low-grade glioma after initial treatment. Part 1 concerns the use of deep learning for the purpose of volume measurement, with both methodological and translational contributions. Part 2 concerns emerging biomarkers and methods, and their relevance in the assessment of low-grade glioma.

Part 1: Glioma segmentation and volume measurement

- Chapter 2 explores the challenge of glioma segmentation with missing data, using a dedicated neural network design that fuses information from different MRI sequences in a later stage. I investigate different designs of the network with different degrees of missing data.
- Chapter 3 is an application of glioma segmentation and volume measurement for post-operative glioblastoma. This study investigates the

relevance of remaining tumor volume after initial resection and before radiotherapy, both in terms of enhancing and non-enhancing regions.

- Chapter 4 describes the clinical application of glioma volume measurement for the management of non-enhancing low-grade glioma. A deep learning method is implemented and evaluated in clinical practice to aid the diagnosis of progression. The initial evaluation also provides insight in remaining challenges for clinical adoption of automated volume measurement.

Part 2: Emerging biomarkers and methods

- Chapter 5 explores a specific imaging marker, called the T2-FLAIR mismatch sign, in a longitudinal setting using the GLASS-NL cohort. The prognostic value of T2-FLAIR mismatch sign and its longitudinal behavior is largely uncertain, and this study provides some indication that there is prognostic value especially in the recurrent setting.

- Chapter 6 concerns a method for the intra-patient longitudinal registration of MR images, in order to quantify the deformation caused by glioma growth. In this preliminary analysis, a novel method based on deep learning is described and compared to existing methodologies.

- Chapter 7 concerns the evaluation of glioma growth predictions for low-grade glioma after resection. The focus of this study is on the evaluation of predictions, where I use a ranking of voxels rather than a binary prediction at a specific volume threshold.

I

Glioma segmentation and volume measurement

2

Multi-modal segmentation with missing MR sequences using pre-trained fusion networks

Science is built up with facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house.

– *Jules Henri Poincaré*

Based on: **Karin A. van Garderen**†, M. Smits, and S. Klein, "Multi-modal segmentation with missing MR sequences using pre-trained fusion networks," in *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, Presented at MICCAI MIL3ID Workshop, vol. 11795 LNCS, Springer, 2019, pp. 165–172

† indicates presenting author

2

Abstract

Missing data is a common problem in machine learning and in retrospective imaging research it is often encountered in the form of missing imaging modalities. We propose to take into account missing modalities in the design and training of neural networks, to ensure that they are capable of providing the best possible prediction even when multiple images are not available. The proposed network combines three modifications to the standard 3D UNet architecture: a training scheme with dropout of modalities, a multi-pathway architecture with fusion layer in the final stage, and the separate pre-training of these pathways. These modifications are evaluated incrementally in terms of performance on full and missing data, using the BraTS multi-modal segmentation challenge. The final model shows significant improvement with respect to the state of the art on missing data and requires less memory during training.

2.1 Introduction

Tumor segmentation is a key task in brain imaging research, as it is a prerequisite for obtaining quantitative features of the tumor. Since manual segmentation by radiologists is time-consuming and prone to inter-observer variation, there is a clear need for effective automatic segmentation methods. Research into these methods for glioma has been accelerated by the recurring BraTS multi-modal segmentation challenge on low-grade glioma (LGG) and glioblastoma (GBM) [37]. The best performing methods in recent editions were all based on 3D convolutional neural networks (CNNs) with the encoder-decoder shape of the UNet.

While the BraTS challenge focuses on improving performance, there are practical problems to overcome before automatic segmentation can be applied in practice. One of these challenges is dealing with missing data. The BraTS benchmark contains four MR modalities: a T1-weighted image (T1W), a T1-weighted image with contrast agent (T1WC), a T2-weighted image (T2W) and a T2-weighted FLAIR image (FLAIR), which are co-registered so that corresponding voxels in the image are aligned and a CNN can learn to segment a tumor from the specific combination of modalities. Although these images are complementary, a radiologist is still able to perform a partial segmentation if one of these modalities is missing, while for a CNN this is not guaranteed. Especially in retrospective and multi-center studies it is not unlikely that images are either missing or have quality issues.

There are two ways in general to deal with the problem of missing data. The most common way is to impute the missing values by an estimate, which can be as simple as the mean value. More advanced techniques for missing image imputation is to generate a new image from remaining modalities, which can be achieved through neural networks [38] [39].

However, it is also possible to train a CNN to be inherently robust to missing data. The HeMis model [40] is an example of this, where the modalities are each passed through a separate pathway before being merged in a so-called abstraction layer which extracts the mean and variance of the resulting features. This network architecture enforces a shared feature representation of the modalities, though it may be of additional value to include a similarity term in the loss function to enforce a true shared representation [41].

2.1.1 Contribution

Building on the existing work on shared representations, we provide a careful experimental evaluation of different aspects that make the network robust to missing images. We evaluate four modifications to a state-of-the-art UNet architecture and evaluate their effect incrementally. A first adaptation is to train with missing data in a curriculum learning approach. Secondly, a multi-

path architecture is evaluated where the information of different modalities is fused in a later stage. Thirdly, within this architecture, a shared representation layer is compared to a concatenation of feature maps. Finally, we propose a training procedure where each pathway is trained separately before combining them and training the final classification layer. This approach enforces each path to form an informative feature represenation. The separate training also reduces the demand on GPU memory, which is the main bottleneck in state-of-the-art segmentation networks. The modified architectures are compared to the baseline architecture, in a situation where it is trained with the entire dataset but also when it is specifically trained for each combination of modalities.

2.2 Methodology

2.2.1 Network architecture

The 3D UNet architecture [42] is a well-established segmentation network and still was one of the best performing architectures at the most recent 2018 BraTS challenge [43]. Therefore the UNet forms the baseline for our research. One UNet is trained on all modalities and evaluated with missing data, but also a dedicated UNet is trained and evaluated for each specific combination of modalities. The number of trainable parameters in the model depends on the number of feature maps in each convolution, which we chose to parameterize by a single variable c. The first convolution has c kernels, and as the size of the feature maps decreases the number of kernels is increased. Fig. 2.1 shows the UNet architecture with the number of feature maps per convolution layer

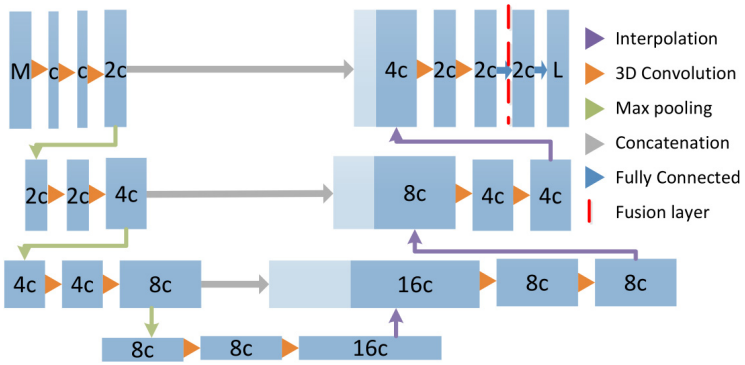


Figure 2.1: Illustration of the UNet architecture. The number of feature maps, as a function of the parameter c, is indicated for each step. The fusion and shared representation networks contain one UNet per modality, which are fused at the indicated location. M indicates the number of input modalities and L the number of output labels. In this study $M = 4$ and $L = 4$.

expressed as a multiple of c.

In the reference UNet architecture each 3D convolution block contains a batch normalization, a 3D unpadded convolution layer with kernels of size 3^3, and Leaky ReLu activation. The last fully connected layers are implemented as a 3D convolution with kernels of size 1^3. The downsampling step is a max-pooling layer of stride 2 and size 2^3 and the upsampling is a tri-linear interpolation. For this UNet architecture each target voxel has a receptive field of 88^3 voxels.

Modality dropout.

To make a network robust to missing data it needs to train with missing data. To this end, a specific modality dropout scheme was implemented which removes entire input channels (MR sequences) with a probability p. The features from missing sequences are removed by setting the input to zero and scaling the other inputs by m_o/M, where m_o is the number of original input images and M is the number of remaining inputs. A curriculum learning approach is used to aid convergence: starting from $p = 0.125$ the probability of dropout is doubled every 50 epochs until it reaches $p = 0.5$. This method is applied directly to the input layer in the Dropout network, but also to the fusion layers in the Multipath and SharedRep networks.

Multipath network.

In this approach the network has one pathway for each of the $M = 4$ modalities and the feature maps of the final convolutional layer are concatenated to an output of $8c$ channels in a fusion layer, which is where the modality dropout is applied. The final prediction is performed again by a 1^3 convolution layer with $4c$ channels.

For a fair comparison it is important to consider the number of trainable parameters, which scales quadratically with the number of channels per layer. To create a multi-path network of the same size as a single reference network, the UNets that form the pathways have half the number of channels per layer. As the UNet was implemented with $c = 32$, the separate pathways are a quarter of the size with $c = 16$. Note that whereas parameter size scales quadratically, the memory usage scales approximately linear with the number of feature maps. The multi-pathway networks (with $M = 4$) therefore require approximately twice the amount of GPU memory during training compared to the single UNet.

Shared Representation.

The Shared Representation (SharedRep) network is a multi-path network with a specific fusion layer, based on the HeMIS model [40]. Instead of concatenating,

the fusion layer takes the mean and variance of each feature map and therefore encourages a common feature represenation between the modalities. To enable fair comparison to the fusion network, the last layer of each pathway has double the amount of feature maps ($4c$), leading to $8c$ features in the fusion layer. The network is trained with modality dropout of the pathways and the variance is set to zero if only a single pathway is available.

Pre-trained paths

Pre-training the paths means that a UNet is trained for each individual MR modality and the separate prediction layers are replaced by one fusion layer. These are trained with modality dropout ($p = 0.5$), while freezing the parameters of the single pathways. When fusing the pathways with a shared representation layer, the final convolutional layers of the networks are also replaced and trained in order to learn a new shared feature representation. Using the pre-training scheme greatly reduces the demand on GPU memory, as the pathways require a quarter of the memory of the whole network and half that of the full UNet with $c = 32$. The combined training scheme took approximately 50% longer than without pre-training, though with parallel training of the paths on separate devices it was even faster than the baseline.

2.2.2 Data and preprocessing

The networks were trained and evaluated on the training set of the BraTS challenge 2018 [44], which is a benchmark dataset of pre-operative scans of 278 patients with low-grade glioma (LGG, 75) or glioblastoma (GBM, 203). The images in this benchmark are skull-stripped, co-registered and resampled to a size of 240 by 240 by 155 voxels. The target areas for evaluation are the whole tumor, tumor core and enhancing core. The non-background voxels of each separate image were normalized to zero mean and unit standard deviation. Random patches of 108^3 voxels were extracted, which correspond to 20^3 target voxels. With a probability of 50% a patch was selected from a tumor area, meaning that the center voxel was part of the tumor, and with 50% probability the center voxel was located outside of the tumor but inside the brain.

2.2.3 Training and evaluation

The networks were optimized with the Adam optimizer [45] and the cross-entropy loss function. An epoch is defined as an iteration over 100 batches with 4 random patches, and the models were trained for 150 epochs. For pre-trained pathways, the separate pathways and the final combination layer were trained for 100 epochs each. The dataset was divided into five cross-validation folds, so that 20% of the subjects were always selected for testing and never used during training. The folds are random, but the same for each experiment. Evaluation

took place on the whole image, although it was classified by the network in patches to limit memory usage. To assess whether the models are indeed more robust to missing data, we evaluated the same models in a situation where any combination of sequences is removed.

2.2.4 Visualizing shared representations

To validate the concept of a shared representation layer in the context of missing data, we would like to know whether the feature representation of such a layer is indeed robust to missing data. We evaluated this in a qualitative way by looking at the t-SNE [46] maps of the activations of the final fully connected layer. Feature maps from the final fully connected layer were extracted for 40,000 random voxels originating from 16 random patches. A t-SNE map was computed to map the 64-dimensional feature vectors to a 2D representation. These maps can be interpreted as a representation of the distances between voxels in the specific feature representation of each model. The same set of voxels was used for both maps.

2.3 Results

Six networks were trained and evaluated in five-fold cross-validation and, as an additional reference, a dedicated UNet was trained for each combination of sequences. The results are summarized in Table 2.1. On the full dataset, the simple UNet without dropout performs best, and every modification to the network comes with a decreased performance in this case. For missing data scenarios, the regular UNet suffers while the other networks are able to maintain a better performance. None of the networks is able to outperform a dedicated UNet trained for each specific combination of sequences.

There is no architecture that consistently outperforms the others, though the pre-trained multipath networks seem to perform best overall and especially on cases with few available modalities. However, when considering performance on the full dataset, the UNet baseline still performs best and the SharedRep model without pretraining performs better than pretrained paths on the tumor core. Training only with modality dropout greatly decreases performance on the full dataset while only providing minor improvement on missing data.

2.3.1 t-SNE visualizations

The resulting t-SNE representations are shown in Fig. 2.2 for the pretrained Multipath and SharedRep model. The predicted and true labels are highlighted in red, showing that the mapped representation is meaningful to the network prediction and ground truth. Also, the feature maps generated with missing data are highlighted to see whether they lead to distinct feature representations.

Whereas the multipath fusion model maps the different missing data scenarios to specific parts of the feature space, the shared representation model seems to have less distinction between complete and incomplete data. This visualization supports the notion that the shared representation layer does indeed lead to a feature representation that is consistent, even when images are removed.

2.4 Discussion and conclusion

We have carefully evaluated different approaches for training a CNN to be robust to missing imaging modalities, in the context of the BraTs multi-modal segmentation challenge with four MR sequences. Applying modality dropout on the input channels is a simple way to achieve some robustness, but it has a significant impact on performance with full data. More advanced multimodal architectures, with a separate pathway for each modality, give a better balance between performance and robustness.

The pathways can be fused either through a simple concatenation or using their statistical moments (mean and variance), thereby enforcing a shared feature representation. Although qualitative visual results show that the shared representation layer forces the feature maps of different combinations of modalities toward a common space, the performance results give no conclusive evidence that it should be preferred over a simple concatenation. The pretraining of the separate paths with a single modality seems to increase the performance mostly in the more difficult cases with fewer modalities. It is also in these cases that a dedicated UNet trained for the specific combination of modalities performs best in comparison, showing that there is still room for improvement.

However, it must be noted that the performance achieved by multipath models do not match the best performance on the most recent BraTS training set, as measured on the full dataset. Further improvements on the UNet core are expected to increase the performance further, on both full and partial datasets.

The evaluation in this paper has focussed on a systematic comparison of model architectures with the same hyperparameters and size. However, the demand on GPU memory is different between networks. The pre-training of paths in the multipath networks drastically reduces the required memory, so they could be trained with more channels per layer, a larger batch size, a larger patch size or simply a less expensive GPU. It should be preferred for this reason and for its consistent good performance with any combination of modalities.

Acknowledgements

This work was supported by the Dutch Cancer Society (project number 11026, GLASS-NL), the Dutch Organization for Scientific Research (NWO) and NVIDIA Corporation (by donating a GPU).

Table 2.1: Numeric results in terms of mean Dice percentage on the three different regions of interest. Color scales are adapted to each region, defined by the best and worst results on that region.

	UNet	Dropout	Multipath	SharedRep	Multipath + Pretraining SharedRep + Pretraining	Dedicated
All	83	77	82	83	84	83
All but T1W	65	76	81	82	83	81
All but T1WC	78	81	82	82	83	81
All but T2W	74	76	77	79	82	79
All but FLAIR	43	59	70	72	75	73
T2W, FLAIR	65	73	80	81	82	79
T1WC, FLAIR	43	62	74	74	78	77
T1WC, T2W	46	59	69	71	74	74
T1W, FLAIR	63	77	77	76	78	76
T1W, T2W	23	61	70	71	73	72
T1W, T1WC	18	33	42	48	56	59
FLAIR	37	51	69	72	72	73
T2W	30	60	63	69	70	71
T1WC	14	21	32	36	49	49
T1W	4	8	25	29	44	48

Whole tumor

	UNet	Dropout	Multipath	SharedRep	Multipath + Pretraining SharedRep + Pretraining	Dedicated
All	71	57	69	70	66	67
All but T1W	47	59	67	69	65	66
All but T1WC	43	33	44	43	42	42
All but T2W	59	56	64	66	64	64
All but FLAIR	46	50	61	64	64	63
T2W, FLAIR	43	36	44	41	42	43
T1WC, FLAIR	26	42	58	60	61	59
T1WC, T2W	36	43	57	59	61	60
T1W, FLAIR	35	20	40	38	40	37
T1W, T2W	23	39	33	37	37	37
T1W, T1WC	26	42	46	50	53	53
FLAIR	28	28	36	30	34	34
T2W	27	40	34	31	36	29
T1WC	9	13	31	38	43	49
T1W	2	9	25	20	29	23

Tumor core

	UNet	Dropout	Multipath	SharedRep	Multipath + Pretraining SharedRep + Pretraining	Dedicated
All	63	57	61	62	62	63
All but T1W	40	56	61	61	62	60
All but T1WC	6	7	7	8	12	10
All but T2W	55	58	58	60	60	58
All but FLAIR	43	55	56	58	60	59
T2W, FLAIR	2	5	5	7	12	12
T1WC, FLAIR	21	39	55	55	57	54
T1WC, T2W	36	46	54	55	58	57
T1W, FLAIR	6	9	7	10	16	9
T1W, T2W	4	8	8	7	6	8
T1W, T1WC	25	44	44	48	50	50
FLAIR	7	4	9	5	17	9
T2W	6	6	6	5	1	8
T1WC	6	13	33	39	39	48
T1W	3	9	9	6	9	9

Enhancing core

Figure 2.2: t-SNE results for pretrained network with fusion by concatenation (left) and shared representation (right). Specific subsets of the voxels are indicated in red.

3

Association of pre-radiotherapy tumor burden and overall survival in newly diagnosed glioblastoma

An underestimated factor in quality of life is the hope of staying alive.

– *Lieke Marsman*

Based on: A. Alafandi, **Karin A. van Garderen**, S. Klein, S. R. van der Voort, D. Rizopoulos, L. Nabors, R. Stupp, M. Weller, T. Gorlia, J. C. Tonn, and M. Smits, "Association of pre-radiotherapy tumor burden and overall survival in newly diagnosed glioblastoma adjusted for MGMT promoter methylation status," *European Journal of Cancer*, vol. 188, pp. 122–130, Jul. 2023

Abstract

Purpose: *We retrospectively evaluated the association between post-operative pre-radiotherapy tumor burden and overall survival (OS) adjusted for the prognostic value of O6-methylguanine DNA methyltransferase (MGMT) promoter methylation in patients with newly diagnosed glioblastoma treated with radio-/ chemotherapy with temozolomide.*

Materials and Methods: *Patients were included from the CENTRIC (EORTC 26071-22072) and CORE trials if post-operative MRI scans were available within a timeframe of up to 4 weeks before radiotherapy, including both pre- and post-contrast T1w images and at least one T2w sequence (T2w or T2w-FLAIR). Post-operative (residual) pre-radiotherapy contrast-enhanced tumor (CET) volumes and non-enhanced T2w abnormalities (NT2A) tissue volumes were obtained by three-dimensional segmentation. Cox proportional hazard models and Kaplan Meier estimates were used to assess the association of pre-radiotherapy CET/NT2A volume with OS adjusted for known prognostic factors (age, performance status, MGMT status).*

Results: *408 tumor segmentations (MGMT methylated, N=270) were included. Median OS in patients with MGMT methylated tumors was 117 weeks versus 61 weeks in MGMT unmethylated tumors ($p<0.001$) without significant correlation between MGMT methylation status and CET volume. When stratified for MGMT methylation status, higher CET volume (HR 1.020; 95% CI [1.013-1.027]; $p<0.001$) and older age (HR 1.664; 95% CI [1.214-2.281]; $p=0.002$) were significantly associated with shorter OS while NT2A volume and performance status were not.*

Conclusion: *Pre-radiotherapy CET volume was strongly associated with OS in patients receiving radio-/chemotherapy for newly diagnosed glioblastoma stratified by MGMT promoter methylation status.*

3.1 Introduction

Glioblastoma is the most common primary malignant brain tumor in adults. Its growth is fast and infiltrative, often with a high tumor burden and a poor prognosis; the majority of patients succumb to the disease within fewer than 18 months despite maximal surgical resection followed by chemo- and radiation chemotherapy [47]. Contrast enhanced magnetic resonance imaging (MRI) is the gold standard for glioblastoma radiological monitoring, radiotherapy treatment planning, and assisting neurosurgeons to achieve a maximal yet safe resection of the contrast enhanced portion of the tumor [48, 49].

Several studies [50, 51] have investigated the prognostic value of residual contrast-enhanced tumor (CET) volume, suggesting that it is associated with survival in glioblastoma along with the previously known prognostic factors such as age and clinical performance. Additionally, 'supratotal' surgical resection which includes the non-enhancing tumor volume has been suggested because of the benefit on survival outcome in some previously published reports [52, 53, 54]. However, there is still uncertainness on how to delineate non-enhancing tumor parts within the non-enhancing T2w-hyperintense MRI signal tissue abnormalities (NT2A) consisting of a combination of tumor and edema.

MGMT promoter methylation status is an important molecular marker in glioblastoma, since tumors with this methylation have better prognosis. Recently, several studies [55, 56, 57] have re-assessed the prognostic value of CET and NT2A residual volume, adjusting for the molecular profile in accordance to the 2016 WHO classification [58]. However, these studies investigated the prognostic associations of resection in the early post-operative stage, and did not take into consideration the tumor progression occurring in the period between resection and start of radiotherapy.

Thus, in this study we aimed to assess the association between post-operative pre-radiotherapy tumor volume and OS, adjusted for MGMT promoter methylation in patients with newly diagnosed glioblastoma receiving radio-/chemotherapy.

3.2 Methods

3.2.1 Patient inclusion criteria

We retrospectively assessed the data of 810 patients with newly diagnosed glioblastoma. The data were collected in the context of the CENTRIC EORTC 26071-22072 phase 3 and CORE phase 2 trials [59, 60]. These trials aimed to explore the efficacy of cilengitide in newly diagnosed glioblastoma with (CENTRIC trial, N=545) and without (CORE trial, N=265) MGMT gene promoter methylation. No effect of cilengitide on overall survival (OS) was found in either one of these trials. The pooled CENTRIC & CORE database

contains the patients' clinical characteristics, i.e., age (dichotomized as younger than versus older than or equal to 50 years), sex, Eastern Clinical Oncology Group performance score (ECOG score), OS; and tumor characteristics, namely centrally determined MGMT gene promoter methylation status and radiological assessments. MRI data acquired during the trials were collected from participating sites and consisted of all trial scans from the moment of surgery. Pre-operative imaging was not collected. Patient characteristics and eligibility criteria for the respective trials have been reported elsewhere [59, 60], for inclusion of this imaging study the following additional criteria apply:

1. The availability of post-operative MRI scans within a timeframe of up to 4 weeks before the start of the radiotherapy (RTX) treatment (the post-operative MRI performed closest to start of RTX was used; the time between this MRI scan and the start of RTX is the RTX time interval);

2. The availability of the relevant MRI sequences for tumor segmentation: both pre- and post-contrast T1-weighted (T1w respectively post-contrast T1w) images and at least one T2-weighted sequence (T2w or T2w-FLAIR);

3.2.2 Tumor segmentation for volumetric measurements

The MRI scans were first uploaded to an XNAT server and then automatically sorted into a structured format using DeepDicomSort [61] software recognizing the scan type (pre-/post-contrast T1w, T2w, T2w-FLAIR, etcetera). Sorted images were manually checked for the availability and correct identification of required MRI sequences and to exclude scans with severe imaging artifacts.

The images were converted from DICOM to Nifty format using dcm2niix[1] [62], co-registered to the post-contrast T1w scan using Elastix[2] [63, 64], skull-stripped using HD-BET[3] [65] and corrected for MR bias field using N4ITK[4] [66]. HD-GLIO[5] [19] was used for automated tumor segmentation if all four required MR sequences (T1w, T1w+c, T2w, T2w-FLAIR) were available. Two radiological manifestations were segmented:

1. contrast-enhanced tumor (CET) as determined by the pre- and post-contrast T1w sequences;

2. hyperintense areas on T2w/T2w-FLAIR images.

[1] dcm2niix v1.0.20181125
[2] Elastix v4.8
[3] https://github.com/NeuroAI-HD/HD-BET git commit 41ebe0d
[4] N2ITK v1.6
[5] https://github.com/NeuroAI-HD/HD-GLIO v1.5

To obtain the volume of non-enhanced T2 abnormalities (NT2A), known to be a combination of tumor infiltration and edema, CET was subtracted from the total volume of T2w hyperintensity. All segmentations were assessed for quality and acceptance, and manually edited in 3 planes (axial, coronal, sagittal) in case of poor segmentation. In cases where automated segmentation was not possible, due to a missing T2w or T2w-FLAIR scan, segmentation was performed manually (N=34). Manual segmentation and editing was performed using ITK-Snap [67].

3.2.3 Data analysis

All statistical analyses were performed using SPSS [6]. Medians and interquartile ranges were used to express the distribution of continuous data. CET volumes were divided into 5 categories (>2cm^3, 2-5cm^3, 5-15cm^3, and >15cm^3) in line with Wijnenga et al [68]. NT2A volumes were divided into 4 categories according to 1st, 2nd and 3rd quartiles. Additionally, the four-tier classification, taking both CET and non contrast-enhancing tumor into account, for supramaximal resection as proposed by RANOresect was applied [69]. Non-parametric tests were performed for non-normally distributed data, Mann-Whitney U tests were used for comparisons of clinical characteristics between the patients with versus without MGMT promoter methylation. A linear regression model was used to explore the association between CET and NT2A volumes after using Log_{10} transformation. Log-rank tests and Kaplan Meier survival analyses were used to evaluate the association of CET and NT2A volumes and a combination thereof with OS; the derived p-values were corrected for multiple-testing using the Bonferroni method. Cox proportional hazard models were created in multivariate analyses, investigating the association of the clinical characteristics and the CET/NT2A volumes with OS stratified by MGMT promoter methylation status. Since information on IDH mutation status was not routinely collected in the trials but has a known prognostic impact [3], all analyses were repeated in sensitivity analyses including only patients with confirmed IDH wild type glioblastoma. The significance level was set at 5%.

3.3 Results

From the 810 patients in the combined CENTRIC & CORE database, 408 met all the inclusion criteria for this analysis (Figure 3.1). Patient characteristics are described in Table 3.1. Median follow-up period from time of randomization until last follow-up or death was up to 190 weeks during the CENTRIC and CORE trials with a median OS of 96 weeks. Median OS was significantly different ($p< 0.001$) between patients with an MGMT promoter methylated

[6] IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.

tumor (117 weeks) versus patients with an MGMT promoter unmethylated tumor (61 weeks). The majority of patients were 50 years or older (74.5%) and male (53.7%). These characteristics were not significantly different between the patients with versus without MGMT promoter methylation. ECOG performance score was found to be significantly different ($p < 0.001$): patients with MGMT promoter methylated glioblastoma more often had an ECOG score of 0 (i.e. better performance). The median time interval between the MRI scan and the start of RTX was 13 days, which was not significantly different between CENTRIC and CORE trials (p=0.912).

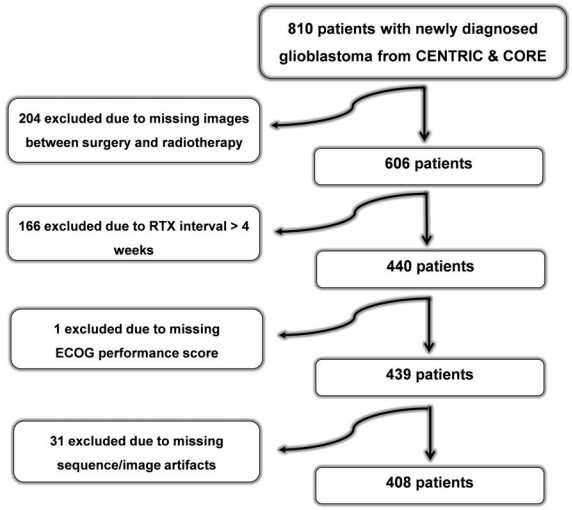


Figure 3.1: Flow chart of patient inclusion.

The median post-operative CET volume was 5.3 cm^3 (p=0.240), and median NTA volume was 19.6 cm^3 (p=0.2300), not significantly different between the MGMT promoter methylated and unmethylated tumors (Table 3.1). The IDH mutational status was known in 195 patients (48%), the vast majority having IDH wild type glioblastoma (n=184; 94%).

3.3.1 Association between contrast-enhanced tumor (CET) and non-enhanced T2w abnormalities (NT2A) volumes

The correlation between the CET and NT2A volumes had a small magnitude (Spearman's $\rho = 0.311$; p<0.001). From the linear regression analysis we found that a 10% increase of CET volume was associated with an 1.8% increase of NT2A (p<0.001; 95%CI [0.124 - 0.255]) (Figure 3.S1).

Table 3.1: Patient Characteristics.

Variables		All patients	MGMT promoter status		P-value
			Methylated	Unmethylated	Mann-Whitney U
		N=408	N=270	N=138	
		(%)	(66.2%)	(33.8%)	
Median OS	weeks [IQR][a]	95.7 [87.2 - 104.2]	116.8 [97.1 - 136.6]	61.1 [55.8 - 66.4]	**<0.001** [b]
Age	<50 years	104 (25.5%)	61 (31.2%)	43 (22.6%)	0.061
Sex	Male	219 (53.7%)	140 (51.9%)	79 (57.2%)	0.302
ECOG Performance [c]	Score 0 [d]	215 (52.7%)	158 (58.5%)	57 (41.3%)	<0.001
Median RTX time interval [e]	days [IQR]	13.0 [8.0 - 18.0]	12.5 [7.0 - 18.2]	13.0 [9.7 - 17.0]	0.912
Median CET volume [f]	cm^3 [IQR]	5.3 [1.6 - 15.3]	4.9 [1.5 - 14.7]	6.5 [1.6 - 16.6]	0.240
Median NT2A volume [g]	cm^3 [IQR]	19.6 [8.7 - 43.1]	18.9 [8.7 - 40.9]	20.9 [9.5 - 46.5]	0.230

[a]IQR= Interquartile range
[b]log rank comparison
[c]ECOG= Eastern Cooperative oncology group
[d]Only patients with scores 0 and 1; no patients with ECOG score 2
[e]RTX=post-operative MRI to start of radiotherapy time interval
[f]CET= Contrast enhanced tumor
[g]NT2A = non-enhanced T2w abnormalities

3.3.2 Overall survival (OS) in relation to MGMT promoter methylation status

We predefined 4 categories of residual CET volumes: <2 cm^3, 2-5 cm^3, 5-15 cm^3, and >15 cm^3. Kaplan Meier survival analysis and log rank comparison was performed for the different categories of CET volumes (Figures 3.2 and 3.3, Table 3.2). Both in patients with MGMT promoter methylated and unmethylated glioblastoma, there was an inverse relationship between CET volumes and OS. In patients with MGMT promoter unmethylated tumors, the difference in OS was significant in patients with CET volumes <2cm^3 compared to patients with CET volumes >5cm^3 and >15cm^3 (82 weeks versus 56 weeks and 47 weeks respectively), while in patients with MGMT promoter methylated tumors this difference was significant between all CET volume categories except for patients with CET volumes of 2-5 cm^3 versus any CET volumes < 15cm^3 and patients with CET volumes >15cm^3 versus 5-15cm^3. Patients with the largest CET volume (>15cm^3) of MGMT promoter methylated tumor still appeared to have better OS compared to patients with the smallest CET volumes (<2cm^3) of MGMT promoter unmethylated tumor, with a median OS of 90 weeks compared to 82 weeks respectively (Table 3.2).

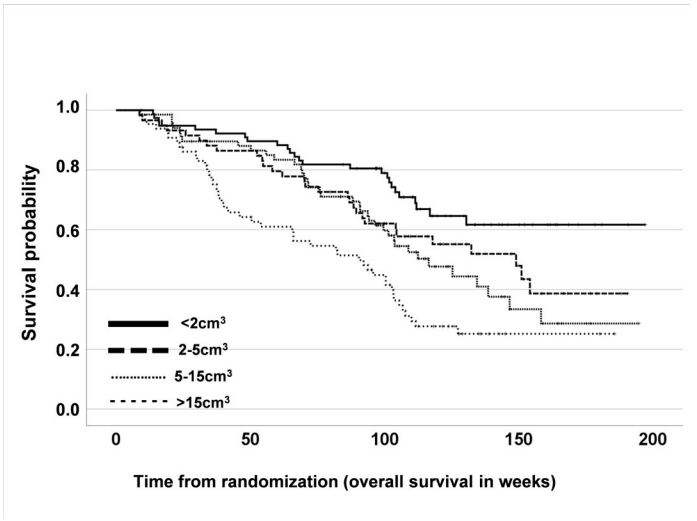


Figure 3.2: Kaplan Meier survival curve of patients with MGMT promoter methylated glioblastoma for different contrast enhanced tumor (CET) volume categories

The extent of non-enhancing residual abnormalities on T2 weighted images (NT2A) did not reveal a significant correlation in OS, irrespective of the MGMT promoter methylation status (Table 3.S1).

The combined CET/NT2A classification showed no significant difference in OS between any of the patients with MGMT promoter unmethylated tumors

Table 3.2: Median overall survival (OS, in weeks) and Log rank comparison (Bonferroni corrected for multiple testing) between contrast-enhanced tumor volume categories.

MGMT promoter status	categories	Patients N=408	OS in weeks	$<2cm^3$ P-value*	$2\text{-}5cm^3$ P-value*	$5\text{-}15cm^3$ P-value*	$>15cm^3$ P-value*
Unmethylated	$<2cm^3$	39	82	-	0.132	**<0.001**	**0.001**
	$2\text{-}5cm^3$	19	74	0.132	-	0.109	0.189
	$5\text{-}15cm^3$	44	56	**<0.001**	0.109	-	0.483
	$>15cm^3$	36	47	**0.001**	0.189	0.483	-
	Overall	138	61				
Methylated	$<2cm^3$	77	-	-	0.066	**0.009**	**<0.001**
	$2\text{-}5cm^3$	59	149	0.066	-	0.454	**0.005**
	$5\text{-}15cm^3$	68	116	**0.009**	0.454	-	0.018
	$>15cm^3$	66	90	**<0.001**	**0.005**	0.018	-
	Overall	270	117				

*The significance level for the Bonferroni adjusted P-value is set at 0.0125.

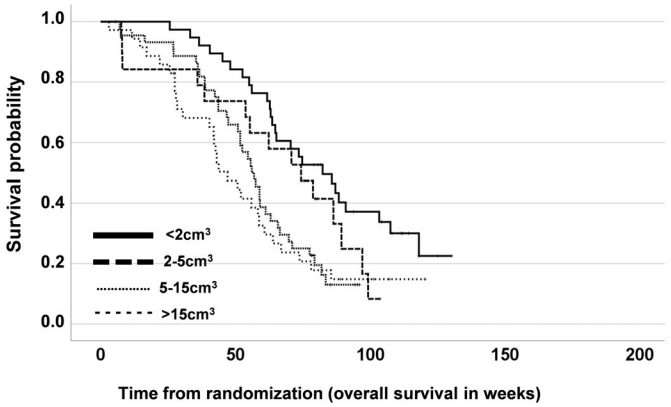


Figure 3.3: Kaplan Meier survival curve of patients with MGMT promoter unmethylated glioblastoma for different contrast enhanced tumor (CET) volume categories

Table 3.3: Multivariate Cox regression model associating CET and NT2A volumes with OS stratified by MGMT promoter methylation status.

Variables	Hazard Ratio [95% CI]	P-value
CET volume[1](cm³)	**1.020 [1.013 - 1.027]**	**<0.001**
NT2A volume[2](cm³)	0.999 [0.995 - 1.003]	0.542
Age	**1.664 [1.214 - 2.281]**	**0.002**
Sex	1.076 [0.830 - 1.394]	0.580
ECOG score[3]	0.964 [0.745 - 1.246]	0.777
RTX-interval[4](days)	0.999 [0.980 - 1.018]	0.902

[1] CET= Contrast enhanced tumor
[2] NT2A = non-enhanced T2w abnormalities
[3] ECOG= Eastern Cooperative oncology group
[4] RTX=post-operative MRI to start of radiotherapy time interval

whereas in patients with MGMT promoter methylated tumors, the difference in OS was significant ($p<0.001$) in patients with maximal CET resection (Class 2) compared to patients with submaximal CET resection (Class 3) (able 3.S2)(Figures 3.S2, 3.S3).

3.3.3 Cox proportional hazard analysis

Stratified by MGMT status, we created multivariate Cox hazard proportional models for CET and NT2A volume. CET volume (in cm^3) showed a highly

significant association with OS with a hazard ratio of 1.020 (95% CI [1.013-1.027]; p<0.001) in addition to age which was also significantly associated with OS (HR 1.664; 95% CI [1.214-2.281]; p=0.002). NT2A volume, sex, ECOG score and RTX interval were not significantly associated with OS (Table 3.3).

3.3.4 Sensitivity analysis

Sensitivity analyses performed in patients with confirmed IDH wild type glioblastoma only (N=184) are provided in Tables 3.S3 and 3.S4. The analyses for associating CET and NT2A showed no different results compared to the main analysis. The Cox regression multivariate models yielded the same results as those in the whole cohort.

3.4 Discussion

In this retrospective study with clinical, genetic and imaging data from two large randomized clinical trials, we found that post-operative pre-radiotherapy CET volume was strongly associated with OS in patients receiving radio-/chemotherapy and that both patients with MGMT promoter methylated and unmethylated tumors fared better the larger the extent of resection (i.e., lower post-operative pre-treatment CET volumes). Patients with the smallest MGMT promoter unmethylated CET volume (<2 cm^3), however, still appeared to have worse survival than patients with the largest MGMT promoter methylated CET volume (>15cm^3) suggesting that the responsiveness to chemotherapy is of greater importance than the extent of resection.

MGMT promoter methylation has been advanced as an important predictive biomarker in neuro-oncology because of the benefit on survival derived from temozolomide in newly diagnosed glioblastoma as observed by Hegi et al [70]. There has since been a growing body of evidence showing the prognostic impact of the MGMT promoter gene on survival in patients with malignant glioma [71, 72, 73]. In Binabaj et al.'s meta-analysis, which included thirty-four studies reporting the association of MGMT promotor methylation status with OS and progression free survival (PFS), MGMT promotor methylation was found to be significantly correlated with favorable outcome on OS (p=0.001). Within our cohort and consistent with the findings of preceding literature, we observed that MGMT methylation was associated with longer OS. Our results suggested that patients with minimal residual CET volume of MGMT promotor unmethylated tumor had almost the same, possibly even somewhat worse, survival as patients with the largest CET volume of MGMT promotor methylated tumor.

Stratified for MGMT promoter methylation status, CET volume was a prognostic factor of OS while adjusting for previously known prognostic factors (age, sex, performance score, RTX interval). We observed that OS in the

four post-operative pre-radiotherapy CET volume categories was significantly different. Previous literature already clarified that post-surgical CET volume negatively impacts the survival outcome in glioblastoma [49, 50]. Ellingson et al showed in a multi-center study of 1,511 patients with newly diagnosed glioblastoma that post-surgical CET volume significantly influenced survival in glioblastoma independently from clinical covariates and the type of therapy employed [50]. Their cohort represents the largest study to evaluate the association of CET volume with OS. However, this study has as a limitation that the MGMT promotor methylation status was not known and thus not taken into consideration.

In contrast to our our study, previous studies investigated the tumor burden utilizing the post-operative MR images collected in the early stage after resection. The assessment of the tumor residual volume at this time point (directly post-surgical images) might influence the clinical outcome and misestimate the prognostic value, because it does not take into consideration tumor growth or tumor recurrence which may occur after surgery prior to the initiation of treatment [74, 75]. In a small study Yamashita et al used computed tomography scanning to examine the tumor growth rate and found that malignant gliomas can double in mass in around 15.0 to 21.1 days [74]. Furthermore, Pirzkall et al assessed the incidence and tumor regrowth between surgery and start of radiotherapy. As many as 53% of their study cohort showed a new contrast-enhanced lesion or increased volume [75]. The difference in tumor burden between the directly post-surgically acquired MRI and initiation of treatment can thus be substantial and is accounted for in our study by using the MRI scan most closely acquired prior to radiotherapy.

There is a growing understanding of the prognostic importance of non-enhanced tissue abnormality in glioblastoma to optimize current treatment strategies and ultimately prolong survival [51, 53, 54, 55, 75, 76]. Lasocki et al reported that non-contrast enhanced lesions in peripherally located glioblastoma were associated with worse survival compared to those with peripheral tumors without this component[77]. Grabowski et al demonstrated the predictive value of T2w/T2w-FLAIR residual volume on survival in both univariate and multivariate analysis [51]. In accordance to these findings, Kotrotsou et al revealed in their multi-center study that high postoperative residual non-enhanced tumor volume ($>70\text{cm}^3$) corresponded to a worse prognosis while patients with low postoperative residual non-enhanced tumor volume had a significant survival benefit (5.6 months)[54]. Because of this increasingly recognized importance of not only contrast-enhanced, but also non-enhanced tumor volume for survival in glioblastoma, we assessed the association of both CET and NT2A volume as well as their combination with OS. We found that CET was significantly correlated with the NT2A volume, showing that 10% increase of CET volume was associated with 1.8% increase of NT2A volume. However, in the multivariate Cox proportional after adjusting for the clinical factors

and MGMT promoter methylation status, we found that only CET volume and the combined CET/NT2A classification were significantly associated with OS, and not the NT2A volume by itself. This disparity between our results and the previous findings might be related to assessing NT2A volume just prior to initiating radiotherapy treatment while the previous studies evaluated the T2w/FLAIR abnormalities as an early post-operative resection percentage in comparison to the pre-operative volume. At the later post-operative stage we assessed, some of the initially non-enhanced tumor tissue may have progressed to enhance. In addition, there is the inherent limitation in any volumetric assessment of the non-enhanced tumor on T2w/T2w-FLAIR given the difficulties to discriminate the non-enhanced tumor region from the other entities causing hyperintense signal intensity on T2w/T2w-FLAIR. Finally, it could be hypothesized that the small percentage of IDH mutated tumors could have influenced the results, as these tumors have proportionally larger areas of NT2A and are associated with better prognosis than IDH wild type glioblastoma. However, our sensitivity analyses performed in patients with confirmed IDH wild type glioblastoma, showing the same results, make this less likely.

Our study had some limitations. IDH mutational status was only known in a proportion of patients, due to the fact that IDH was not part of the diagnostic criteria for glioblastoma when the trials were performed. As mentioned above, this also resulted in a small percentage of IDH mutated tumor in the study cohort, with different prognosis from what is now considered glioblastoma according to the WHO 2021 criteria[78]. We addressed this issue by performing sensitivity analyses in patients with confirmed IDH wild type tumor only, which yielded similar results to the main analysis.

A further limitation is the retrospective nature of the study, leading to inherent difficulties of including a homogenous cohort of patients controlling for all the prognostic factors. However, this study concerned data from two prospectively conducted clinical trials in which such prognostic factors were also important for the primary outcome. The groups were well-balanced in terms of age and sex, but a larger proportion of patients with MGMT promoter methylated tumors had a better performance status introducing some bias. Therefore, clinical characteristics were adjusted for in the survival analysis. Also, this limitation is offset by the large size of the study population.

Finally, the MRI scan acquisition was heterogeneous due to the multicenter nature of these trials, being performed before standardized imaging protocols were implemented. This concerned both the timing with respect to the surgery and radiotherapy, and the MRI acquisition. The latter was addressed by meticulously checking and correcting all tumor segmentations. The former raises the question whether tumor growth could have occurred between the surgery and the pre-radiotherapy scan. However, there was no difference between the patient groups in the time interval between the start of

radiotherapy and the pre-treatment scan, which was a median of only 13 days.

In conclusion, we found that lower pre-radiotherapy contrast-enhanced tumor volume after surgery was associated with longer OS. While MGMT promoter methylation was clearly associated with better survival, both patients with MGMT promoter methylated glioblastoma and those with an MGMT promoter unmethylated tumor fared better with lower pre-radiotherapy tumor volumes.

3.5 Supplementary materials

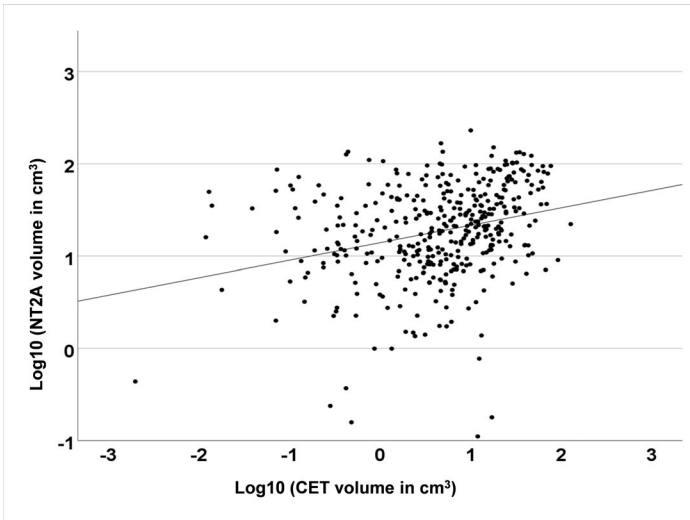


Figure 3.S1: Scatter plot of contrast enhanced tumor (CET) and non-enhanced T2-weighted abnormalities (NT2A) volumes after Log10 transformation

Table 3.S1: Median overall survival (OS, in weeks) and Log rank comparison (Bonferroni corrected for multiple testing) between NT2A volume categories.

MGMT promoter status	categories	OS in weeks	<9cm3 P-value*	9-19cm3 P-value*	19-43cm3 P-value*	>43cm3 P-value*
Unmethylated	<9cm3	63.0	-	0.198	0.826	0.749
	9-19cm3	58.8	0.198	-	0.792	0.896
	19-43cm3	62.2	0.826	0.792	-	0.968
	>43cm3	30.749	0.896	0.968	-	
	Overall	61.1				
Methylated	<9cm3	132.2	-	0.444	0.674	0.217
	9-19cm3	125.2	0.444	-	0.829	0.633
	19-43cm3	146.5	0.674	0.829	-	0.526
	>43cm3	110.0	0.217	0.633	0.526	
	Overall	116.8				

*The significance level for the Bonferroni adjusted P-value is set at 0.0125.

Table 3.S2: Median overall survival (OS, in weeks) and Log rank comparison between the combined CET/NT2A categories according to RANOresect [69].

MGMT promoter status	categories	Patients N=407[a]	OS in weeks	RANO Class 2 P-value*	RANO Class 3 P-value*
Unmethylated	RANO Class 2	24	82.2	-	0.182
	RANO Class 3	114	58.8	0.182	-
	Overall	138	61.1		
Methylated	RANO Class 2	58	N.A.[b]	-	**<0.001**
	RANO Class 3	211	106.7	**<0.001**	-
	Overall	270	116.8	-	

*The significance level for the Bonferroni adjusted P-value is set at 0.0125.

[a]RANO Class 1 was excluded: N=1

[b]survival curve did not reach 0.5: 69% were still alive (40 out of 58) at end of follow-up. Of those patients that were no longer alive, 10 had a short OS of <69 weeks.

Table 3.S3: Patient characteristics with confirmed IDH-wild type glioblastoma.

Variables		**All patients**	**MGMT promoter status**		**P-value**
			Methylated	Unmethylated	Mann-Whitney U
		N=184	N=123	N=61	
		(%)	(66.8%)	(33.2%)	
Median OS	weeks [IQR] [a]	89.4	105.4	57.4	**<0.001** [b]
Age	<50 years	45 (24.5%)	26 (21.1%)	19 (31.1%)	0.138
Sex	Male	98 (53.3%)	64 (52%)	34 (55.7%)	0.636
ECOG Performance [c]	Score 0 [d]	95 (51.6%)	75 (61%)	20 (32.8%)	**<0.001**
Median RTX time interval [e]	days [IQR]	13.0 [9.0 - 18.0]	14.0 [10- 20.0]	12.0 [7.5 - 14.0]	0.020
Median CET volume [f]	cm^3 [IQR]	4.8 [1.6-13.9]	3.4 [1.3-14.1]	9.0 [2.3-13.5]	0.073
Median NT2A volume [g]	cm^3 [IQR]	22.0 [8.8-44.8]	21.7 [8.6-44.9]	25.5 [9.0-45.4]	0.581

[a] IQR= Interquartile range
[b] log rank comparison
[c] ECOG= Eastern Cooperative oncology group
[d] Only patients with scores 0 and 1; no patients with ECOG score 2
[e] RTX=post-operative MRI to start of radiotherapy time interval
[f] CET= Contrast enhanced tumor
[g] NT2A = non-enhanced T2w abnormalities

Table 3.S4: Multivariate Cox regression model associating CET and NT2A volumes with OS stratified by MGMT promoter methylation status for confirmed IDH-wild type glioblastoma.

Variables	Hazard Ratio [95% CI]	P-value
CET volume[1](cm3)	**1.024 [1.013-1.040]**	**<0.001**
NT2A volume[2](cm3)	0.995 [0.987-1.003]	0.237
Age	**1.676 [1.035-2.655]**	**0.036**
Sex	1.220 [0.835-1.781]	0.303
ECOG score[3]	0.892 [0.603-1.320]	0.568
RTX-interval[4](days)	1.016 [0.985-1.048]	0.301

[1] CET= Contrast enhanced tumor
[2] NT2A = non-enhanced T2w abnormalities
[3] ECOG= Eastern Cooperative oncology group
[4] RTX=post-operative MRI to start of radiotherapy time interval

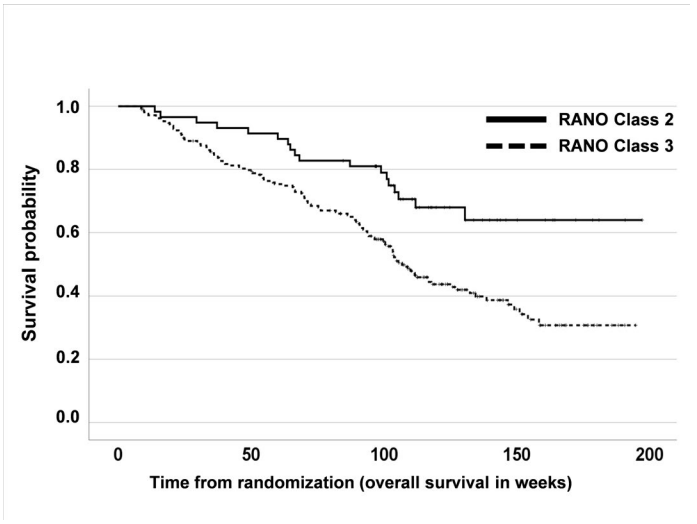


Figure 3.S2: Kaplan Meier survival curve of patients with MGMT promoter methylated glioblastoma for the combined CET/NT2A categories according to the RANOresect [69] classification.

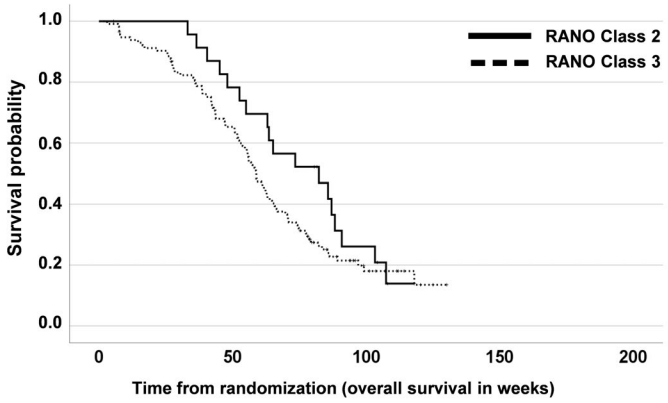


Figure 3.S3: Kaplan Meier survival curve of patients with MGMT promoter unmethylated glioblastoma for the combined CET/NT2A categories according to the RANOresect [69] classification.

4

EASE: Clinical Implementation of Automated Tumor Segmentation and Volume Quantification for Adult Low-Grade Glioma

It is through science that we prove, but through intuition that we discover.

– Jules Henri Poincaré

Based on: **Karin A. van Garderen**, S. R. van der Voort, A. Versteeg, M. Koek, A. Gutierrez, M. van Straten, M. Rentmeester, S. Klein, and M. Smits, "EASE: Clinical implementation of automated tumor segmentation and volume quantification for adult low-grade glioma," *Frontiers in Medicine*, p. 1791, Oct. 2021

Abstract

The growth rate of non-enhancing low-grade glioma has prognostic value for both malignant progression and survival, but quantification of growth is difficult due to the irregular shape of the tumor. Volumetric assessment could provide a reliable quantification of tumor growth, but is only feasible if fully automated. Recent advances in automated tumor segmentation have made such a volume quantification possible, and this work describes the clinical implementation of automated volume quantification in an application named EASE: Erasmus Automated SEgmentation. The visual quality control of segmentations by the radiologist is an important step in this process, as errors in the segmentation are still possible. Additionally, to ensure patient safety and quality of care, protocols were established for the usage of volume measurements in clinical diagnosis and for future updates to the algorithm. Upon the introduction of EASE into clinical practice, we evaluated the individual segmentation success rate and impact on diagnosis. In its first three months of usage, it was applied to a total of 55 patients, and in 36 of those the radiologist was able to make a volume-based diagnosis using three successful consecutive measurements from EASE. In all cases the volume-based diagnosis was in line with the conventional visual diagnosis. This first cautious introduction of EASE in our clinic is a valuable step in the translation of automatic segmentation methods to clinical practice.

4.1 Introduction

Magnetic resonance (MR) imaging plays a key role in the management of low-grade glioma (LGG) as a method for measuring treatment response and for regular surveillance during periods of watchful waiting. LGG are known to show constant slow growth [79], until – in adults – they inevitably transform to a more malignant type. The early growth rate of the T2-weighted hyperintense region is a known prognostic factor for malignant progression [80] and overall survival [81], so the reliable quantification of growth may be a valuable tool for clinical decision making [82]. However, due to the anisotropic growth and irregular size it can be difficult to evaluate slow growth on consecutive imaging using a visual assessment or 2D measurement [83]. Volumetric measurements are preferred for the assessment of early growth due to their reproducibility and sensitivity to subtle changes [6], but a manual segmentation would require an effort that is unrealistic in clinical practice.

Automatic segmentation of glioma has shown great advances in recent years due to the release of public datasets and the development of artificial intelligence [20]. A recent method described in Kickingereder et al. [19] has been shown to be a reliable alternative for the prognostication of glioma, comparable to the current clinical standard of 2D measurement according to the RANO criteria. Although these criteria apply specifically to high-grade glioma and the measurement of enhancing tumor [6], the performance evaluation in Kickingereder et al. also shows an almost perfect quantification of non-enhancing abnormalities on T2-weighted FLAIR imaging. This makes it potentially suitable for the assessment of volume changes in non-enhancing low-grade glioma.

Due to the clear clinical need of volume quantification in LGG, we decided to implement a segmentation pipeline and integrate it in the existing clinical workflow of the Brain Tumor Center, Erasmus MC Cancer Institute, Rotterdam. This introduced a new measurement tool in the radiologists' toolbox, which we named EASE: Erasmus Automated SEgmentation. With a new tool come potential risks to patient safety and quality of care, which need to be considered in the design of the software and protocols for its use.

For the clinical implementation of this segmentation pipeline, we identified potential risks and practical challenges. The main concern was that of incorrect tumor segmentations resulting in incorrect volume measurements. Further risks were found in software updates over time, potentially leading to unreliable or inconsistent volume measurements, and finally in the incorrect interpretation of volume measurements at time of diagnosis. These risks and the design choices to address these are described in more detail in section 4.2 and 4.3, and an overview is shown in Table 4.1.

This work describes the design of both the technical implementation of EASE and its integration into the clinical workflow, to ensure quality of results

and prevent incorrect interpretation of the resulting volume measurements. Furthermore, an initial evaluation of the software was performed in which both the success rate and clinical impact of the volumetric assessment were measured.

4.2 Materials and equipment

This section describes the software implementation of EASE. Each scan assessed with EASE goes through a number of processing steps: 1) The images (pre- and postcontrast T1-weighted, T2-weighted and T2-weighted FLAIR) are received and stored (section 4.2.1) ; 2) The segmentation is generated (4.2.2); 3) The segmentation is checked by a radiologist (section 4.2.3) ; 4) A report is generated and sent back to the PACS (section 4.2.4). A data and state management tool is used to manage the state of each scan and launch processing tasks, in order to balance the workload on the server and enable monitoring of errors in the process. The global software design and data flow are shown in Figure 4.1. The software components for data management, processing and annotation are all open-source, both as separate components and as an adaptable containerized framework using Docker [84].

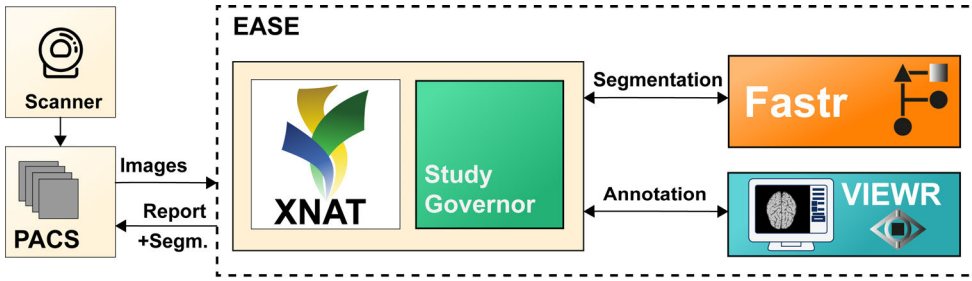


Figure 4.1: Illustration of the different components of EASE. Images are sent from the PACS and added to the XNAT [85] database. The data and state manager (Study Governor) triggers the processing using Fastr [86]. After successful processing, the results can be checked in the VIEWR. A report, including the delineations, is sent back to the PACS.

Table 4.1: Overview of identified risks and measures to address those risks.

RISK	MEASURE
Segmentation errors	Quality check in annotation interface (4.2.3)
Inconsistencies due to updates	Reference dataset and version control (4.3.2)
Incorrect interpretation of volumes	Design guidelines for usage (4.3.1)
	Storage of segmentations in PACS (4.2.4)

4.2.1 Data management

The scan is sent from the PACS (Vue PACS, Carestream Health, v12.2.2.1025) to a dedicated workstation where the scan protocol is automatically checked and the required MR sequences (see section 4.2.2) are automatically selected. The images are then stored on a local XNAT database (v1.7) [85], which forms the common database for all further processing steps. The images are stored for a maximum of 6 months to allow for monitoring of the algorithm performance over time, while avoiding unnecessary risk to patient privacy.

4.2.2 Segmentation

The input for the segmentation consists of four MR sequences: pre- and post-contrast T1-weighted, T2-weighted and T2-weighted FLAIR imaging. The pipeline consists of the following steps: first, the images are converted from DICOM to Nifty images using dcm2niix (v1.0.20171215) [62] and co-registered to the postcontrast T1-weighted scan using Elastix (v4.8) [87]. Then, they are skull-stripped using HD-BET (git commit 98339a2) [65] and MR bias fields are corrected using N4ITK (using SimpleITK v2.0.2 for Python) [88]. The resulting images are used as input for HD-GLIO (v1.5) [17, 19], producing the final delineation of both the enhancing tumor and non-enhancing hyperintensities on T2-weighted FLAIR. Although bias correction is not included in the recommended preprocessing for HD-GLIO, initial tests showed that this improves the performance of the segmentations for scans from our clinic. This pipeline was found, in initial experiments, to perform well on representative images in our center. The Fastr workflow engine (v3.2) [86] was used to integrate these different tools in a robust pipeline.

4.2.3 Quality assessment

Although the underlying segmentation algorithm, HD-GLIO, was evaluated in a large number of scans and found to be reliable [19], an initial evaluation in our center found that our pipeline does not provide perfect segmentations in all scans of low-grade glioma (see section 3.2). The manual quality assessment of segmentations is therefore essential for the use of EASE in clinical practice. To enable this assessment within a clinical workflow, a dedicated interface was developed for the radiologist to easily assess the segmentation. The main purpose of the quality assessment is to prevent failed segmentations from being used for a volume-based diagnosis. Additionally, the same quality assessment can be used for the initial validation of the algorithm, prospective evaluation, and continuous monitoring of the segmentation quality. Therefore, besides a binary check on the usability of the segmentation, a more refined quality assessment scoring system was included. Important factors in the design were usability and prevention of human errors. The interface shows

the segmentation as an overlay over all four co-registered scans, and allows for basic interaction through scrolling, manipulation of the contrast, and selecting sequences and imaging planes. The radiologist is asked to evaluate the segmentation both in a binary way (ACCEPTABLE/UNACCEPTABLE) and on an ordinal scale (rating of 1-5, where 5 is the best score). As an additional sanity check, specifically to prevent unnoticed false positives, the interface also lists the number of connected components in the segmentation together with their volumes. Segmentations deemed UNACCEPTABLE cannot be used for diagnosis. A screenshot of the interface is shown in Figure 4.2.

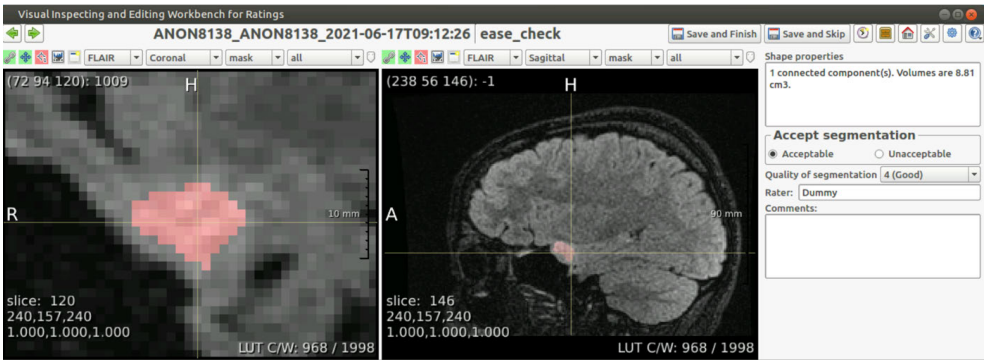


Figure 4.2: Screenshot of the annotation interface. Both image panels can be controlled to show different scan sequences, imaging planes, to change the contrast, to zoom in or out, or to set the overlay transparency. Besides the required annotation of quality, the panel on the right shows the volume of each connected component in the segmentation, and allows for free text comments that are included in the report.

4.2.4 Reporting

Results of the EASE assessment are sent back to the PACS in the form of a report (see Figure 4.3) exported as DICOM file. This report contains the quality assessment, current software version and details of the scan session. Volume measurements are included only if the segmentation is deemed acceptable, to make sure rejected segmentations are not used for diagnosis. In addition to the report, the segmentations are shown as delineations on the T2-weighted FLAIR and post-contrast T1-weighted scan. It would have been possible to store results as a DICOM Structured Report and DICOM SEG respectively, but conventional DICOM images were preferred as not all viewers used in the clinic supported these formats.

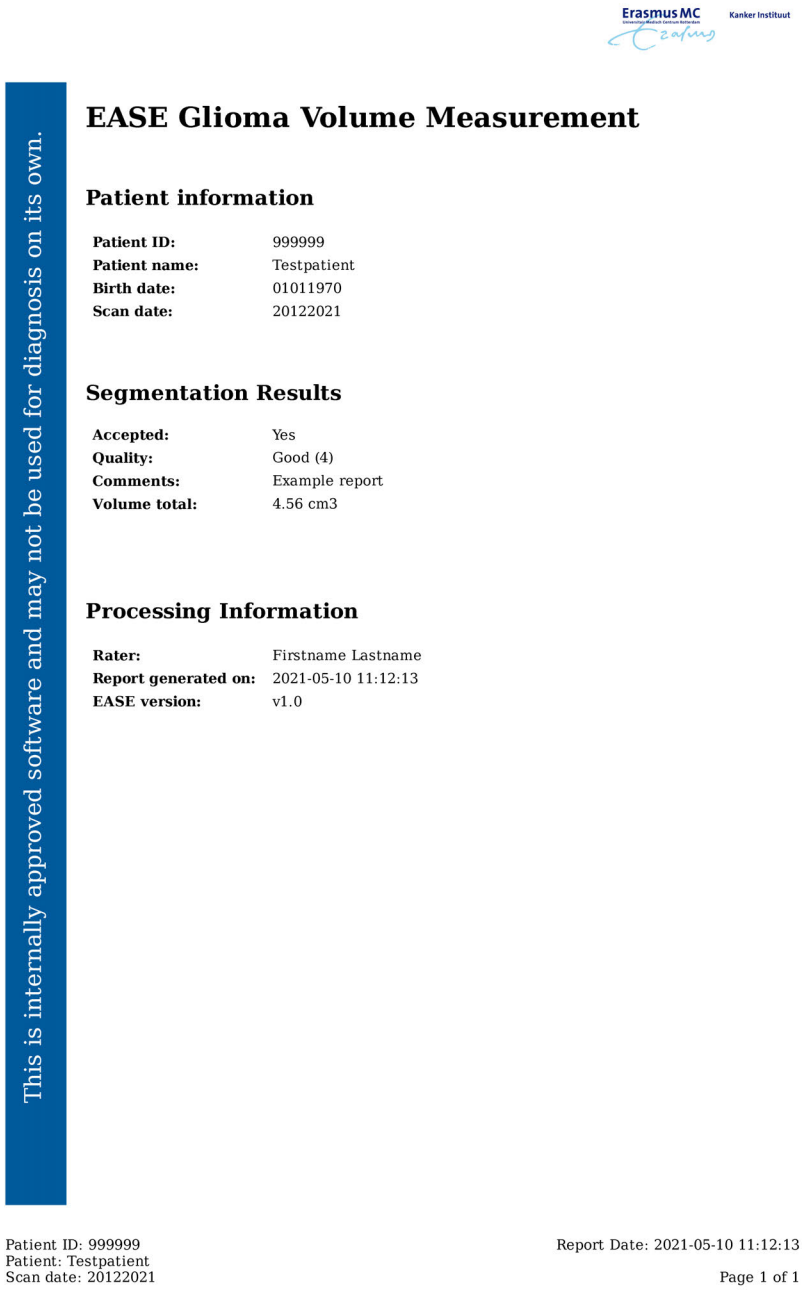


Figure 4.3: Example of the EASE report.

4.3 Methods

This section describes the protocols for usage of EASE in diagnosis (section 4.3.1), the measures for software validation and version control (section 4.3.2), and the method for initial evaluation in clinical practice (section 4.3.3).

4.3.1 Diagnosis

The purpose of volume measurements produced by EASE is to assess therapy response or progression by estimating tumor growth. The standard clinical procedure for estimating growth is to compare the current measurement to two previous measurements and measure the difference in size, with a manual quantitative measurement of two perpendicular diameters if possible, as described in the RANO guidelines (6). The EASE software provides an automated 3D alternative to the existing measurement. However, as the EASE software has not been tested extensively in this setting, we decided that the existing 2D method should still be performed before using EASE. The volume measurements provided by EASE can lead to further insight and even a different diagnosis, but if there is a discrepancy between the two assessment methods leading to a different conclusion, the diagnosis should be made in consensus with a second radiologist. The following protocol is in place for the interpretation of automatic volume measurements in clinical practice. The complete workflow is illustrated in Figure 4.4.

1. Two prior reference scans are selected for the assessment (in addition to the current scan).
2. The radiologist assesses the scan using the routine 2D RANO measurement.
3. EASE is applied to all three scans and the segmentations are checked for quality and acceptance. If any scan was already processed and checked previously, this does not have to be repeated.
4. If any of the segmentations are rejected, a volumetric assessment is not possible.
5. If all segmentations are accepted, the volumes can be compared.
6. If the volume measurements lead to a change in interpretation compared to the initial assessment after step 2, a second radiologist must be consulted. This second rater first forms an independent opinion of the diagnosis. If this is in line with the first radiologist's opinion, this finalizes the conclusion. If not, both radiologists discuss together how their findings are best described in the report, clearly indicating the uncertainty regarding the findings.

The radiological report clearly describes how each assessment is done (2D RANO, 3D EASE) and how the conclusion is reached. If there was a discrepancy between the two methods, leading to a consensus diagnosis, this should be reflected in the report.

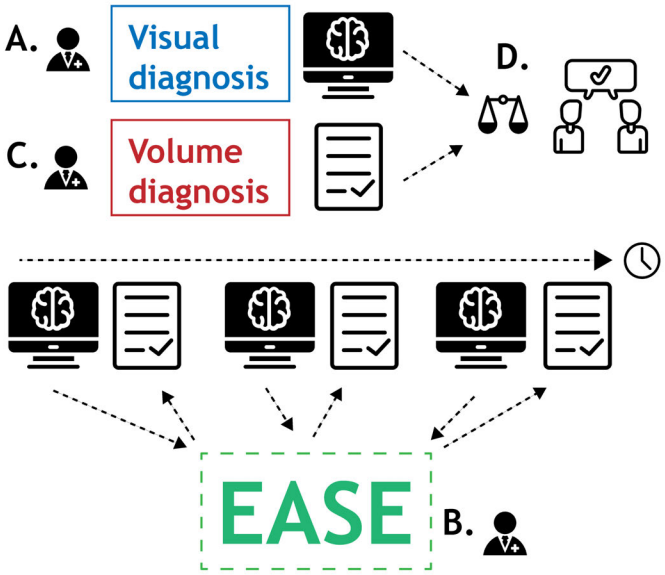


Figure 4.4: Graphical representation of the protocol for the use of EASE in clinical practice. A) Radiologist applies conventional method for visual diagnosis; B) Segmentations are produced by EASE and assessed separately; C) Radiologist interprets volume measurements; D) If volume measurements lead to change in diagnosis, a second radiologist is consulted for a consensus conclusion.

4.3.2 Validation and version control

Before deploying the EASE workflow/pipeline, and after any subsequent update, the segmentation quality should be tested in a reference dataset that is representative of the target domain. For this purpose, 20 scans were selected of patients with non-enhancing LGG. All sessions were surveillance scans of patients who had undergone surgical resection, but no further treatment, of LGG. For these scans, the same quality assessment as described in section 4.2.3 was performed by an experienced neuroradiologist.

It is essential that updates to the software do not cause a bias in volume that might skew the diagnosis. Therefore, a protocol for software updates was established that allows updates of the processing pipeline while ensuring the

continued quality and consistency of the volume measurements. The protocol is as follows:

1. In case of an update, the reference dataset of 20 segmentations is processed again with EASE.

2. The segmentation results are compared to earlier versions of the software. If there is no change in the segmentation, the update can be deployed.

3. If there is a change in results, the manual validation is repeated with the new results.

4. If the qualitative scores are equal or improved with respect to the previous version, the update can be deployed.

5. If the update causes substantial differences in volume (defined as a difference >25%) in any of the accepted segmentations in the reference dataset, the new version is considered incompatible with previous versions and volume results cannot be compared between versions. A warning is included in subsequent EASE reports, so that radiologists know when they have to re-assess previously segmented reference scans with the updated version of EASE.

4.3.3 Evaluation in clinical practice

To evaluate the impact of automated segmentation and volume quantification, an observational study was performed for three months from first introduction of the software in the clinic. The study protocol was reviewed and approved by the internal review board (MEC-2021-0530). Users were asked to complete a survey after each patient in whom EASE was applied, to measure the success rate of EASE in practice and the rate at which volume quantification leads to a change in diagnosis.

To assess the treatment response or tumor progression in non-enhancing LGG three consecutive volume measurements are required, as the standard clinical procedure is to compare the current scan to two former scans. Therefore, patients were excluded if EASE was applied to the first scan after surgery. Furthermore, patients were excluded if any contrast enhancement was found, which would automatically lead to a diagnosis of tumor progression irrespective of volume measurements.

For each of the included patients, the radiologist was first asked whether EASE had led to a successful diagnosis. Although the success rate of a single segmentation can be extracted from the quality assessments made in the user interface, the success of a full diagnosis requires three accepted segmentations from the same patient. If the diagnosis was unsuccessful, the user was asked to submit the reason for failure.

When the volumetric diagnosis was successful, the radiologist was asked to categorize both the visual (2D) diagnosis and the volume-based diagnosis (through EASE) as progression, stable disease or treatment response. These results, combined with the quality assessments made in EASE for the individual scans, were used to measure the success rate of EASE and the impact on the clinical diagnosis. The full user survey is shown in Figure 4.5 in the form of a flowchart.

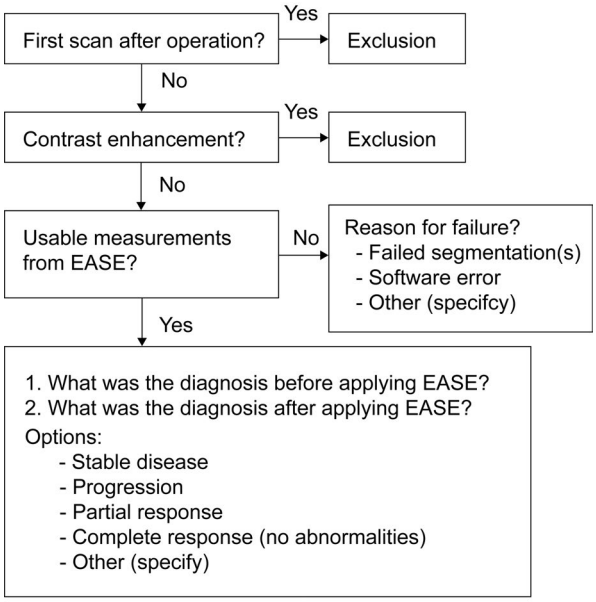


Figure 4.5: Flowchart of the survey on the usage of EASE in clinical practice.

Additionally, for the purpose of a quantitative comparison, measurements were made according to the 2D RANO-LGG guidelines [6] if at all possible, measuring two perpendicular diameters of the lesion. As these lesions are often irregular in shape, the diameters were measured in the portion of the lesion that could be measured most reliably.

4.4 Results

Of the 20 scans in the reference set, which were processed and evaluated before deployment of EASE, 13 (65%) were considered acceptable for clinical volume measurement. The quality scores are summarized in Table 4.2.

EASE was released for local use in Erasmus MC on 25 May 2021, and the evaluation in clinical practice was performed from 1 June 2021 until 19 August 2021.

Table 4.2: Results of reference dataset of 20 MR scans at initial release of EASE. Scans were annotated for acceptability and quality by an experienced neuroradiologist.

Acceptance		
	ACCEPTABLE	13 (65%)
	UNACCEPTABLE	7 (35%)
Quality		
	Perfect (5)	0 (0%)
	Good (4)	10 (50%)
	Fair (3)	6 (30%)
	Poor (2)	2 (10%)
	Terrible (1)	2 (10%)

During the evaluation period, 55 patients were included in the clinical evaluation, meaning that their visual diagnosis was performed and a volume-based diagnosis was attempted. The patient characteristics are summarized in Table 4.3. A successful diagnosis requires three consecutive scans per patient, and in total 162 scans were segmented by EASE and checked by a radiologist. In one of the patients, the two reference scans were not submitted to EASE after the first segmentation was already rejected and in another scan the segmentation failed due to a software error.

Of the 162 segmentations generated by EASE, 124 (77%) were accepted by the radiologist. The distribution of quality scores can be found in Table 4.4. A successful volume-based diagnosis was reached in 36 out of 55 patients. In all patients where volume-based diagnosis was successful, the volume-based diagnosis made by the radiologist was the same as the conventional visual diagnosis, even though in some cases there was a discrepancy between 2D and 3D measurements as shown in Figure 4.7. Figure 4.6 shows an overview of the volume differences detected by EASE, separated by diagnosis (stable disease vs. progression). Figure 7 shows a comparison to the 2D RANO measurements for those patients in whom both measurements were possible. Three patients are not included in this figure because the lesion was too small to measure according to RANO guidelines. In four patients, EASE measurements indicated a volume increase of more than 40% while the final diagnosis was SD. These differences in volume could be explained by inconsistencies between the segmentations, possibly caused by differences in intensities on T2-FLAIR, and therefore the radiologist maintained the original visual diagnosis of SD. There were no other reported reasons for considering volumetric measurements longitudinally unreliable.

Of the failed cases, 19 could be attributed to the rejection of one of the segmentations and two failed diagnoses were attributed to a different reason. Specifically, in one case a segmentation was missing due to a software error, and in another case all segmentations were accepted by the radiologist but the

Table 4.3: Characteristics of 55 patients included in the evaluation of EASE in clinical practice.

Patient characteristics: (total)	55
Age (years)	
median (min - max)	54 (26 - 76)
Sex	
Female	24
Male	31
Tumor type	
Oligodendroglioma	19
Astrocytoma	25
Oligo-astrocytoma	2
Presumed low-grade glioma (no tissue diagnosis)	9
Time after surgery (months)	
median (min - max)	80 (5 – 307)
Time after last treatment (months)	
median (min - max)	67 (5 – 307)
Treatment	
Radiotherapy	33
Chemotherapy	33
Surgical resection	41
Time between scans from current scan (months)	
vs. first reference scan, median (min - max)	14 (7 – 32)
vs. second reference scan, median (min - max)	7 (3 – 20)
Tumor volume found in successful diagnosis (mL)	
median (min – max)	13.2 (1.3 – 77.1)
Oligodendroglioma, median (min – max)	18.3(2.1 – 77.1)
Astrocytoma, median (min – max)	16.1(2.1 – 60.0)
Oligo-astrocytoma	27.8 (9.5 – 46.1)
Presumed low-grade glioma, median (min – max)	2.0 (1.3 – 14.9)

Table 4.4: Results of annotations entered in EASE in clinical practice. During the first three months of usage, 162 scans were annotated for acceptability and quality by 5 different radiologists.

Acceptance		
	ACCEPTABLE	124 (77%)
	UNACCEPTABLE	38 (23%)
Quality		
	Perfect (5)	15 (9%)
	Good (4)	87 (54%)
	Fair (3)	33 (20%)
	Poor (2)	17 (10%)
	Terrible (1)	10 (6%)

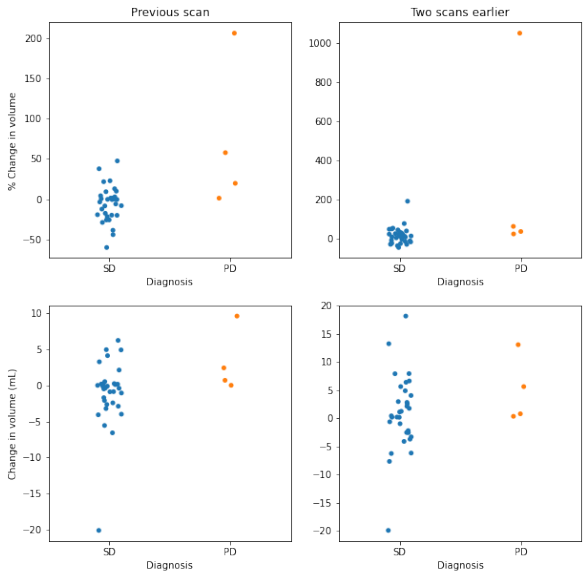


Figure 4.6: Overview of volume changes in successful volume-based diagnosis. Changes per patient with respect to the previous scan (left) and two scans earlier (right). Values are given in percentage change (top) and change in volume (bottom), separated by diagnoses categorized as stable disease (SD) and progressive disease (PD).

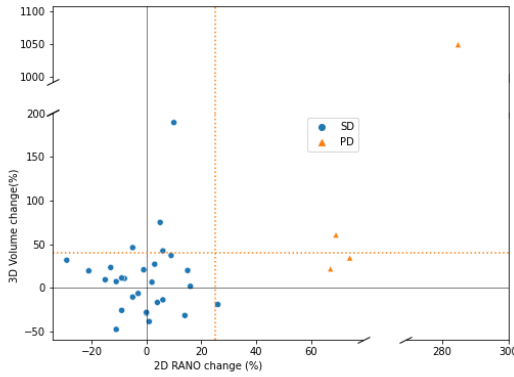


Figure 4.7: Comparison of measurements using EASE (volume) and 2D RANO (product of two diameters), in percentage change with respect to the first (t-2) scan, for patients where both measurements were successful (33 patients). Dotted lines indicate the recommended thresholds for diagnosis of PD.

final volume results were considered unusable due to inconsistencies between the segmentations across the three timepoints. Figure 4.8 shows examples of segmentations made by EASE: two consecutive delineations that were considered inconsistent and two consecutive delineations from a successful volume-based diagnosis.

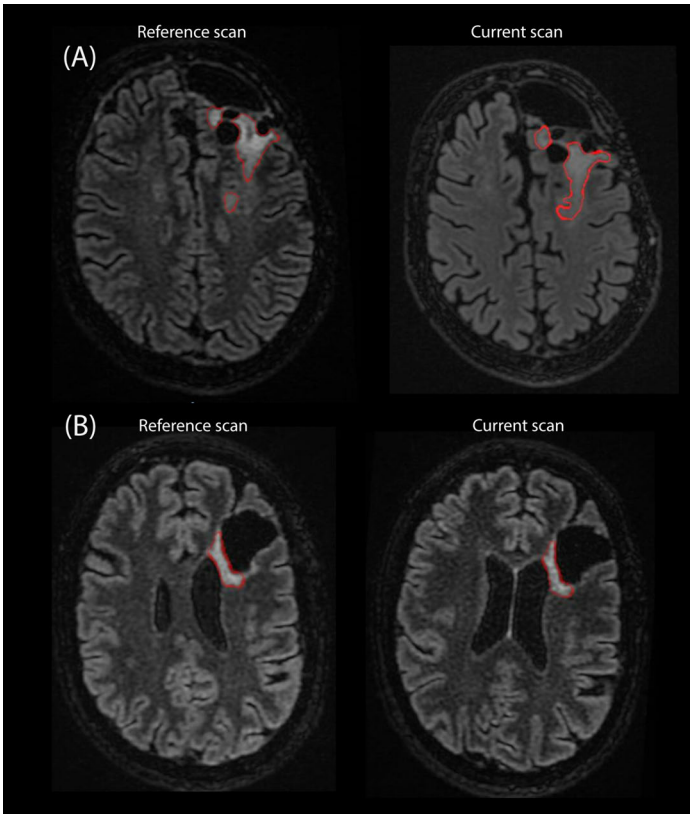


Figure 4.8: Example of segmentations as they are stored in PACS as an overlay on the T2-FLAIR scan from two consecutive timepoints. A) Two consecutive scans of patient where EASE segmentations were considered inconsistent by the radiologist B) Two consecutive scans where a volume-based diagnosis of stable disease could be made.

4.5 Discussion

A clinical segmentation pipeline 'EASE' was implemented to perform automated 3D volume measurements in LGG. As the effect of such a measurement on clinical decision making is still unknown, and perfect performance of the algorithm cannot be expected, several steps were taken to ensure patient safety and monitor results.

The main purpose of this work is to establish the protocols and tools to allow the first introduction of a new, potentially valuable diagnostic tool into clinical practice. From the initial reference dataset, with 7 out of 20 segmentations rejected, it is clear that the quality assessment remains an essential step in the usage of EASE. First results from clinical practice indicate a similar success rate of 74% for individual scans, and approximately half of the patients could be successfully diagnosed with three consecutive volume measurements. However, since the sample size is limited, with almost exclusively diagnoses of stable disease, so further validation of the performance is required to draw firm conclusions on the expected success rate. It must be noted that the segmentation of non-enhancing LGG is particularly difficult due to their diffuse border and varying signal intensity on particularly T2-FLAIR imaging. Furthermore, the underlying deep learning solution, HD-GLIO, was evaluated mostly on high-grade glioma. The current application is therefore aimed at a different, and possibly more challenging patient group and while our results show that a clinical application is feasible, but a more reliable segmentation is needed to facilitate efficient diagnosis.

The results confirm that automatic segmentation of low-grade glioma during follow-up is not a solved problem, and therefore highlight the importance of the quality assurance protocols and manual checks that are presented in this work, and which are ideally part of any introduction of new assessment tools into clinical practice. EASE facilitates a quantitative measurement of lesions that are often impossible to measure accurately even in 2D, due to their irregular shape, and therefore serves a long-standing wish from the neuro-oncological community to move to a potentially more accurate 3D measurement. In this light, a successful diagnosis in over half of the patients is already a valuable step forward.

The initial evaluation in clinical practice provides valuable feedback on the use of automatic segmentation in low-grade glioma. Notably, it shows that an automatic segmentation method is no guarantee for consistent results. Even though the inter-rater variation is removed through automation, the diffuse border of low-grade glioma can still cause ambiguity in the segmentation. Ideally, an automatic segmentation method would be consistent in its choice of where to set the border, but results from EASE show that slight variations in image intensities between consecutive scans can lead to longitudinal inconsistencies. This means that a critical assessment by the radiologist is still needed even if all segmentations are checked and accepted on an individual basis. In EASE, this is ensured by a workflow that can be easily applied in the clinical routine and the protocol for clinical decision-making described in section 4.3.1. Future technical improvements in the automatic segmentation of LGG should focus not only on improving the quality of individual segmentations, but also on longitudinal stability. For this, assessing the reproducibility of the entire process from scan to measurement would be of value, although this would

require repeated measurements within a close enough timeframe to assume no change in tumor volume. Such a set-up is not consistent with clinical practice and would require a dedicated study with funding for additional scanning procedures and full consideration of whether the burden this incurs on patients is justified reproducibility of the entire process from scan to measurement.

This work describes a first and careful implementation of automatic segmentation of LGG in clinical practice. Although the results leave room for improvement for the segmentation method, it is already being applied successfully in approximately half of the patients. In all patients diagnosed thus far, the volume measurements confirm the conventional visual diagnosis, as would be expected, but the volume quantification increases confidence in the diagnosis. Essentially, results show that radiologists are cautious in their use of the measurements. The fact that the segmentations are verified and stored for future reference not only decreases the risk of a false diagnosis, but also increases the confidence of the radiologist when using such deep learning solutions in their clinical practice.

Only four patients were included with a diagnosis of progressive disease (PD), which can be attributed to the fact that the most common sign of PD is the presence of contrast enhancement. This is often accompanied by concurrent volume increase, but these cases were excluded from the study in order to address the diagnostic uncertainty regarding non-enhancing lesions. When comparing the volume change between patients with SD and PD, there is no clear threshold to separate the two categories. Although the RANO guidelines recommend a threshold of 25% change for 2D measurements, which would correspond to a 40% change in volume, the final interpretation is left to the discretion of the radiologist and may depend on other factors, such as baseline volume, the presence or absence of treatment-related white matter abnormalities and the consistency of segmentations longitudinally.

When looking at the 2D RANO measurements there is a clear distinction between SD and PD, even though these measurements do not capture the full extent of the irregular shape and diffuse infiltration of these lesions. From these results it seems that the existing visual diagnosis is still being used as the primary tool to determine tumor growth, but are too few patients showing progression in either method to draw a firm conclusion. Also, it must be noted that these results were gathered in the first months after EASE was released for clinical use.

EASE was put into service prior to the date of application of EU regulation 2017/745 on medical devices (MDR). We are aware that in case of substantial changes in the design or intended purpose of EASE, the requirements of this regulation are applicable. Our approach to ensure quality of results and prevent incorrect interpretation is already in line with the general aim of the MDR.

We think this implementation provides a potential benefit to both the clinicians and researchers, as radiologist receive a valuable tool for the quan-

tification of glioma volume, even if not fully perfected,while researchers receive valuable feedback from clinical practice. In its current form, EASE does not allow for correction of failed segmentations through manual intervention of the radiologist, as this is not feasible in clinical practice. However, the feedback from clinical practice could enable further improvement in the segmentation, whether that is in the preprocessing or by improving the HD-GLIO model in a transfer learning approach, while the clearly defined protocol for software updates ensures patient safety during such future improvements.

II

Emerging biomarkers and methods

5

Longitudinal characteristics of T2-FLAIR mismatch in IDH-mutant Astrocytoma: relation to grade, histopathology and overall survival in the GLASS-NL cohort

Je kunt zeggen: we hebben de luxe niet om even niks te doen. Maar dat is geen luxe, dat is noodzaak als je een probleem wilt oplossen.

– *Lieke Marsman*

Based on: **Karin A. van Garderen***, W. R. Vallentgoed*, A. Lavrova, J. M. Niers, W. W. de Leng, Y. Hoogstrate, I. de Heer, B. Ylstra, E. van Dijk, S. Klein, K. Draaisma, P. A. Robe, R. G. Verhaak, B. A. Westerman, P. J. French, M. J. van den Bent, M. C. Kouwenhoven, J. M. Kros, P. Wesseling, and M. Smits, "Longitudinal characteristics of T2-FLAIR mismatch in IDH-mutant astrocytoma: Relation to grade, histopathology and overall survival in the GLASS-NL cohort," *Neuro-Oncology Advances*, vol. 5, no. 1, vdad149,

* Contributed equally

Abstract

Background: *The T2-FLAIR mismatch sign is defined by signal loss of the T2-weighted hyperintense area with FLAIR (Fluid-Attenuated Inversion Recovery) on MRI, causing a hypointense region on FLAIR. It is a highly specific diagnostic marker for IDH-mutant astrocytoma, and is postulated to be caused by intercellular microcystic change in the tumor tissue. However, not all IDH-mutant astrocytomas show this mismatch sign and some show the phenomenon in only part of the lesion. The aim of the study is to determine whether the T2-FLAIR mismatch phenomenon has any prognostic value beyond initial non-invasive molecular diagnosis.*

Methods: *Patients initially diagnosed with histologically lower grade (2 or 3) IDH-mutant astrocytoma and with at least two surgical resections were included in the GLASS-NL cohort. T2-FLAIR mismatch was determined, and the growth pattern of the recurrent tumor immediately before the second resection was annotated as invasive or expansive. The relation between the T2-FLAIR mismatch sign and tumor grade, microcystic change, overall survival (OS) and other clinical parameters was investigated both at first and second resection.*

Results: *The T2-FLAIR mismatch sign was significantly related to grade 2 (80% vs 51%), longer post-resection median OS (8.3y vs 5.2y), expansive growth and lower age at second resection. At first resection no relation was found between the mismatch sign and OS. Microcystic change was associated with areas of T2-FLAIR mismatch.*

Conclusions: *T2-FLAIR mismatch in IDH-mutant astrocytomas is correlated with microcystic change in the tumor tissue, favorable prognosis and grade 2 tumors at time of second resection.*

5.1 Introduction

GLASS-NL is a multicenter consortium in the Netherlands (NL) and part of the international Glioma Longitudinal AnalySiS (GLASS) initiative [89]. GLASS-NL (Vallentgoed et al., manuscript in preparation) focusses on changes underlying malignant progression in astrocytomas, IDH-mutant (henceforth 'astrocytomas') through the analysis of molecular characteristics of repeated resections and longitudinal magnetic resonance imaging (MRI). A highly specific imaging feature of this type of glioma is the presence of a near-complete mismatch between the signal on T2-weighted (T2w) and T2-weighted Fluid Attenuated Inversion Recovery (FLAIR) MRI, also called the T2-FLAIR mismatch sign. In this study we investigate the longitudinal characteristics of this mismatch phenomenon, and its relation to tumor malignancy grade and patient prognosis.

The signal on FLAIR is similar to T2w within brain tissue, but the free fluid signal on is dark FLAIR whereas on T2w it is hyperintense. A non-enhancing lesion is generally hyperintense on both sequences, but in low-grade glioma it has been noted that areas where T2w shows a distinct high signal, FLAIR signal may sometimes be relatively hypointense. The T2-FLAIR mismatch sign describes the phenomenon where nearly the entire lesion shows this signal intensity mismatch except for a bright outer rim. It was first reported by Patel et al. [90] as an imaging marker for non-invasive diagnosis of astrocytoma, and subsequently validated in multiple studies [91, 92] to be highly specific (99%) for these tumors within the group of adult-type diffuse low-grade gliomas (LGGs). However, the sensitivity of the T2-FLAIR mismatch sign was found to be low, as it is described to be present in approximately half of all astrocytomas at initial diagnosis. A clear relationship was reported between the T2-FLAIR mismatch sign and the presence of microcysts in histological slides of the tumors [90, 93, 94]. Previous studies reported no significant difference in outcome or clinical parameters between cases with and without the mismatch sign [91, 93, 94]. However, these analyses were performed on studies with small sample sizes ($n < 50$) and only MRI information at initial diagnosis was used. To date, no longitudinal study has been performed on the T2-FLAIR mismatch sign.

So far, it remains unclear whether the T2-FLAIR mismatch sign has clinical relevance beyond an MRI-based (i.e., non-invasive) indication of astrocytoma in adults. Astrocytomas that start as lower grade tumors are known to sooner or later undergo malignant transformation, so a non-invasive marker for tumor grade can potentially inform treatment decisions for recurrent disease. The GLASS-NL cohort provides a unique opportunity to relate T2-FLAIR mismatch to tumor grade at progression in a clinically well-defined cohort. Furthermore, the longitudinal nature of the data acquisition allows us to assess the growth pattern of the recurrent tumor and the change in T2-FLAIR mismatch over

time.

The aim of this work is to validate the correlation between the T2-FLAIR mismatch phenomenon and microcystic change in the GLASS-NL-cohort, and to investigate the clinical relevance of the mismatch phenomenon during the course of tumor evolution. Specifically, we aim to investigate whether T2-FLAIR mismatch is related to tumor grade and overall survival (OS), and whether it has added prognostic value when considering the presence of contrast-enhancement and tumor grade. It is also possible that a part of the lesion shows T2-FLAIR mismatch, but the mismatch is not 'near-complete' as is required for the T2-FLAIR mismatch sign. In this study we consider also these cases as a T2-FLAIR mismatch *area*, distinct from the mismatch *sign*, as this distinction may be relevant for the relation with histopathology and clinical parameters.

5.2 Methods

5.2.1 Patient inclusion

The GLASS-NL cohort is a retrospective multi-center cohort from the Netherlands. Patients were included from Amsterdam UMC, Erasmus MC and UMC Utrecht according to the following inclusion criteria:

1. Patient was first diagnosed as an adult (>18 years old);
2. The initial histological diagnosis was lower grade (grade II or III) astrocytoma, IDH-mutant according to the CNS WHO-2016 classification6;
3. Patient underwent surgical resection at least twice, with second surgery performed after progression and with at least 6 months' time difference;
4. Both surgeries yielded tumor tissue sufficient for molecular diagnosis;

   For this study, samples from the GLASS-NL cohort were selected where pre-operative MR imaging was available, meeting the following requirements:
5. The following sequences were available: 1) T2-weighted, 2) T2-weighted FLAIR, 3) T1-weighted and 4) T1-weighted after administration of a gadolinium-based contrast agent (post-contrast T1-weighted).
6. The image quality was sufficient to delineate the lesion.

The latest available pre-operative scan meeting these criteria was used for image analysis and annotation. This study was approved by the ethical review board of Amsterdam UMC (VUMC 2019.085). The study was performed in accordance with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

5.2.2 Clinical characteristics

The location of the lesion and presence in or near eloquent areas according to Sawaya et al. [95] were annotated by a radiologist (A.L.) for each scan, both at first and second resection. Clinical characteristics and treatment history were retrieved from electronic health records. Overall survival (OS) and progression-free survival (PFS) were measured from both the date of first (OS-R1, PFS-R1) and second resection (OS-R2, PFS-R2) till time of death or progression respectively, or censored at last follow-up date. As clinical practice has changed over time, with an initial period of watchful waiting being more common in earlier samples, the time of resection was preferred over the time of diagnosis as a reference date for survival analysis. Tumor grade was re-evaluated according to the WHO-2021 classification [3].

5.2.3 Volume measurement and annotation of enhancement

The tumor was delineated automatically using HD-GLIO [17, 19] and corrected semi-automatically using ITK-Snap [96]. The resulting two-class segmentation was used to compute the contrast-enhancing volume (CET) and whole tumor (WT) volume (T2-weighted abnormalities + CET). If parts of the abnormalities on T2-weighted imaging were clearly attributable to treatment effects, they were excluded from the volume. The presence and thickness of the contrast-enhancing margin was annotated by a radiologist (A.L.) according to the VASARI features [97]. If there was a disconcordance between the annotation of contrast enhancement and the presence of contrast-enhancing volume in the segmentation in recurrent lesions, which may be a result of post-surgery changes, a board-certified and expert-neuroradiologist (M.S.) was consulted to decide whether the lesion was enhancing or non-enhancing.

5.2.4 T2-FLAIR mismatch annotation

A distinction was made between the presence of a T2-FLAIR mismatch area and the T2-FLAIR mismatch sign. A lesion was annotated as having a mismatch area if an area was found with the following characteristics:

1. The T2w sequence is homogeneously hyperintense in this area;
2. The FLAIR sequence is clearly hypointense in this area in comparison to FLAIR-hyperintense areas elsewhere in the same lesion (e.g., a hyperintense rim);
3. The T2-FLAIR mismatched area is not contrast-enhancing, necrotic or a cyst, although these aspects may be present near or within the mismatched area. In order to be classified as having the mismatch sign, the following criteria were used:

4. In initial lesions: almost the entire lesion shows T2-FLAIR mismatch area, except for a thin hyperintense rim;

5. In recurrent lesions: the lesion identified as recurrent tumor almost entirely shows T2-FLAIR mismatch, with presence of a hyperintense rim at the interface of recurrent lesion and healthy-appearing brain.

This is according to the recommendations by Jain et al. [98] for initial lesions. However, the criteria for recurrent lesions were adapted as preexisting and potentially treatment-induced abnormalities can be hyperintense on the T2-weighted FLAIR scan and thereby result in an incomplete T2-FLAIR mismatch phenomenon. Both the T2-FLAIR mismatch area and the T2-FLAIR mismatch sign were annotated for all included pre-operative scans as (YES/NO). Note that a mismatch area is a prerequisite for the mismatch sign, so all cases with a mismatch sign necessarily also show a mismatch area. If the first rater (K.A.v.G.) was not sure whether a T2-FLAIR mismatch (sign or area) was present, a board-certified and expert-neuroradiologist (M.S.) was consulted to reach a decision in consensus.

5.2.5 Growth pattern annotation

The growth pattern was annotated for the recurrent lesions, by the same rater as the T2-FLAIR mismatch (K.A.v.G.), by comparing the pre-operative scan to a prior reference scan, selected to show the most recent visible growth of the tumor. The following categories were used:

1. Mostly invasive: the recurrent lesion mostly infiltrates formerly healthy appearing tissue;

2. Mostly expansive: the recurrent lesion barely seems to invade surrounding preexistent brain tissue, but rather shows expansive growth, thereby variably displacing/compressing the surrounding brain tissue;

3. Mixed: both patterns of growth are present and neither is clearly dominant;

4. Not sure: the growth pattern cannot be distinguished, e.g. because there is not enough growth visible or the scan quality is insufficient.

5. Not available: there is no imaging available.

Only the category ‘Not available’ was considered missing values and excluded from the statistical analysis.

5.3 Histopathology

Histological slides of the tumors were stained with hematoxylin-eosin (H&E) and digitized using a whole slide scanner (Hamamatsu NanoZoomer 2.0 HT). The presence/absence of microcysts was assessed on the section originally used for histopathological diagnosis in all samples where sufficient pre-operative imaging was available for the assessment of T2-FLAIR mismatch sign, by a board-certified, expert-neuropathologist (J.M.K.) who was blinded to MRI data. If only a small part of the samples showed microcysts, or if the microcystic change was incipient but visible, it was still annotated as present. If the scanned samples did not have sufficient quality to identify microcysts, the sample was annotated as NOT SURE and excluded from the analysis.

5.3.1 Statistical analysis

Statistical analysis was performed using python (v3.8.13) and the packages SciPy (v1.8.0) and lifelines (v0.27.4). The group of tumors with T2-FLAIR mismatch sign and those with T2-FLAIR mismatch area, the latter by definition also including tumors with the T2-FLAIR mismatch sign, were compared to the group of tumors without T2-FLAIR mismatch sign/area. The analysis was performed at first and second resection, including for each resection all patients with available imaging data at that time. Additionally, the clinical characteristics of the entire cohort (including those without imaging data) were reported at time of first and second resection. Fisher's exact test was used to test differences in categorical variables. The difference in continuous variables (volume and age) between groups was tested using the Mann-Whitney-U test. A threshold of $p<0.05$ was used for significance. In case of missing values in the clinical parameters, the number of patients missing the parameter was reported and they were excluded from statistical testing for that specific parameter. The first and second resection were analyzed separately as availability of imaging was different at both time-points.

Survival analysis was performed by visualizing the Kaplan-Meier estimates for groups and comparing these using the log-rank test. The groups were defined based on the presence of T2-FLAIR mismatch area and sign, tumor grade according to the WHO-2021 classification [3] and the presence of microcysts. As the analysis at second resection may include cases with clear malignant transformation, a stratified analysis of clinical characteristics and OS was performed in the subset of non-enhancing samples at second resection. Additionally, a stratified analysis of OS was performed in cases of CNS WHO-2021 grade 2.

To assess whether the longitudinal changes in T2-FLAIR mismatch affected survival, another comparison was made for OS-R2 for T2-FLAIR mismatch area and sign in longitudinal categories, creating four groups for each: patients

where the mismatch sign/area was present in both first and second resection ('preserved'), those where it was present at second resection but not at first ('gained'), those where it was present at first resection but not at second ('lost') and those where it was never present ('never').

5.4 Results

5.4.1 Patient inclusion and characteristics

A total of 101 patients were included in GLASS-NL, with a total of 224 tissue samples. Although there were tissue samples of at at least two resections for each patient, for some patients the tissue of the first or second resection was missing. A total of 98 samples were included at first resection and 97 at second resection. For 32 and 4 patients, a sample could be included from a third and fourth resection, respectively. Figure 5.4.1 shows examples of the T2-FLAIR mismatch sign and area at second resection, including the corresponding H&E slides and growth pattern.

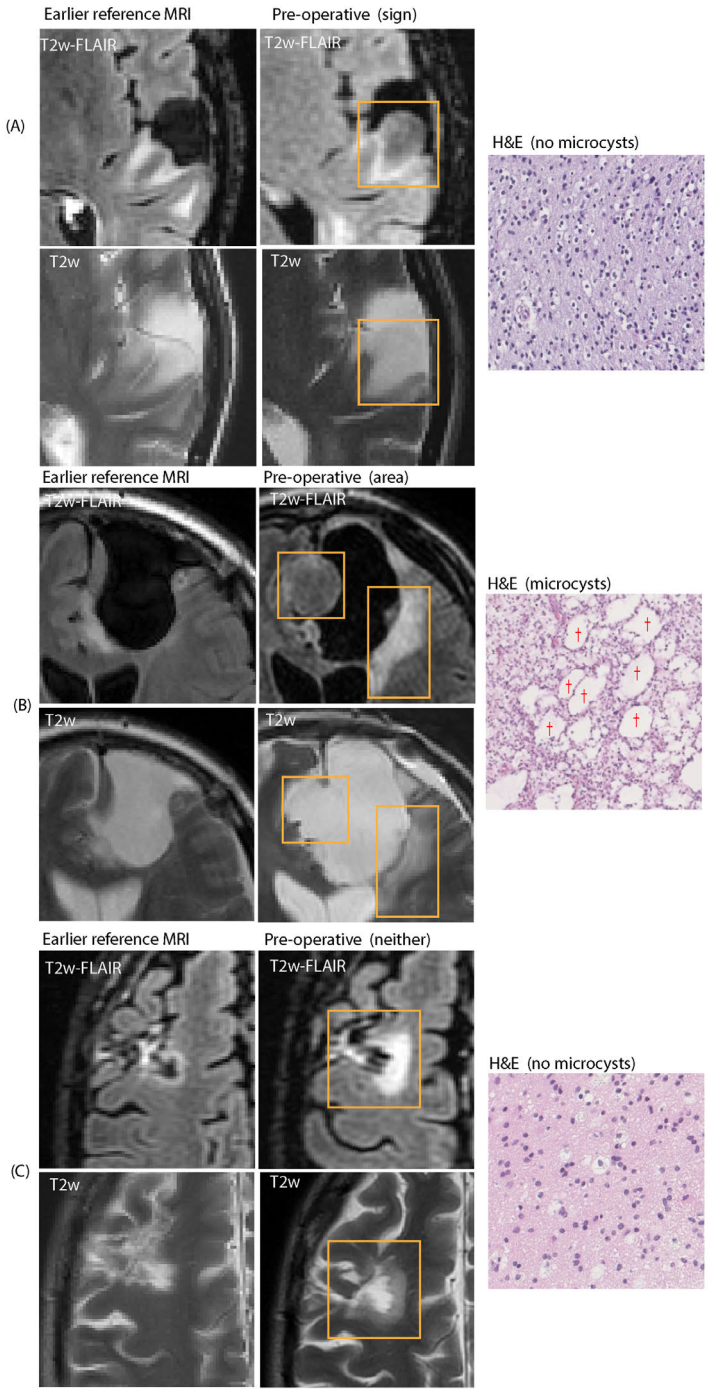
Clinical characteristics at first resection

At first resection, 45 samples were included. A complete overview of of patient characteristics at first resection, stratified by the presence/absence of the T2-FLAIR mismatch sign is included in Supplementary Table 5.S1. Supplementary Table 5.S2 contains the same characteristics stratified by T2-FLAIR mismatch area. The mean age at diagnosis was 32 years, median OS was 9.6 years (95% CI: 8.8 – 10.7) and median PFS was 2.8 years (95% CI: 2.6 – 3.3). The T2-FLAIR mismatch sign was found in 13 patients (29%) and a T2-FLAIR mismatch area was found in an additional 26 patients (58%). The remaining 6 patients (13%) showed no T2-FLAIR mismatch at all. There was no significant difference between patients with and without the T2-FLAIR mismatch sign for any of the clinical parameters at first resection. Patients with a T2-FLAIR mismatch area received chemotherapy less often than those without either a T2-FLAIR mismatch area or sign (1/39 vs 3/6 patients, 3% vs 50%) and were less often diagnosed with grade 4 (1/39 vs 3/6 patients, 3% vs 50%).

Clinical characteristics at second resection and beyond

At second resection, 76 samples could be included. Table 5.1 contains an overview of patient characteristics at second resection, stratified by the presence/absence of the T2-FLAIR mismatch sign. Supplementary Table 5.S3 contains the same overview stratified by T2-FLAIR mismatch area. Median OS-R2 was 5.8 years (95% CI: 4.2 – 7.4) and median PFS-R2 was 2.5 years (95% CI: 1.6 – 3.5).

The T2-FLAIR mismatch sign was found in 25 patients (33%), and an additional 11 patients (14%) showed a T2-FLAIR mismatch area but not the sign. The remaining 40 patients (53%) did not show any T2-FLAIR mismatch. Both the presence of the T2-FLAIR mismatch area and sign were related to a lower mean age at initial diagnosis and a lower probability of the lesion being in an eloquent area. The T2-FLAIR mismatch area and sign were related to a lower contrast-enhancing volume and higher probability of being non-enhancing and of an expansive growth pattern. The median whole tumor volume was significantly lower in patients with the T2-FLAIR mismatch sign (12.1 vs. 22.5 mL, p=0.01), but not in patients with the T2-FLAIR mismatch area (16.9 vs 20.4 mL, p=0.29).



(Caption on next page.)

Figure 5.1: Examples of FLAIR MRI and H&E slides for three cases. The last pre-operative MRI is shown with an earlier reference MRI used to assess the growth pattern. Recurrent lesions are outlined with a rectangle. Examples of microcysts are indicated by a '†' symbol. A) Recurrent tumor with mismatch sign; pre-operative MRI shows pre-existing treatment effect, but recurrent growth is mismatched with hyperintense rim; growth pattern is mostly expansive; H&E does not show microcysts. B) Recurrent tumor that is partly mismatched, so this case shows a T2-FLAIR mismatch area but no mismatch sign; growth pattern is mixed; H&E shows large microcysts. C) Recurrent tumor without T2-FLAIR mismatch; growth pattern is mostly invasive; H&E shows no microcysts.

Table 5.1: Patient characteristics and annotation results at second resection, stratified by the presence of T2-FLAIR mismatch sign. First column (all) includes patients where no sufficient imaging was available. P-value compares columns absent and present. Missing values are reported as 'Not available', but not included in the computation of percentages and p-values.

T2-FLAIR mismatch sign	**Absent (n=51)**	**Present (n=25)**	**p-value**
Female sex (%)	17 (33%)	15 (60%)	0.05
Age at diagnosis (y) median (range)	34.0 (18.0 - 70.0)	28.0 (19.0 - 53.0)	**0.03**
Median time since diagnosis in y (range)	4.5 (1.2 - 23.5)	3.9 (1.0 - 14.9)	0.3
Median overall post-resection survival (OS-R2) in y (95% CI)	5.2 (2.8 - 5.9)	8.3 (6.4 – N/A)	**0.001**
Median progression-free post-resection survival (PFS-R2) in y (95% CI)	1.6 (1.3 - 2.8)	4.1 (2.5 – N/A)	**0.01**
Median time to second resection in y (range)	3.6 (1.0 - 17.5)	3.6 (0.9 - 13.9)	0.8
CNS WHO-2021 grade			
- Grade 2	26 (51%)	20 (80%)	**0.02**
- Grade 3	7 (14%)	2 (8%)	0.71
- Grade 4	18 (35%)	3 (12%)	0.05
KPS before surgery			
- 100	21 (41%)	14 (56%)	0.33
- 90	19 (37%)	8 (32%)	0.8
- <90	11 (22%)	3 (12%)	0.37
LOCATION			
Side of lesion center			
- Left	24 (47%)	13 (52%)	0.81
- Right	27 (53%)	12 (48%)	0.81
Location in or near eloquent regions (Sawaya et al.)			
- Eloquent (III)	36 (71%)	9 (36%)	**0.006**
- Near-eloquent (II)	8 (16%)	7 (28%)	0.23
- Non-eloquent (I)	7 (14%)	9 (36%)	**0.04**
Tumor site (multiple sites possible)			
- Frontal lobe	39 (76%)	22 (88%)	0.36
- Temporal lobe	23 (45%)	6 (24%)	0.09
- Insula	21 (41%)	5 (20%)	0.08
- Corpus callosum	7 (14%)	0 (0%)	0.09
- Parietal lobe	15 (29%)	3 (12%)	0.15
- Occipital lobe	3 (6%)	0 (0%)	0.55
- Brainstem	1 (2%)	0 (0%)	1
- Basal ganglia	1 (2%)	0 (0%)	1
- Thalamus	1 (2%)	0 (0%)	1

Table 5.1: (cont.)

T2-FLAIR mismatch sign	**Absent (n=51)**	**Present (n=25)**	**p-value**
TREATMENT			
Extent of resection			
- Partial resection	36 (71%)	13 (52%)	0.13
- Complete resection	15 (29%)	12 (48%)	0.13
Radiotherapy	15 (29%)	12 (48%)	0.13
Chemotherapy	18 (35%)	8 (32%)	1.00
PRIOR TREATMENT			
Radiotherapy	21 (41%)	5 (20%)	0.08
Number of radiotherapy treatments			
- 1	20 (39%)	5 (20%)	0.12
- 2	1 (2%)	0 (0%)	1.00
Chemotherapy	11 (22%)	1 (4%)	0.09
Number of chemotherapy treatments			
- 1	7 (14%)	1 (4%)	0.26
- 2	2 (4%)	0 (0%)	1.00
Biopsy	6 (12%)	1 (4%)	0.41
RADIOLOGICAL FEATURES			
Median contrast-enhancing volume * in mL (range)	0.1 (0.0 - 63.8)	0.0 (0.0 - 0.3)	**<0.001**
Median whole tumor volume in mL (range)	22.5 (1.7 - 149.3)	12.1 (1.3 - 29.0)	**0.01**
Thickness of enhancing margin			
- Not Applicable	25 (49%)	20 (80%)	**0.01**
- Thin (<3mm)	11 (22%)	4 (16%)	0.76
- Thick/Nodular (=>3mm)	10 (20%)	1 (4%)	0.09
- Solid	5 (10%)	0 (0%)	0.16
Growth pattern			
- Mostly invasive	13 (27%)	4 (18%)	0.55
- Mostly expansive	2 (4%)	5 (23%)	**0.03**
- Not sure	18 (38%)	7 (32%)	0.79
- Mixed	15 (31%)	6 (27%)	0.79
- Not available	3	3	1
Mismatch sign at first resection			
- Yes	4 (14%)	7 (50%)	**0.02**
- No	24 (86%)	7 (50%)	**0.02**
- Not available	23	11	1

Longitudinal T2-FLAIR characteristics

The prevalence of T2-FLAIR mismatch area decreased with second resection (87% vs. 47% without mismatch, p<0.001), but the prevalence of the T2-FLAIR mismatch sign did not change significantly between first and second resection (29% vs. 33%, p=0.69). There was a significant positive relation between the presence of T2-FLAIR mismatch sign at first and second resection (p=0.02), although 4 patients lost the T2-FLAIR mismatch sign and 7 patients gained the T2-FLAIR mismatch sign at second resection. At third resection, 20 samples could be matched with a pre-operative scan and annotated, of which one showed T2-FLAIR mismatch sign. At fourth resection, with 2 matching scans available, none of the scans showed the T2-FLAIR mismatch phenomenon anymore. See Sankey diagram in Figure 5.2 for visualization of the number of cases showing T2-FLAIR mismatch sign and/or area in the sequential resection specimens.

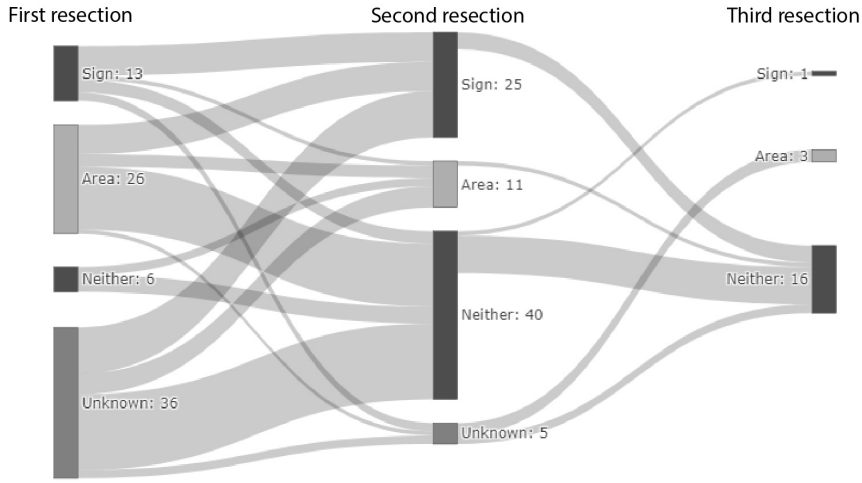


Figure 5.2: Sankey diagram of T2-FLAIR mismatch sign and area over repeated resections (left: first, middle: second, right: third). Area indicates that there was an area of T2-FLAIR mismatch, but the lesion did not meet the criteria for the mismatch sign. Resections of the same patients are connected by grey bands.

5.4.2 Survival analysis

Figure 5.3 shows the Kaplan-Meier OS curves for presence/absence of both the T2-FLAIR mismatch area and sign at first (OS-R1) and second resection (OS-R2). There was no significant difference in OS-R1 in patients with or without a T2-FLAIR mismatch area (p=0.19) or sign (p=0.07). The tumor grade at first resection, according to WHO-2021 guidelines, was also not a

significant prognostic factor in this cohort (p=0.12 CNS WHO grade 2 vs. 3 or 4, p=0.73 grade 2 or 3 vs. 4), see Supplementary Figure 5.S1. The median PFS-R1 was significantly higher in patients with than in patients without the T2-FLAIR mismatch sign (3.5 vs. 2.2 years, p=0.002). When looking at the T2-FLAIR mismatch area, the difference in PFS-R1 was not significant (2.8 vs. 2.8 years, p=0.51).

At second resection, both the presence of T2-FLAIR mismatch area and sign were related to longer OS-R2 (p=0.001 sign, p=0.009 area) (Figure 5.3) and PFS-R2 (p=0.01, p=0.07 area). Tumor grade in general was also a prognostic factor at second resection (p<0.001 grade 2 vs. 3 or 4, p<0.001 grade 2 or 3 vs. 4). When considering only cases of CNS WHO grade 2, there was no significant difference in OS-R2 between patients with or without the T2-FLAIR mismatch sign at second resection (p=0.21). When considering only non-enhancing lesions, the presence of the T2-FLAIR mismatch sign was still a strongly significant prognostic factor (p=0.009). Figure 5.4 contains the Kaplan-Meier curves for these stratified analyses. No significant difference in OS was found between patients with and without microcysts at first resection (OS-R1) (p=0.12) or second resection (OS-R2) (p=0.11), see Supplementary Figure 5.S2.

For the longitudinal categories of T2-FLAIR mismatch sign/area, there was a significant difference in OS-R2 for patients where the T2-FLAIR mismatch area was preserved versus gained (p=0.004). For the T2-FLAIR mismatch sign there was no significant difference between preserved or gained. Figure 5.S3 contains the Kaplan-Meier curves for all four categories.

5.4.3 Analysis in non-enhancing recurrent lesions

When considering only the non-enhancing lesions at second resection 20 out of 43 patients (47%) showed the T2-FLAIR mismatch sign. The group with the T2-FLAIR mismatch sign had a longer OS-R2 (p=0.009) after resection, but there was no significant difference in PFS-R2 (p=0.13), tumor grade (p=0.18), age at diagnosis (p=0.90) or tumor volume (p=0.53). The complete overview of clinical characteristics for non-enhancing lesions at second resection can be found in Table **??**. When we compared the lack of contrast enhancement to the T2-FLAIR mismatch sign as a marker for grade 2, we found that the positive predictive value (PPV) was higher (80%) for the T2-FLAIR mismatch sign than the absence of contrast enhancement (74%), although the sensitivity was lower (43% vs. 70%). When we combine the two markers, so considering lesions that are non-enhancing and show the T2-FLAIR mismatch sign, the PPV for grade 2 increases to 85%, while the sensitivity is 39%. The confusion matrices for these three imaging markers are shown in Supplementary Table 5.S4.

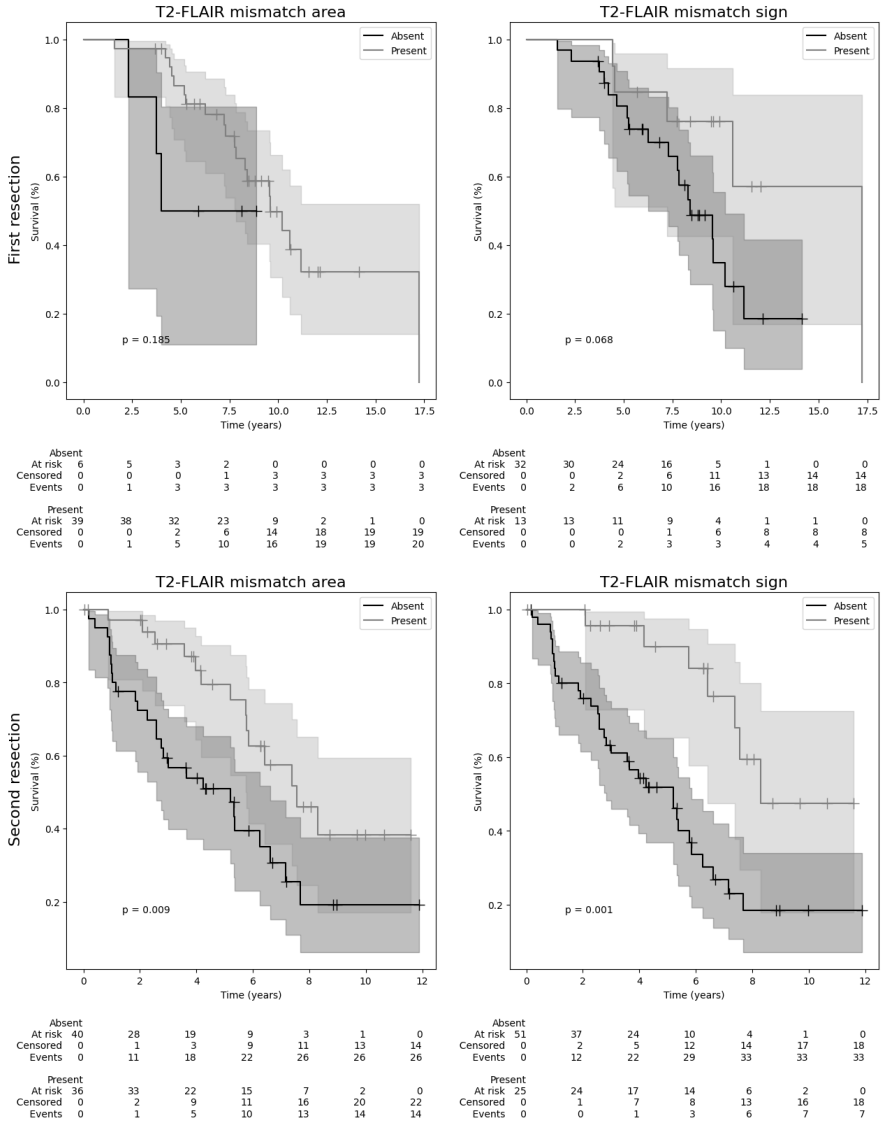


Figure 5.3: Kaplan Meier curves of mismatch area (left) and mismatch sign (right) at first (top) and second (bottom) resection. Starting date for the analysis is the date of first and second resection respectively, and T2-FLAIR mismatch area / sign was annotated on the last available MRI before resection. Censored patients indicated by a '+' at date of last follow-up. Shaded areas indicate 95% confidence intervals.

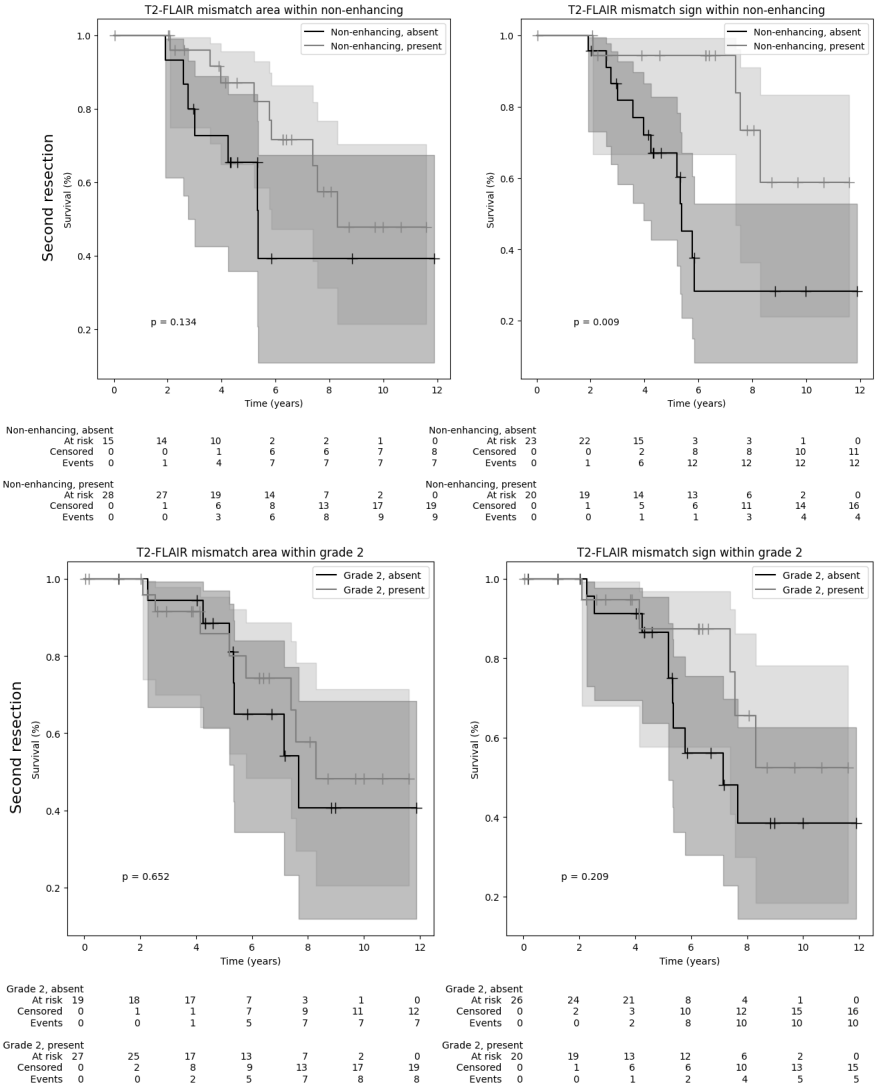


Figure 5.4: Kaplan Meier curves of mismatch area (left) and sign (right) combined with other indicators of good prognosis at second resection. Top: Non-enhancing lesions. Bottom: WHO-2021 grade 2. Starting date of the analysis is date of second resection. T2-FLAIR mismatch area / sign and enhancement were annotated on the last available MRI before resection. Censored patients indicated by a '+' at date of last follow-up. Shaded areas indicate 95% confidence intervals.

Table 5.2: Patient characteristics and annotation results at second resection for non-enhancing lesions only, stratified according to the presence of T2-FLAIR mismatch sign. Missing values are reported as 'Not available', but not included in the computation of percentages and p-values.

T2-FLAIR mismatch sign	**Absent (n=23)**	**Present (n=20)**	**p-value**
Female sex (%)	9 (39%)	12 (60%)	0.23
Median age at diagnosis in y (range)	27.0 (18.0 - 54.0)	28.0 (19.0 - 41.0)	0.90
Median time since diagnosis in y (range)	3.6 (1.2 - 18.7)	3.8 (1.3 - 14.9)	0.80
Median overall post-resection survival (OS-R2) in y (95% CI)	5.4 (4.0 – N/A)	N/A (7.4 – N/A)	**0.009**
Median progression-free post-resection survival (PFS-R2) in y (95% CI)	2.7 (1.4 - 7.7)	5.2 (1.8 – N/A)	0.13
CNS WHO-2021 grade			
- Grade 2	15 (65%)	17 (85%)	0.18
- Grade 3	4 (17%)	2 (10%)	0.67
- Grade 4	4 (17%)	1 (5%)	0.35
KPS before resection			
- 100	14 (61%)	11 (55%)	0.76
- 90	8 (35%)	6 (30%)	1.00
- <90	1 (4%)	3 (15%)	0.32
LOCATION			
Side of lesion center			
- Left	13 (57%)	12 (60%)	1.00
- Right	10 (43%)	8 (40%)	1.00
- Center	0 (0%)	0 (0%)	1.00
Location in or near eloquent regions (Sawaya et al.)			
- Eloquent (III)	14 (61%)	7 (35%)	0.13
- Near-eloquent (II)	5 (22%)	6 (30%)	0.73
- Non-eloquent (I)	4 (17%)	7 (35%)	0.29
Tumor site (multiple sites possible)			
- Frontal lobe	16 (70%)	18 (90%)	0.14
- Temporal lobe	11 (48%)	5 (25%)	0.21
- Insula	7 (30%)	4 (20%)	0.50
- Corpus callosum	5 (22%)	0 (0%)	0.05
- Parietal lobe	4 (17%)	1 (5%)	0.35
- Occipital lobe	1 (4%)	0 (0%)	1.00
TREATMENT			
Extent of resection			
- Partial resection	18 (78%)	10 (50%)	0.06
- Complete resection	5 (22%)	10 (50%)	0.06
Radiotherapy	9 (39%)	9 (45%)	0.76
Chemotherapy	5 (22%)	6 (30%)	0.73
PRIOR TREATMENT			
Radiotherapy	3 (13%)	2 (10%)	1.00

Table 5.2: (cont.) * The growth pattern could not be assessed if no reference scan was available prior to progression (see section 5.2.5).

T2-FLAIR mismatch sign	**Absent (n=23)**	**Present (n=20)**	**p-value**
Chemotherapy	0 (0%)	0 (0%)	1.00
Biopsy	1 (4%)	1 (5%)	1.00
RADIOLOGICAL FEATURES			
Median contrast-enhancing volume in mL (range)	0.0 (0.0 - 0.1)	0.0 (0.0 - 0.1)	0.79
Median whole tumor volume in mL (range)	11.5 (1.7 - 73.1)	11.4 (1.3 - 29.0)	0.53
Thickness of enhancing margin			
- Not Applicable	21 (91%)	18 (90%)	1.00
- Thin (<3mm)	2 (9%)	2 (10%)	1.00
Growth pattern			
- Mostly invasive	7 (32%)	3 (17%)	0.46
- Mostly expansive	0 (0%)	4 (22%)	**0.03**
- Mixed	8 (36%)	5 (28%)	0.74
- Not sure	7 (32%)	6 (33%)	1.00
- Not available*	1	2	0.59

5.4.4 Microcysts

When combining all samples, including third and fourth resections, 88 out of 137 samples (64%) contained microcysts and the relation with T2-FLAIR mismatch area was significant (p=0.03). However, there was no significant relation with the T2-FLAIR mismatch sign (p=0.16).

At first resection, 42 samples were evaluated of which 23 (55%) were found to have microcysts. Two samples had to be excluded due to insufficient quality. In the samples with microcysts, all MRI scans also showed a T2-FLAIR mismatch area. Only six (26%) of the samples with microcysts had the T2-FLAIR mismatch sign on MRI. In the samples without microcysts, 13 MRI scans (68%) still showed a T2-FLAIR mismatch area and five (26%) showed the T2-FLAIR mismatch sign. There was a significant correlation between the presence of microcysts and T2-FLAIR mismatch area (p=0.005), but there was no significant relation with the T2-FLAIR mismatch sign (p=0.99). At second resection, all 76 samples could be annotated and 52 (68%) contained microcysts. In the samples with microcysts, 29 (56%) also showed a T2-FLAIR mismatch area and 21 (40%) also showed the T2-FLAIR mismatch sign on MRI. There was a significant difference in the presence of microcysts for both the T2-FLAIR mismatch area (p=0.03) and sign (p=0.04). Supplementary Table 5.S5 shows the confusion matrix for microcysts and T2-FLAIR mismatch sign/area at first and second resection.

5.5 Discussion

In this study, the presence of T2-FLAIR mismatch area and sign was analyzed in the GLASS-NL cohort, a longitudinal study of astrocytomas, IDH-mutant. In general, we find that the T2-FLAIR mismatch sign is related to a higher probability of grade 2 and better prognosis at recurrence. Considering the risk of malignant progression in this patient group, the presence of the T2-FLAIR mismatch sign is a potential clinically relevant marker for low grade recurrence that is highly specific but not sensitive. Comparing to the presence/absence of contrast enhancement, which is a well established indicator of malignant progression, in astrocytoma the T2-FLAIR mismatch sign can be considered an additional strong indicator of low grade recurrence and good prognosis.

The interpretation of OS in this cohort is difficult, as it is prone to survivorship bias and confounding factors. Patients included in this cohort survived at least up to six months after the first resection and were eligible for a second resection as per the inclusion criteria, indicating a relatively good condition and a tumor at location accessible for surgery. This may also contribute to our finding that tumor grade at first resection was not a prognostic marker for OS in this cohort. Although OS is generally measured from date of diagnosis, the time between diagnosis and first or second resection may vary due to changing treatment standards. Therefore, the dates of resection were used as a starting point for survival analysis. Considering that the patients became eligible for inclusion in the GLASS-NL study only at time of the second resection, we took this date as a starting point for the analysis as well. However, in interpreting the results at second resection we must be careful to consider the potential confounding effect of treatment decisions. It is possible that a longer survival is caused by a difference in intervention rather than overall prognosis, especially when considering radiological parameters. These lesions more often showed an expansive growth pattern and tended to appear more well-delineated, which could make them more likely to be considered for resection and amenable to gross total resection. However, there was no significant difference in the time between diagnosis and first or second resection for patients with or without the T2-FLAIR mismatch sign.

The T2-FLAIR mismatch sign at second resection was also correlated to age, location, whole tumor volume and contrast-enhancing tumor volume, which are potential confounding variables for OS. However, this analysis includes enhancing lesions that show clear malignant progression (such as contrast enhancement). When considering only non-enhancing lesions, the T2-FLAIR mismatch sign was still strongly correlated with OS while the correlations with age, volume and location were no longer found. Although the GLASS-NL cohort is unique in its availability of tissue of at least two time points in the disease process, it is not suitable to draw firm conclusions about prognosis. It is possible that the findings at recurrence also extend to the initial presentation,

but the homogeneity of the cohort and low availability of MRI at first resection in our study likely cause insufficient power to distinguish any correlation between the presence of the T2-FLAIR mismatch sign and tumor grade or OS. When considering the T2-FLAIR mismatch sign and area at both resections combined, we found evidence of an improved prognosis for those who showed a mismatch area already at first resection, versus those who gained it at second resection. For future research, it would be worthwhile to study the relation between T2-FLAIR mismatch at initial presentation and recurrence in a large cohort of astrocytomas without selection for treatment or initial histological grade.

The distribution of T2-FLAIR mismatch area and sign was different at first and second resection. At the first resection, the presence of a T2-FLAIR mismatch area (but no sign) was common, while the absence of any T2-FLAIR mismatch was rare. At the time of second resection, the absence of a T2-FLAIR mismatch area was more common, but a T2-FLAIR mismatch area (without sign) was rare and the T2-FLAIR mismatch sign was found at approximately the same rate as in the first resection. This is an indication that the criteria for the T2-FLAIR mismatch sign as further refined by Jain et al. [98] was strictly interpreted at initial diagnosis, while consideration of treatment-related changes caused the T2-FLAIR mismatch area (without sign) to be annotated less often. In general, the potential ambiguity in the definition of the T2-FLAIR mismatch sign at recurrence is a limitation for its use in clinical practice and interpretation in research. Recent studies also show that the detection of the mismatch phenomenon may be highly dependent on acquisition parameters of the FLAIR scan [99], which could cause cases of T2-FLAIR mismatch sign or area to remain undetected.

From this study it is unclear how the cases with a T2-FLAIR mismatch area, but without meeting all criteria for the T2-FLAIR mismatch sign at recurrence should be interpreted. In future research it would be worthwhile to measure the mismatch sign as a continuous variable and investigate the percentage of T2-FLAIR mismatch in the lesion as a prognostic marker. However, this would not be feasible for clinical practice without a robust automated volume measurement. In general, the T2-FLAIR mismatch sign shows a stronger relation to prognosis and grade than a T2-FLAIR mismatch area, suggesting that a strict interpretation of the marker and adherence to the criteria should be preferred while taking into account potential treatment effects in recurrent lesions.

The presence of microcysts as observed by histopathological analysis was significantly correlated with the presence of a T2-FLAIR mismatch area, which supports the hypothesis that the mismatch phenomenon is a direct result of microcystic change. A limitation of this analysis is that the exact location of acquisition of the histopathology slides was unknown, making it impossible to exactly correlate the T2-FLAIR mismatch area with microcystic

change. Previous studies have shown a correlation between the T2-FLAIR mismatch sign and microcystic change [90, 93, 94], but did not include recurrent samples or a distinction between T2-FLAIR mismatch area sign. Small areas of T2-FLAIR mismatch that do not meet the criteria for the T2-FLAIR mismatch sign are not specific to astrocytoma, and neither is microcystic change, so this correlation would likely be found in other low-grade gliomas such as oligodendroglioma, IDH-mutant and 1p19q-codeleted. The longitudinal analysis shows a correlation between the T2-FLAIR mismatch sign at first and second resection, suggesting that there could be a distinct property that causes the T2-FLAIR mismatch sign in a subgroup of astrocytomas. Further research would be needed to find an underlying cause or property that explains the presence of the T2-FLAIR mismatch sign.

To conclude, in the GLASS-NL cohort we found that the T2-FLAIR mismatch sign is related to a low grade and better prognosis at recurrence, and that it has prognostic relevance in addition to the absence of contrast enhancement. However, it is possible that the treatment regimen affected the results in this retrospective study, as the appearance of lesions with the T2-FLAIR mismatch sign may have affected the diagnosis of progression and the decision to undergo a second resection. Due to the design of the cohort, we can not draw firm conclusions concerning the prognostic value of the T2-FLAIR mismatch sign at initial presentation. We conclude that the T2-FLAIR mismatch sign is a potential additional marker for favorable prognosis in recurrent astrocytoma, IDH-mutant that should be investigated further.

5.6 Supplementary material

Table 5.S1: Patient characteristics and annotation results at first resection, stratified according to the presence of T2-FLAIR mismatch sign. First column (all) includes patients where no sufficient imaging was available. P-value compares columns absent and present. Missing values are reported as 'Not available', but not included in the computation of percentages and p-values.

T2-FLAIR mismatch sign	**Absent (n=32)**	**Present (n=13)**	**p-value**
Female sex (%)	14 (44%)	8 (62%)	0.34
Age at diagnosis in (y) median (range)	27.5 (19.0 - 70.0)	27.0 (18.0 - 45.0)	0.6
Median time since diagnosis in y (range)	0.4 (0.0 - 17.0)	0.3 (0.0 - 1.3)	0.39
Median overall post-resection survival (OS-R1) in y (95% CI)	8.4 (7.3 - 10.2)	17.2 (7.2 - 17.2)	0.07
Median progression-free post-resection survival (PFS-R1) in y (95% CI)	2.2 (1.3 - 3.0)	3.5 (2.8 - 6.2)	**0.002**
Median time to second resection in y (range)	2.9 (1.0 - 8.7)	4.2 (0.9 – 8.6)	**0.04**
CNS WHO-2021 grade			
- Grade 2	20 (62%)	11 (85%)	0.18
- Grade 3	9 (28%)	1 (8%)	0.24
- Grade 4	3 (9%)	1 (8%)	1
KPS before surgery			
- 100	13 (41%)	4 (31%)	0.74
- 90	16 (50%)	8 (62%)	0.53
- <90	2 (6%)	1 (8%)	1
- Not available	1	0	1
LOCATION			
Side of lesion center			
- Left	15 (47%)	7 (54%)	0.75
- Right	17 (53%)	6 (46%)	0.75
Location in or near eloquent regions (Sawaya et al.)			
- Eloquent (III)	22 (69%)	8 (62%)	0.73
- Non-eloquent (I)	5 (16%)	4 (31%)	0.41
- Near-eloquent (II)	5 (16%)	1 (8%)	0.66

Table 5.S1: (cont.)

T2-FLAIR mismatch sign	**Absent (n=32)**	**Present (n=13)**	**p-value**
Tumor site (multiple sites possible)			
- Frontal lobe	22 (69%)	11 (85%)	0.46
- Temporal lobe	11 (34%)	4 (31%)	1
- Insula	8 (25%)	6 (46%)	0.29
- Corpus callosum	1 (3%)	1 (8%)	0.5
- Parietal lobe	10 (31%)	2 (15%)	0.46
- Occipital lobe	3 (9%)	0 (0%)	0.55
- Cerebellum	0 (0%)	0 (0%)	1
- Basal ganglia	3 (9%)	1 (8%)	1
- Thalamus	1 (3%)	0 (0%)	1
TREATMENT			
Extent of resection			
- Partial resection	24 (75%)	9 (69%)	0.72
- Complete resection	8 (25%)	4 (31%)	0.72
Radiotherapy	6 (19%)	3 (23%)	0.70
Chemotherapy	4 (12%)	0 (0%)	0.31
PRIOR TREATMENT			
Radiotherapy	1 (3%)	0 (0%)	1
Chemotherapy	1 (3%)	0 (0%)	1
Biopsy	1 (3%)	0 (0%)	1
RADIOLOGICAL FEATURES			
Median contrast-enhancing volume in mL (range)	0.0 (0.0 - 7.7)	0.0 (0.0 - 0.0)	0.34
Median whole tumor volume in mL (range)	47.5 (2.8 - 163.0)	33.5 (4.7 - 154.1)	0.22
Thickness of enhancing margin			
- Not Applicable	27 (84%)	12 (92%)	0.66
- Thick/Nodular (=>3mm)	1 (3%)	0 (0%)	1
- Solid	1 (3%)	0 (0%)	1
- Thin (<3mm)	3 (9%)	1 (8%)	1
Mismatch sign at second resection			
- Yes	24 (77%)	4 (36%)	**0.02**
- No	7 (23%)	7 (64%)	**0.02**
- Not available	1	2	0.20

Table 5.S2: Patient characteristics and annotation results at first resection, stratified according to the presence of T2-FLAIR mismatch area. First column (all) includes patients where no sufficient imaging was available. P-value compares columns absent and present. Missing values are reported as 'Not available', but not included in the computation of percentages and p-values.

T2-FLAIR mismatch area	**Absent (n=6)**	**Present (n=39)**	**p-value**
Female sex (%)	2 (33%)	20 (51%)	0.67
Median age at diagnosis in y (range)	30.5 (19.0 - 45.0)	27.0 (18.0 - 70.0)	0.74
Median time since diagnosis in y (range)	0.8 (0.1 - 17.0)	0.3 (0.0 - 12.6)	0.75
Median overall post-surgery survival (OS-R1) in y (95% CI)	N/A (2.3 – N/A)	9.6 (7.8 - 17.2)	0.18
Median post-surgery progression-free survival (PFS-R1) in y (95% CI)	2.8 (1.1 - 5.9)	2.8 (2.1 - 3.3)	0.51
Median time to second resection in y (range)	3.2 (1.4 - 5.9)	3.4 (0.9 - 8.7)	0.60
CNS WHO-2021 grade			
- Grade 2	3 (50%)	28 (72%)	0.36
- Grade 3	0 (0%)	10 (26%)	0.31
- Grade 4	3 (50%)	1 (3%)	**0.005**
KPS before surgery			
- 100	2 (33%)	15 (39%)	1.00
- 90	4 (67%)	20 (53%)	0.67
- <90	0 (0%)	3 (8%)	1.00
- Not available	0	1	1.00
LOCATION			
Side of lesion center			
- Left	2 (33%)	20 (51%)	0.67
- Right	4 (67%)	19 (49%)	0.67
Location in or near eloquent regions (Sawaya et al.)			
- Eloquent (III)	5 (83%)	25 (64%)	0.65
- Near-eloquent (II)	1 (17%)	5 (13%)	1.00
- Non-eloquent (I)	0 (0%)	9 (23%)	0.32
Tumor site (multiple sites possible)			
- Frontal lobe	4 (67%)	29 (74%)	0.65
- Temporal lobe	1 (17%)	14 (36%)	0.65
- Insula	1 (17%)	13 (33%)	0.65
- Corpus callosum	1 (17%)	1 (3%)	0.25
- Parietal lobe	2 (33%)	10 (26%)	0.65
- Occipital lobe	1 (17%)	2 (5%)	0.36
- Basal ganglia	0 (0%)	4 (10%)	1.00
- Thalamus	0 (0%)	1 (3%)	1.00

Table 5.S2: (cont.)

T2-FLAIR mismatch area	**Absent (n=6)**	**Present (n=39)**	**p-value**
TREATMENT AFTER RESECTION			
Extent of resection			
- Partial resection	3 (50%)	30 (77%)	0.32
- Complete resection	3 (50%)	9 (23%)	0.32
- Not available	0 (0%)	0 (0%)	1.00
Radiotherapy	2 (33%)	7 (18%)	0.58
Chemotherapy	3 (50%)	1 (3%)	**0.005**
PRIOR TREATMENT			
Radiotherapy	1 (17%)	0 (0%)	0.13
Chemotherapy	1 (17%)	0 (0%)	0.13
Biopsy	1 (17%)	0 (0%)	0.13
RADIOLOGICAL FEATURES			
Median contrast-enhancing volume (CET) in mL (range)	0.0 (0.0 - 7.7)	0.0 (0.0 - 0.1)	0.05
Median whole tumor volume (WT) in mL (range)	37.7 (2.8 - 72.9)	46.5 (4.7 - 163.0)	0.55
Thickness of enhancing margin			
- Not Applicable	4 (67%)	35 (90%)	0.18
- Thick/Nodular (=>3mm)	0 (0%)	1 (3%)	1.00
- Solid	1 (17%)	0 (0%)	0.13
- Thin (<3mm)	1 (17%)	3 (8%)	0.45
Mismatch sign at second resection			
- Yes	0 (0%)	14 (39%)	0.08
- No	6 (100%)	22 (61%)	0.08
- Not available	0	3	1.00

Table 5.S3: Patient characteristics and annotation results at second resection, stratified according to the presence of T2-FLAIR mismatch area. First column (all) includes patients where no sufficient imaging was available. P-value compares columns absent and present. Missing values are reported as 'Not available', but not included in the computation of percentages and p-values.

T2-FLAIR mismatch area	**Absent (n=40)**	**Present (n=36)**	**p-value**
Female sex (%)	15 (38%)	17 (47%)	0.49
Median age at diagnosis in y (range)	34.0 (19.0 - 70.0)	28.0 (18.0 - 55.0)	0.04
Median time since diagnosis in y (range)	4.5 (1.2 - 23.5)	4.1 (1.0 - 22.7)	0.39
Median overall post-surgery survival (OS-R2) in y (95% CI)	5.2 (2.6 - 6.6)	7.6 (5.8 – N/A)	**0.009**
Median post-surgery progression-free survival (PFS-R2) in y (95% CI)	1.5 (0.9 - 2.8)	3.4 (2.4 - 5.2)	0.07

Table 5.S3: (cont.) *The growth pattern could not be determined if no reference scan was available prior to progression (see section 5.2.5).

T2-FLAIR mismatch area	**Absent (n=40)**	**Present (n=36)**	**p-value**
CNS WHO-2021 grade			
- Grade 2	19 (48%)	27 (75%)	**0.02**
- Grade 3	7 (18%)	2 (6%)	0.16
- Grade 4	14 (35%)	7 (19%)	0.20
KPS before surgery			
- 100	16 (40%)	19 (53%)	0.36
- 90	15 (38%)	12 (33%)	0.81
- <90	9 (22%)	5 (14%)	0.39
LOCATION			
Side of lesion center			
- Left	17 (42%)	20 (56%)	0.36
- Right	23 (57%)	16 (44%)	0.36
Location in or near eloquent regions			
- Eloquent (III)	29 (72%)	16 (44%)	0.02
- Near-eloquent (II)	5 (12%)	10 (28%)	0.15
- Non-eloquent (I)	6 (15%)	10 (28%)	0.26
Tumor site (multiple sites possible)			
- Frontal lobe	30 (75%)	31 (86%)	0.26
- Temporal lobe	19 (48%)	10 (28%)	0.10
- Insula	17 (42%)	9 (25%)	0.15
- Corpus callosum	5 (12%)	2 (6%)	0.44
- Parietal lobe	14 (35%)	4 (11%)	**0.02**
- Occipital lobe	3 (8%)	0 (0%)	0.24
- Brainstem	0 (0%)	1 (3%)	0.47
- Basal ganglia	1 (2%)	0 (0%)	1.00
- Thalamus	1 (2%)	0 (0%)	1.00
TREATMENT AFTER RESECTION			
Extent of resection			
- Partial resection	26 (65%)	23 (64%)	1.00
- Complete resection	14 (35%)	13 (36%)	1.00
- Radiotherapy	9 (22%)	18 (50%)	**0.02**
- Chemotherapy	12 (30%)	14 (39%)	0.47
PRIOR TREATMENT			
Radiotherapy	18 (45%)	8 (22%)	0.05
Number radiotherapy of treatments			
- 1	17 (42%)	8 (22%)	0.09
- 2	1 (2%)	0 (0%)	1.00
Chemotherapy	10 (25%)	2 (6%)	**0.03**
Number chemotherapy of treatments			
- 1	6 (15%)	2 (6%)	0.27
- 2	2 (5%)	0 (0%)	0.49
Biopsy	4 (10%)	3 (8%)	1.00

Table 5.S3: (cont.) *The growth pattern could not be determined if no reference scan was available prior to progression (see section 5.2.5).

T2-FLAIR mismatch area	**Absent (n=40)**	**Present (n=36)**	**p-value**
RADIOLOGICAL FEATURES			
Median contrast-enhancing volume (CET) in mL (range)	0.3 (0.0 - 63.8)	0.0 (0.0 - 4.1)	**<0.001**
Median whole tumor volume (WT) in mL (range)	20.4 (1.7 - 149.3)	16.9 (1.3 - 76.8)	0.29
Thickness of enhancing margin			
- Not Applicable	16 (40%)	29 (81%)	**<0.001**
- Thin (<3mm)	9 (22%)	6 (17%)	0.58
- Thick/Nodular (=>3mm)	10 (25%)	1 (3%)	**0.008**
- Solid	5 (12%)	0 (0%)	0.06
Growth pattern			
- Mostly invasive	10 (26%)	7 (22%)	0.78
- Mostly expansive	1 (3%)	6 (19%)	**0.04**
- Mixed	11 (29%)	10 (31%)	1.00
- Not sure	16 (42%)	9 (28%)	0.32
- Not available*	2	4	0.414

Table 5.S4: Confusion matrices of enhancement and T2-FLAIR mismatch sign at second resection versus CNS WHO 2021 grade.

	Enhancement		**T2-FLAIR sign**		**Total**
	Non-enhancing	**Enhancing**	**Yes**	**No**	
Grade 2	32	14	20	26	46
Grade 3	6	3	2	7	9
Grade 4	5	16	3	18	21
Total	43	33	25	51	76
	Non-enhancing plus sign				**Total**
	Yes	**No**			
Grade 2	17	29			46
Grade 3	2	7			9
Grade 4	1	20			21
Total	20	56			76

Table 5.S5: Confusion matrix of microcysts and T2-FLAIR mismatch, for samples obtained at first and second resection. P-value result of Fisher's exact test between the presence of microcysts (YES / NO) and the mismatch sign (vs. rest) and area (vs rest).

	First resection			**Second resection**		
Microcysts	**YES**	**NO**	**p-value**	**YES**	**NO**	**p-value**
Mismatch sign	6	5	1.0	21	4	0.04
Mismatch area (including sign)	23	13	0.005	29	7	0.03
Neither	0	6		22	17	
Total	23	19		51	24	

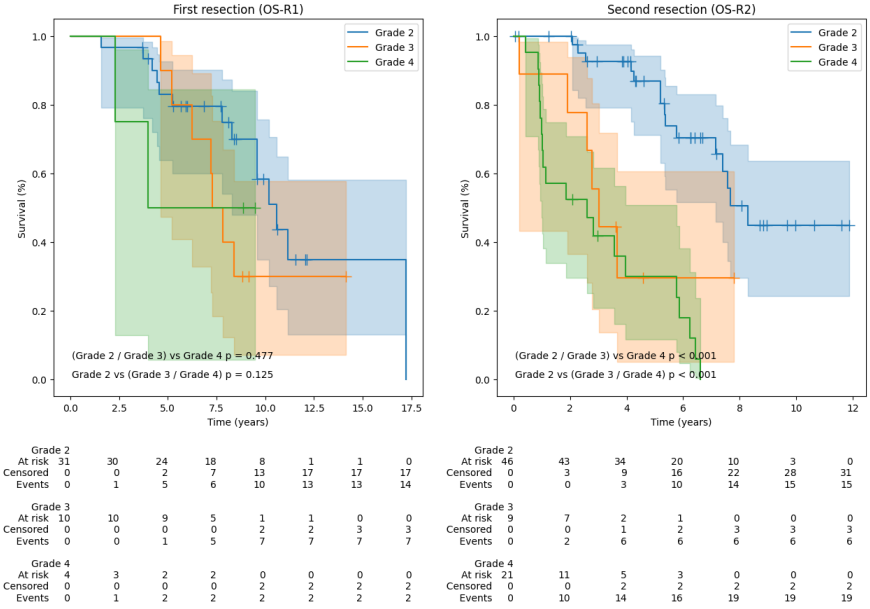


Figure 5.S1: Kaplan Meier curves of different WHO 2021 grades at first and second resection. Censored patients indicated by a '+' at date of last follow-up. Shaded areas indicate 95% confidence intervals.

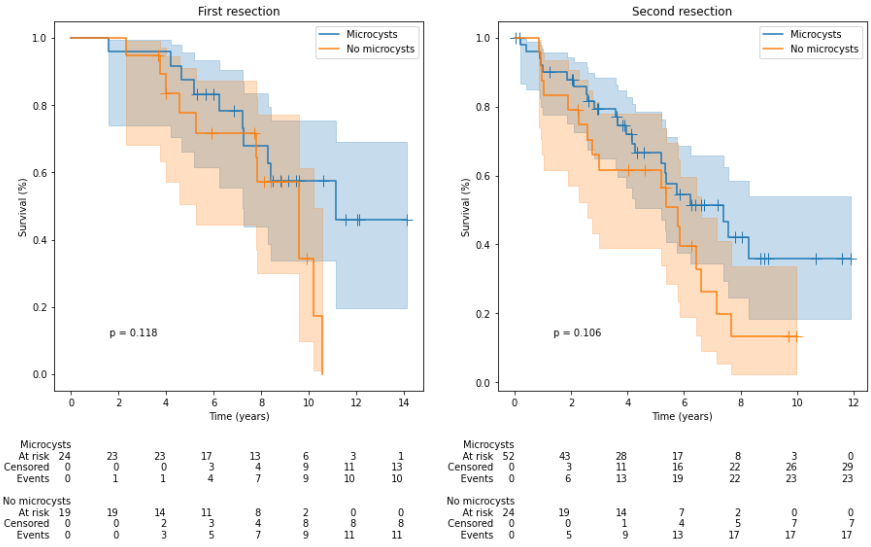
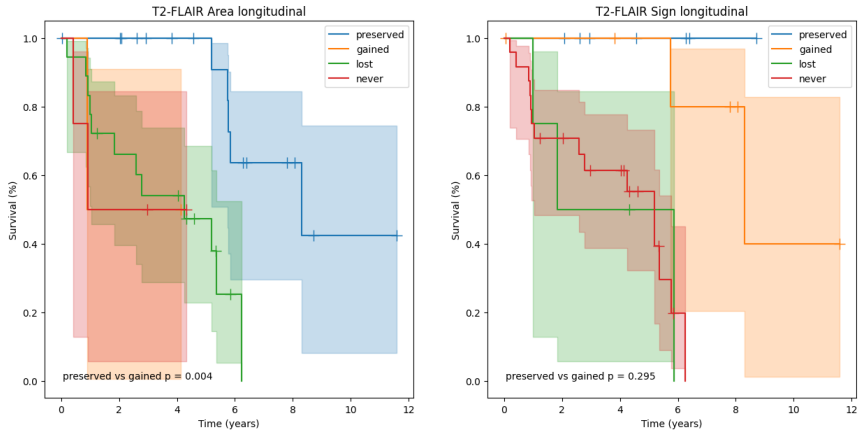


Figure 5.S2: Kaplan Meier curves of microcysts at first and second resection. Censored patients indicated by a '+' at date of last follow-up. Shaded areas indicate 95% confidence intervals.



preserved							
At risk	18	17	12	7	4	1	0
Censored	0	1	6	7	10	12	13
Events	0	0	0	4	4	5	5
gained							
At risk	2	1	1	0	0	0	0
Censored	0	0	0	1	1	1	1
Events	0	1	1	1	1	1	1
lost							
At risk	18	11	9	1	0	0	0
Censored	0	1	1	6	6	6	6
Events	0	6	8	11	12	12	12
never							
At risk	4	2	1	0	0	0	0
Censored	0	0	1	2	2	2	2
Events	0	2	2	2	2	2	2

preserved							
At risk	7	7	4	3	1	0	0
Censored	0	0	3	4	6	7	7
Events	0	0	0	0	0	0	0
gained							
At risk	7	6	5	4	3	1	0
Censored	0	1	2	2	3	4	5
Events	0	0	0	1	1	2	2
lost							
At risk	4	2	2	0	0	0	0
Censored	0	0	0	1	1	1	1
Events	0	2	2	3	3	3	3
never							
At risk	24	16	12	1	0	0	0
Censored	0	1	3	9	9	9	9
Events	0	7	9	14	15	15	15

Figure 5.S3: Kaplan Meier curves of longitudinal changes in T2-FLAIR mismatch area (left) and T2-FLAIR mismatch sign (right) for OS-R2. Preserved: present at both first and second resection, gained: present at second resection, absent at first, lost: present at first resection, absent at second, never: absent at both first and second resection. Censored patients indicated by a '+' at date of last follow-up. Shaded areas indicate 95% confidence intervals.

6

Deep learning-based groupwise registration for longitudinal MRI analysis in glioma

Measure what is measurable, and make measurable what is not so.

– Galileo Galilei

Not everything that can be counted counts, and not everything that counts can be counted.

–William Bruce Cameron, 1963

Based on: C. C. Hammecher†, **Karin A. van Garderen**, M. Smits, P. Wesseling, B. Westerman, P. French, M. Kouwenhoven, R. Verhaak, F. Vos, E. Bron, and B. Li, "Deep learning-based groupwise registration for longitudinal MRI analysis in glioma," *Abstract presented at ISMRM 2023*. arXiv: 2306.10611

† indicates presenting author

6.1 Introduction

Glioma progression is monitored by routine MR scanning, enabling that tumor growth can be evaluated with respect to earlier time-points. This growth may present both as a mass effect and as an extension of abnormalities into previously healthy tissue. To accurately assess tumor growth and tumor-induced deformations, longitudinal intrasubject image registration is often used. However, such registration in cases with large deformations and tissue change is highly challenging.

Longitudinal image registration may benefit from groupwise strategies in which multiple images are concurrently aligned. This avoids introducing bias towards an a priori selected reference image [100]. However, existing learning-based methods for image registration mostly concern pair-wise approaches [101]. Moreover, the few proposed learning-based methods for groupwise registration are designed for analysis of images without pathologies, and are prone to fail registering glioma images. To bridge this gap, we present a learning-based method for non-linear registration of longitudinal glioma images.

6.2 Methods

We used T2-weighted FLAIR MRI scans of 61 participants from the multi-center GLASS-NL study [102]. Participants were initially diagnosed with lower-grade (grade 2 or 3) IDH-mutant astrocytoma and underwent multiple surgical resections. Images were affinely aligned to the ICBM 2009a nonlinear asymmetric atlas [87], skull-stripped and intensity normalized. We obtained tumor [17, 19] and normal-appearing tissue segmentations [103]. For each subject, we grouped available scans before or after a surgical resection into all possible permutations of three time-points. The data was split into 46:15 patients (3349:90 permutations) for training and testing.

We expanded an existing learning-based registration approach [101] to take tumor presence and growth into account. During training, the method estimates the diffeomorphic deformations to the permutation's mean-space, maximizes the local cross-correlation across the warped images, and encourages a smooth and continuous deformation. To be robust against possible intensity alterations in the tumor region, a loss-function masking strategy was implemented to compute the loss value only in the normal-appearing region across the three time-points. In addition, to register large local mass-effects caused by gliomas, we estimated the deformation at two resolutions, to firstly register the general structures in down-sampled images, and secondly refine the residual deformations at full resolution (Fig. 6.1).

We evaluated the proposed method against state-of-the-art classical groupwise registration methods: Elastix [104], NiftyReg [105], and ANTs [106]. These

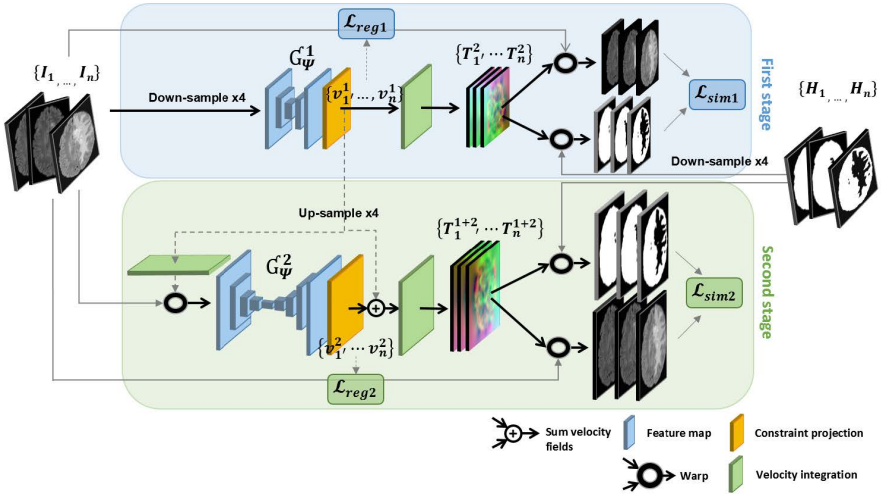


Figure 6.1: Schematic representation of the proposed framework. The first stage is trained on the down-sampled FLAIR images I_n. After training, the velocity fields v_n^1 are up-sampled to warp I_n and serve as an input to the second stage. Here, the residual deformation fields are obtained from $v_n^1 + v_n^2$ and applied to I_n to reduce interpolation error. In the second stage, the training parameters G_Ψ^1 are fixed. In both stages, the normal-appearing masks Hn are included in the loss function.

were run with default parameters, providing normal tissue masks as input when this option was available (i.e., Elastix and NiftyReg).

The similarity across the warped images was assessed by the Dice coefficients, and the average structural similarity index measure (SSIM) between warped image and the average image [107]. Also, the centrality was evaluated by the average norm of the three resulting deformations. What is more, the smoothness of the deformations was measured by the number of foldings (negative values) in the Jacobian maps and their average standard deviation [104]. All metrics were computed in the normal-appearing tissue.

6.3 Results

Figure 6.2 presents the average Dice and SSIM scores of all test permutations by the initial affine registration, the classical methods, and the proposed framework. Our single-stage method ('mask only') performed comparably to the classical methods in terms of Dice coefficient. The average SSIM obtained by our method was higher than for these classical methods, except Elastix. On the other hand, our multi-stage implementation ('mask+multi-stage') improves both Dice and SSIM coefficients with respect to the single-stage.

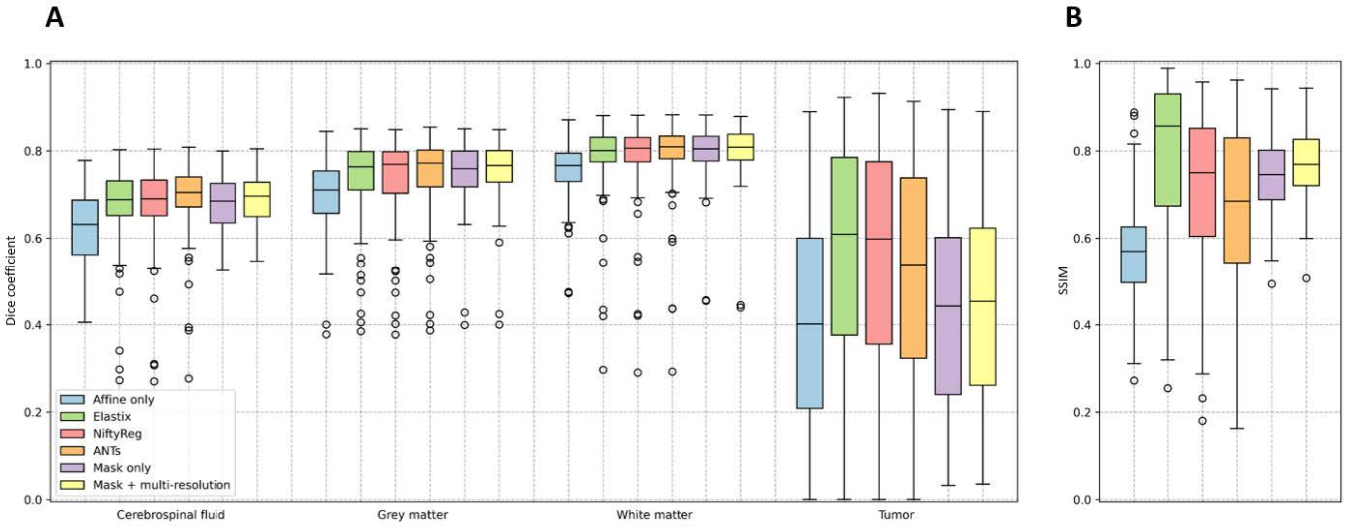


Figure 6.2: Registration accuracy for affine alignment only, Elastix, NiftyReg, ANTs, and our proposed method with and without multi-resolution implementation. In A, the boxplots of Dice coefficients for CSF, grey matter, white matter, and tumor region are shown. In B, the boxplots for average SSIM in normal-appearing tissue are depicted.

Elastix presents the best centrality, followed by our multi-resolution strategy (Table 6.1). The proposed strategies show improvement in smoothness and have inference runtimes of under a minute, significantly faster than the classical approaches. In a qualitative example (Fig. 6.3 and Fig. 6.4), the stronger deformations of Elastix lead to more overlap of the tumor across images, but with non-anatomically plausible deformations near the tumor edge. The proposed methods accurately align the normal-appearing tissue, but did not align the resection volume.

6.4 Discussion

The proposed method is able to register glioma images despite the presence of non-correspondences across the time-points by focusing on the normal-appearing tissue similarity. The obtained GM and WM Dice coefficients are comparable to those of state-of-the-art toolboxes, but with higher SSIM values, suggesting that the registrations are more detailed. Elastix and NiftyReg show larger tumor Dice but stronger deformations, which could indicate anatomically implausible registration of non-correspondences. Qualitatively, our method shows stronger misalignment of the resection volume. This could indicate that changes in such volume are identified as non-correspondences instead of mass-effect.

Our method also achieves smoother deformations with the least foldings. An important advantage of our network approach is that new images can be registered in seconds, which is much faster than the classical methods (e.g., 28 hours by ANTs). We showed that the multi-stage strategy combined with the tumor masks yields higher registration accuracy than without this strategy, as this allows large, smooth deformations while avoiding local minima. However, for the cases with extremely large mass-effect, further refinement of the method could be considered.

6.5 Conclusion

The proposed deep learning-based unbiased group-wise registration method can serve as an alternative to existing classical toolboxes for the analysis of glioma growth in longitudinal MRI.

Table 6.1: Centrality and smoothness of the estimated deformations, and the required runtime in minutes using the methods Elastix, NiftyReg, ANTs, and the proposed framework with and without multi-resolution implementation. Results are computed within the normal tissue and averaged over the test set.

	↓ Centrality	↓ Smoothness		↓ Run time
		$\lvert J_\Phi\rvert \leq 0[\%]$	SD $\lvert J_\Phi\rvert$	[min]
Elastix	**1.0e-14**	6.7e-02	0.13	22
NiftyReg	4.9e-02	1.2e-02	0.16	33
ANTs	6.7e-02	3.2-03	0.11	1647
Proposed (mask only)	1.0e-03	**0.0**	0.092	**<1**
Proposed (mask + multi-resolution)	1.7e-03	6.0e-07	**0.085**	**<1**

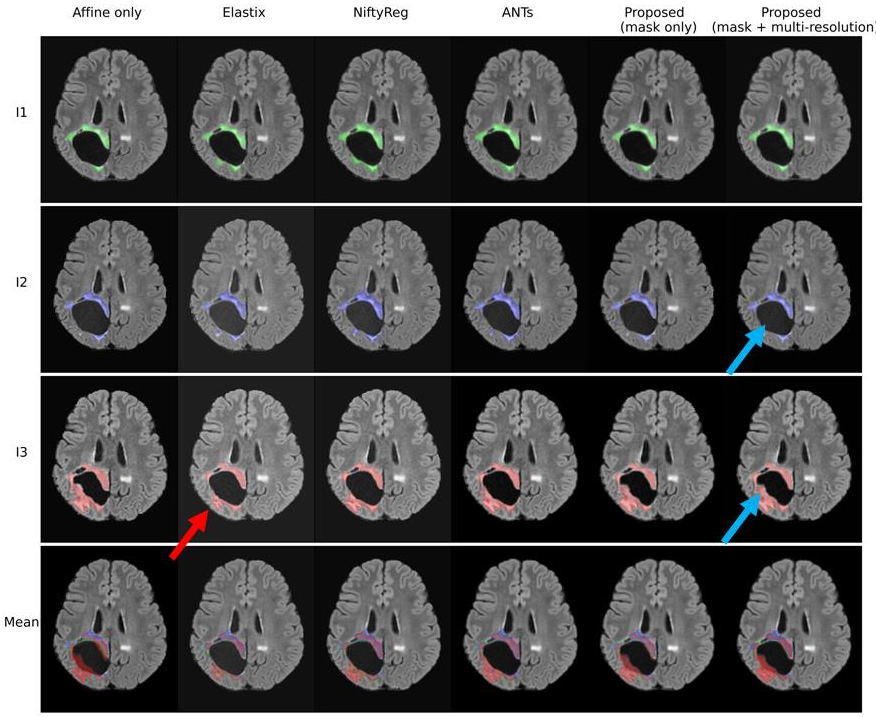


Figure 6.3: Results of one longitudinal permutation with images I_1, I_2, and I_3 taken 3, 16, and 36 months after surgery. Overlaid on the axial slices the warped tumor segmentations. The last row shows the average image across the warped images and all tumor segmentations. Red arrow: excessive compression of tumor. Blue arrows: resection cavity not aligned.

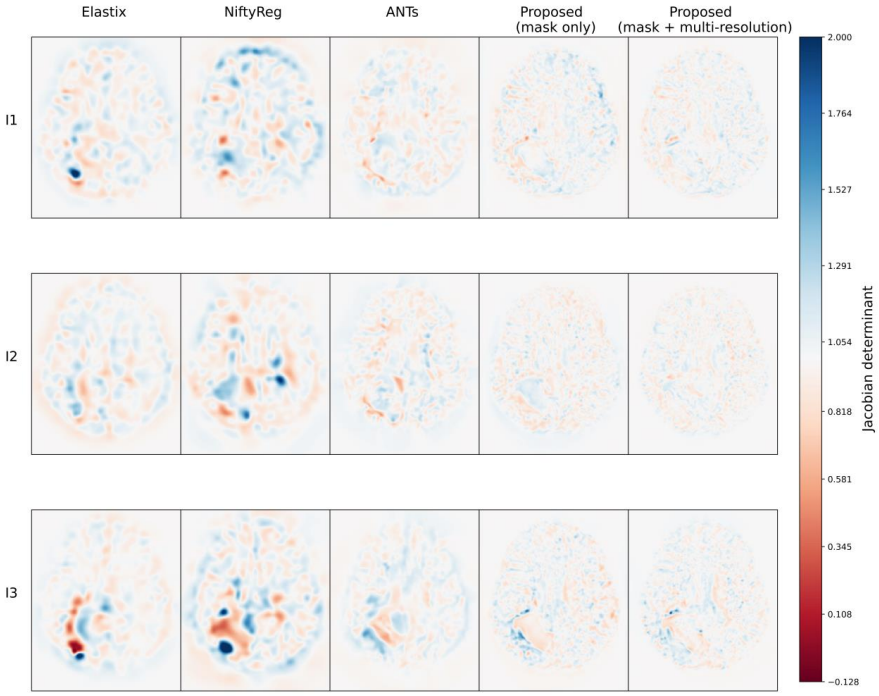


Figure 6.4: Example axial slices Jacobian maps, corresponding to the results observed in Figure 3. Expansions with respect to the mean-space are depicted in red, while shrinking is in blue.

7

Evaluating the predictive value of glioma growth models for low-grade glioma after tumor resection

Now I know what a ghost is. Unfinished business, that's what.

–Salman Rushdie, The Satanic Verses

Based on: **Karin A. van Garderen**, S. R. van der Voort, M. M. Wijnenga, F. Incekara, A. Alafandi, G. Kapsas, R. Gahrmann, J. W. Schouten, H. J. Dubbink, A. J. Vincent, M. van den Bent, P. J. French, M. Smits, and S. Klein, "Evaluating the predictive value of glioma growth models for low-grade glioma after tumor resection," *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, pp. 253–263, 2024

Abstract

Tumor growth models have the potential to model and predict the spatiotemporal evolution of glioma in individual patients. Infiltration of glioma cells is known to be faster along the white matter tracts, and therefore structural magnetic resonance imaging (MRI) and diffusion tensor imaging (DTI) can be used to inform the model. However, applying and evaluating growth models in real patient data is challenging. In this work, we propose to formulate the problem of tumor growth as a ranking problem, as opposed to a segmentation problem, and use the average precision (AP) as a performance metric. This enables an evaluation of the spatial pattern that does not require a volume cut-off value. Using the AP metric, we evaluate diffusion-proliferation models informed by structural MRI and DTI, after tumor resection. We applied the models to a unique longitudinal dataset of 14 patients with low-grade glioma (LGG), who received no treatment after surgical resection, to predict the recurrent tumor shape after tumor resection. The diffusion models informed by structural MRI and DTI showed a small but significant increase in predictive performance with respect to homogeneous isotropic diffusion, and the DTI-informed model reached the best predictive performance. We conclude there is a significant improvement in the prediction of the recurrent tumor shape when using a DTI-informed anisotropic diffusion model with respect to istropic diffusion, and that the AP is a suitable metric to evaluate these models. All code and data used in this publication are made publicly available.

7.1 Introduction

As the automated image-based diagnosis and delineation of glioma has improved, partially due to the emergence of machine learning techniques and the availability of large public datasets [18], the next major step is that of predicting the disease trajectory. A long history of modelling tumor growth as a biophysical process of diffusion and proliferation has shown promise in predicting the development of glioma in real patient data [33]. However, there is currently no consensus in how to precisely formulate and evaluate the tumor growth problem. The large variety in approaches, from image processing to model fitting, and the lack of public data make it difficult to compare models and estimate their predictive value. Furthermore, the predictive value of tumor growth models in treated low-grade glioma (LGG) is not well studied.

Predicting the spatial patterns of tumor growth can aid diagnosis and treatment in several ways [33]. Local treatment such as radiotherapy can be informed by the most likely pattern of recurrence and the location and extent of tumor infiltration beyond the visible boundaries [34, 108, 109]. Additionally, a growth model can be used to aid automated image analysis methods such as tumor segmentation and image registration [110, 111]. Furthermore, especially in the case of models that are rooted in the biophysical understanding of tumor growth, a better prediction indicates a better understanding of the disease or a better fit to specific patient cases. Knowing the infiltrative behavior of glioma cells, and especially being able to differentiate that behavior between patients, could have prognostic value [112, 113]. Furthermore, an accurate model of tumor growth could aid in the differentiation between true progression and pseudoprogression by identifying changes that do not adhere to the expected growth pattern [114].

The modelling of LGG especially presents a large potential for clinical application. Patients presenting with LGG have a better prognosis compared to high-grade glioma (HGG), but there is no consensus on the optimal treatment [115]. A pro-active treatment regimen might be effective at increasing overall survival, but there is also a risk of unnecessary treatment burden. With their potential to predict and increase the efficacy of treatment, predictive modelling may be of value in the application to LGG [113].

The problem of tumor growth is challenging in many ways, not in the least due to limited observations used to calibrate and evaluate tumor growth models. Magnetic resonance (MR) imaging is commonly used as both the input for calibration of model parameters and evaluation of results, as it provides a non-invasive visualization of both the tumor and surrounding tissue. However, the relation between imaging characteristics and actual tumor cell density is difficult to characterize exactly [116]. While more precise observations exist, such as tissue samples [108] or PET imaging [34], for most clinical cases the best available approximation of glioma infiltration is the delineation of abnormalities

on MR imaging. It is well-recognized that LGG show differences in growth behavior from HGG. In LGG, often the entire lesion is non-enhancing, but tumor cells are known to infiltrate even beyond the visible boundaries on MR imaging [117]. A number of studies have evaluated growth models on patients with LGG [118, 119, 120], but the lower incidence and slower growth compared to HGG make it difficult to accumulate large datasets and to observe growth patterns accurately.

Due to the nature of the available data, growth predictions are often evaluated as a segmentation problem, for example by using an overlap metric such as the Dice similarity coefficient (DSC) based on a single sample in time [31, 121]. Although this metric comes natural to the ground-truth data, it may be less representative of the underlying problem. The main disadvantage of overlap-based metrics is that they treat all voxels equally, independent of their location. Intuitively, we would want to assign more significance to false predictions at a large distance to the predicted tumor boundary, as changing their prediction would require a larger change to the model. This intuition is represented in metrics based on the segmentation boundary, such as the symmetric surface distance [119], but even a distance metric considers only a single binary prediction. Using a boundary metric also becomes less appropriate when the ground truth contains new, disconnected lesions.

The personalized fitting of model parameters provides an additional challenge in the comparison of different models. A good prediction is the result of both a good fit to the initial situation and a prediction of future behavior. Although the ideal model would fit both perfectly, more complex models are at an advantage with more degrees of freedom to fit the initial tumor shape. Given the ill-posedness of that initial problem, there is no guarantee that the improved fit always translates to a better estimate of future behavior. At the same time, the nature of the problem makes it difficult to distinguish the two, especially when there is limited growth.

In this work, we present a novel approach to the evaluation of tumor growth models and aim to evaluate whether tumor growth models have predictive value in LGG, compared to a trivial baseline of isotropic expansion. Our main contributions are:

1. We propose a novel evaluation metric for tumor growth predictions that elimates the need for a specific volume threshold. As the problem of spatial infiltration is in its essence a ranking problem, we propose to use the average precision as an evaluation metric.

2. Using this metric, we compare models with different assumptions on the spatial diffusion characteristics. Instead of applying patient-wise tuning of the model parameters, which may cause a bias towards models with more degrees of freedom, we selected parameters that represent

common assumptions in diffusion-proliferation models and compare their predictive performance in a clinically representative dataset.

3. All code and data used in our experiments are made publicly available, to aid future innovations in this field.

This work builds upon a preliminary version presented at the 2021 MICCAI Brainlesion workshop [122]. Although the approach to the evaluation as a ranking problem is repeated in this work, we have refined the model definition, evaluation, image processing and patient selection.

7.2 Methods

7.2.1 Tumor growth model

The tumor growth models in this work are all based on a diffusion-proliferation model with parameters to include both an isotropic and anisotropic diffusion component. The model is defined by a partial differential equation for the cell density c, which is updated with each timestep dt according to:

$$\frac{dc}{dt} = \nabla(\mathbb{D}\nabla c) + \rho c(1-c), \tag{7.1}$$

$$\mathbb{D}\nabla c \cdot n_{\delta\Omega} = 0, \tag{7.2}$$

where ρ is the growth factor, $n_{\delta\Omega}$ is the normal vector at the boundary between the brain and cerebrospinal fluid (CSF), and $\mathbb{D}$ is a tensor comprising an isotropic and anisotropic component:

$$\mathbb{D} = \kappa(x)\mathbb{I} + \tau F(x)\mathbb{T}(x), \tag{7.3}$$

where κ and τ are parameters to weigh the two components, $\mathbb{I}$ is the identity matrix, $F(x)$ is the local fractional anisotropy (FA) and $\mathbb{T}$ is the normalized diffusion tensor obtained by dividing the diffusion tensor by the mean diffusivity (as described in [123]).

The isotropic diffusion depends on the local tissue type through separate diffusion factors κ_w and κ_g for white matter and grey matter respectively (as described in [36]). Because the tissue segmentation may contain partial volume effects, the two diffusion parameters are weighted by the local tissue probabilities p_w and p_g:

$$\kappa(x) = \kappa_w p_w(x) + \kappa_g p_g(x) \tag{7.4}$$

The brain boundary is also derived from the local tissue probabilities by setting a threshold p_b on the combined tissue probabilities $p_w(x) + p_g(x) < p_b = 0.8$. This threshold was chosen so that the sulci are optimally visible in

the CSF segmentation, preventing the model from growing across the cortical folds.

From the prediction $c(t, x)$ we can derive a segmentation $S(t)$ by applying a visibility threshold $c(t, x) > c_v$, which is set at $c_v = 0.5$ in this work. The initial condition of the model is provided by an initial cell density $c(t = 0)$ which is defined as a gaussian distribution centered at a seed location x_s with a standard deviation of 1mm. When applying to patient data, the seed location is set to the center of gravity of the initial tumor segmentation.

The model was implemented in FEniCS [124] in a cubic mesh of 1mm isotropic cells, using a finite element approach and Crank-Nicolson approximations for the time stepping. The model has four parameters (ρ, τ, κ_w, κ_g) and an additional implicit parameter in the form of the seed location x_s.

7.2.2 Model selection

Generally, each patient will present with a different rate of diffusion and proliferation, and the variation in relative diffusivities in white and gray matter is not known. However, to fit both the initial location and model parameters on a single observation of the tumor is an ill-posed inverse problem. Rather than optimizing the parameters for each individual patient, we opted to design three different models:

- *BASE* As a baseline. The diffusion tensor is isotropic ($\tau = 0$) and the same in both gray and white matter ($\kappa_w = \kappa_g$), and only limited by the boundaries of the brain.
- *TISSUE* This model also has an isotropic diffusion tensor ($\tau = 0$), but the rate of diffusion depends on the tissue type ($\kappa_w > \kappa_g$).
- *DTI* The *DTI* model has an anisotropic diffusion tensor ($\tau > 0$), informed by the local DTI tensor and FA measurement. The weight of the isotropic element of the diffusion tensor ($\kappa(x)$) is the same for both gray and white matter ($\kappa_g = \kappa_w$), assuming that the difference in diffusion is captured in the anisotropic element.

The parameters (ρ, τ, κ_g, κ_w) were selected to achieve a similar relative diffusivity ($\rho / \operatorname{Tr}(\mathbb{D})$) in white matter and overall growth speed, while showing a clear difference in tumor shape. The tissue type and local DTI measurements were inferred from a healthy brain atlas (see Section 7.2.6). Table 7.1 shows the tuned parameter settings for the three growth models.

Table 7.1: Growth parameters selected for the different models.

Model	ρ $1/day$	τ mm^2/day	κ_w mm^2/day	κ_g mm^2/day
BASE	0.005	0	0.1	0.1
TISSUE	0.005	0	0.1	0.01
DTI	0.005	0.5	0.01	0.01

7.2.3 Evaluation metric

In this section, we propose that tumor growth prediction could be framed as a ranking problem, aimed at predicting the relative time-to-invasion of each voxel in the brain.

We assume that a growth model could produce a segmentation of the tumor $S(t)$ at any time $t > 0$. It may therefore assign to every location x in the brain a time $T(x)$, which is the first time t when the tumor reaches that location. We only require that the estimated $T(x)$ is a ranking of voxels in the brain, such that:

$$T(x_a) > T(x_b) \Leftrightarrow \exists t : x_a \notin S(t), x_b \in S(t). \tag{7.5}$$

Based on this perspective, we propose to use the average precision (AP) as an evaluation metric to assess the spatial accuracy of growth prediction. This metric separates the spatial accuracy from the temporal axis, such that an accurate estimate of the growth speed is not required. This is deliberate, as an estimate of both spatial and temporal growth, from a single initial scan, is beyond the possibilities of most growth models. A separate metric could be used to evaluate the temporal accuracy. The ranking can be evaluated with the ground-truth segmentation S' using the AP, which is defined as the area under the precision-recall (PR) curve:

$$AP = \int_0^1 P(R)dR = \int_0^\infty P(t)\frac{dR(t)}{dt}dt, \tag{7.6}$$

where R and P are the recall and precision, and t is a threshold on the time-to-invasion ranking T, leading to the predicted segmentation $S(t) = \{x : T(x) \leq t\}$, and comparing to the reference segmentation S':

$$P(t) = \frac{|S(t) \cap S'|}{|S(t)|}, R(t) = \frac{|S(t) \cap S'|}{|S'|}. \tag{7.7}$$

Note that although S' is time-dependent, as it depends on the time at which the patient was scanned, t is defined as a threshold on the prediction and unrelated to the actual timing of S'. The AP metric weighs the precision scores with the difference in recall, so that all tumor volume predictions $S(t)$ are taken

into account from the tumor onset to the threshold where the recall is one, which means that the ground-truth segmentation is completely encompassed by the prediction. Although it is possible that the prediction never reaches a perfect precision, e.g. due to poor initialization, we assume that a perfect recall is always possible since each voxel would be assigned some $t < \infty$.

To compare the AP to an overlap-based metric, we consider using the dice similarity coefficient (DSC) at a threshold of equal volume. An evaluation based on a single threshold t would represent a point on the PR curve. If we take a sample at the threshold t_v where the estimated tumor volume equals the observed tumor volume ($|S(t_v)| = |S'|$), the recall is equal to the precision ($R(t_v) = P(t_v)$) and therefore the equal-volume-based DSC (DSC_v) can be expressed as:

$$DSC_v = \frac{2|S(t_v) \cap S'|}{|S(t_v)| + |S'|} = R(t_v) = P(t_v), \tag{7.8}$$

$$t_v = t : |S(t)| = |S'|. \tag{7.9}$$

When comparing two models in terms of their prediction, the AP takes into account the precision and recall at each possible cut-off value in $T(x)$. Even if the binary prediction of a voxel is the same for both models in terms of the DSC_v, a change in its rank $T(x)$ will still affect the AP metric. On the other hand, voxels that are wrongly predicted according to the DSC_v, but are ranked close to the decision threshold t_v, will have a relatively small effect on the AP compared to voxels that are either ranked very early but negative (causing the PR curve to drop early) or ranked very late but positive (causing a low tail of the PR curve). This is illustrated in Fig. 7.1.

Formulating the problem as a ranking and using the AP has a number of additional qualitative advantages. First, the ranking T is correlated to the speed of the tumor boundary. If the ranking is smooth, the gradient of T represents the local movement of the visible tumor boundary. Second, we can quantify the agreement between T and S locally, by using the rank of each voxel $T(x)$ as a threshold on the PR curve and calculating the local precision $P(t = T(x))$ and recall $R(t = T(x))$. Fig. 7.2 illustrates the computation of the AP metric and the resulting local precision and recall.

7.2.4 Simulations

To illustrate these models and their effect on tumor shape and growth speed, simulations were performed with different seed locations both in the deep white matter and cortical gray matter of a healthy brain atlas. The tumor growth was tracked in terms of mean tumor diameter (MTD) ($MTD = 2(3V/4\pi)^{1/3}$, where V is the volume $|S(t)|$), which is known to increase approximately linearly with time, to compare the effective growth rates in simulations. The

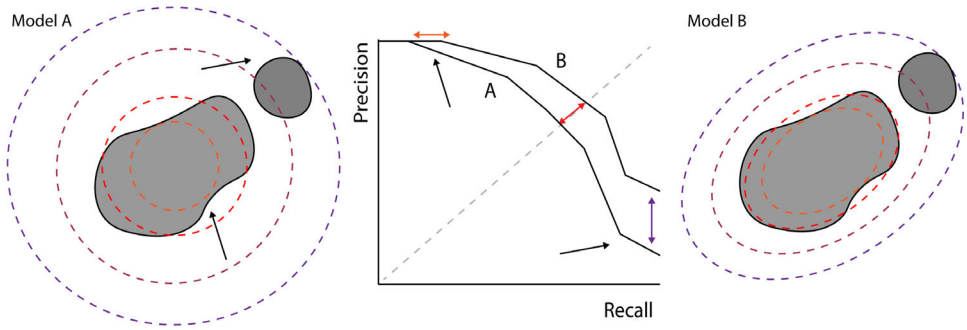


Figure 7.1: Illustration of the AP metric for two model predictions A and B, evaluated with the same ground-truth segmentation (shown in grey). Dotted outlines indicate predictions at different times, so thresholds on $T(x)$. The red dotted outline illustrates the volume-based threshold for DSC_v. The purple (outer) dotted outline shows the threshold at maximum recall. The orange (inner) dotted outline indicates the maximum threshold with perfect precision. The precision-recall curve for both models is illustrated on the right, with the changes at the three thresholds indicated by an arrow in their respective colors. The separate lesion does not affect the DSC_v between model A and B, but it does have an effect on the AP. Arrows indicate the early drop in precision and tail of low recall in model A and their corresponding causes.

effective relative diffusivity can be estimated from the local gradient of $c(x)$, which will be steeper in less diffuse models. Therefore the cell densities $c(x)$ were compared at the time t where the volume $|S(t)|$ was approximately 20 mL for all models. To quantify the effect on tumor shape, the resulting tumor shapes S from the different models were compared in terms of the AP and DSC_v. For this comparison, the $BASE$ model was used as a ground truth S' to evaluate the rankings $T(x)$ generated by the DTI and $TISSUE$ models. In order to investigate the effect of the timing of the follow-up scan, several different cut-off volumes were used to generate S'.

7.2.5 Patient data

A retrospective dataset was selected from Erasmus MC of patients referred for awake craniotomy who a) were diagnosed with a low-grade, IDH-mutant glioma and b) were treated with surgical resection, but received no chemo- or radiotherapy. Three MRI scans were selected: a pre-operative scan (t_0) used for treatment planning, which includes a 3D T1-weighted (T1w) scan and a 2D or 3D T2-weighted (T2w), and two scans acquired during follow-up (t_1 and t_2) both with a pre- and postcontrast T1w scan, a T2w scan and a T2w-FLAIR scan (2D or 3D). A DTI scan was not required as the healthy brain template

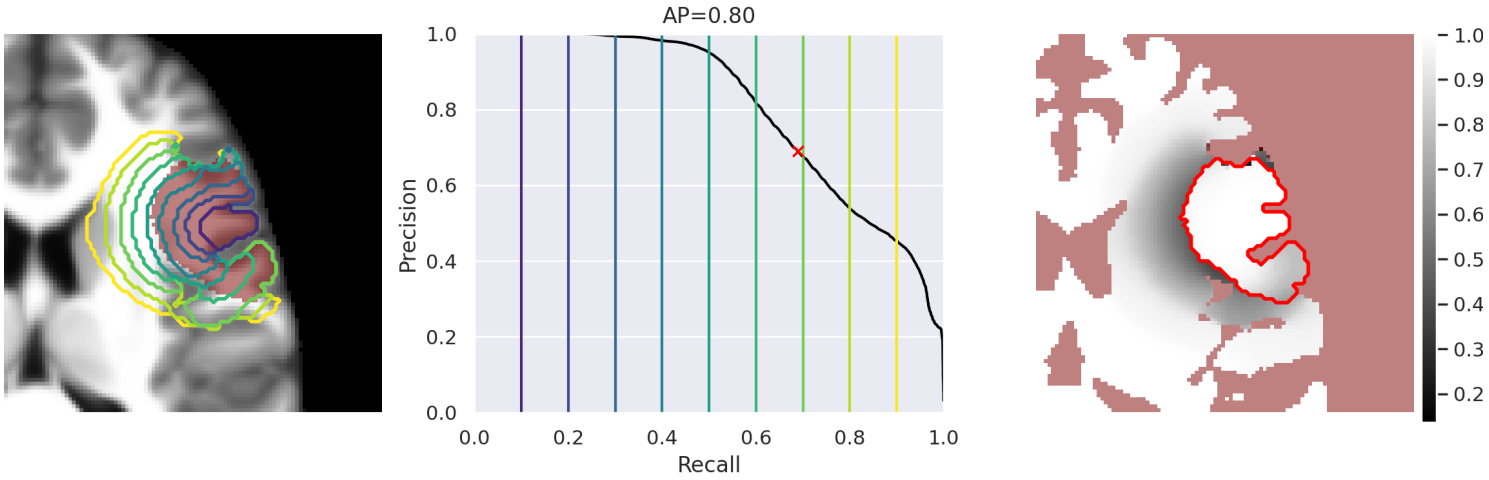


Figure 7.2: Left: cross-section with thresholds on the prediction $T(x)$ generated by a tumor growth model (patient ID 04, *BASE* model). The ground-truth segmentation S' is indicated by a red overlay. Middle: corresponding precision-recall curve with the same thresholds indicated by vertical lines. The value of DSC_v, achieved by thresholding the prediction at equal volume, is indicated by a red x. Right: quantification of agreement by $R(T(x))$ outside S' and $P(T(x)))$ for voxels inside S'. The boundary of S' is indicated by a red outline.

was provided by an atlas.

Patients treated with radiotherapy were excluded due to the possibility of treatment effects, making it impossible to accurately distinguish recurrent tumor growth from treatment-induced abnormalities. The pre-operative scan was used for the initial tumor shape S'_0, while the first available post-operative scan was used to measure the residual volume ($|S'_1|$). The last available follow-up scan before the start of a new treatment was selected to define the recurrent tumor S'_2. Patients were excluded from the analysis if the measured tumor growth ($|S'_2| - |S'_1|$) was less than 5 mL. Any contrast enhancement at t_2 was noted but not a reason for exclusion.

This study design was reviewed by the Erasmus MC Medical Ethics Committee (MEC-2016-419) and performed according to the declaration of Helsinki and the Dutch regulations on medical research.

7.2.6 Image processing and annotation

The application and evaluation of a tumor growth model requires a healthy brain template, since the tumor infiltration affects the diffusion properties and appearance of the brain anatomy. This is commonly approximated by registering to an atlas [34, 36, 110]. In this study, the IIT Human Brain Atlas was used [125, 126] as a template for the DTI tensor and tissue probabilities (p_w, p_g). The patient-specific imaging was used to establish the tumor boundaries and to exclude resected regions of the brain.

The T1w scans at t_0 and t_2 were registered to the atlas using Elastix [63, 127]. As the tumor growth and resection may cause large deformations of the brain, a non-rigid registration was applied after the initial rigid and affine deformations. All registration steps used the mutual information as a metric, and the final non-rigid registration was performed using a b-spline transformation with a 25mm grid spacing. As a sanity check, and to estimate the effect of the non-rigid registration, the change in tumor volumes due to non-rigid registration was computed. A decrease in volume would indicate that the registration has compensated for mass effect in the tumor, while an increase in volume would be unexpected as it represents a shrinking of the tumor-infiltrated tissue. Additionally, the tumor outlines after registration were compared to the original images to visually inspect the registration quality. A heatmap was generated to visualize the combined spatial distribution of the dataset.

Voxels outside of the brain (i.e. CSF) were masked from the evaluation, which means that those voxels were excluded entirely in the computation of the DSC_v and AP. In pre-operative imaging (t_0) the boundaries of the brain were determined from the atlas as described in section 7.2.1. For the post-operative imaging (t_2), a patient-specific CSF mask was needed to capture the resection cavity. For this purpose, FSL FAST [128] was used to extract

the CSF from the T2w-FLAIR scan. FAST can be used to cluster brain voxels in a predetermined number of classes, depending on the tissue contrast and presence of lesions. Due to the variation in intensity in the lesion and healthy tissue within the T2w-FLAIR image, we found a total of 6 tissue classes to be optimal to separate the CSF.

For the pre-operative images, which did not include a T2w-FLAIR sequence, the initial tumor S'_0 was segmented manually on the T2w scan. Tumor segmentations S'_1 and S'_2 for consecutive images were produced using HD-GLIO [17, 19] and corrected manually, if needed, using ITK-Snap [96].

7.2.7 Evaluation

The three models, *BASE*, *TISSUE*, and *DTI*, were applied to the patient data by setting the initial location x_s to the center of gravity of the initial tumor segmentation S'_0. The rankings $T(x)$ generated by the growth models were compared to the initial tumor segmentation S'_0 and follow-up segmentation S'_2, in terms of the AP and the DSC_v. The reference segmentations S'_0 and S'_2 and the volume threshold used for the DSC_v metric were computed after non-rigid registration to the atlas and masking of voxels outside the atlas brain volume or in the CSF (for S'_2), as described in section 7.2.6. The performance metrics were compared using the Wilcoxon signed rank test, using $p<0.05$ as a threshold for significance. To investigate the effect of the brain boundaries on model performance, we repeated the evaluation with $p_b = 0.5$ and $p_b = 0.9$.

7.2.8 Data availability

All code was made publicly available through Gitlab[1]. The imaging data and patient-specific results, such as the segmentations, registered volumes and predictions, were made publicly available through Health-RI XNAT[2].

7.3 Results

7.3.1 Simulations

The resulting growth patterns of the simulation experiments, started from three different seed locations, are visualized in Fig. 7.3 in terms of the cell density distribution $c(x)$ and in Fig. 7.4 in terms of the growth speed. We can observe that the effective growth speed (MTD over time) and the effective diffusivity, which is represented by the gradient of $c(x)$ at the edge of the tumor, are similar between models. Fig. 7.5 compares the isodensity contours of the different models at the three seed locations. Fig. 7.6 shows the error

[1] https://gitlab.com/neuroonco/growthranking
[2] https://xnat.bmia.nl/data/projects/lgg-grow

with respect to the $BASE$ model in terms of the DSC and AP metrics as it depends on the volume of the ground-truth segmentation S'. Based on these simulations, the estimated difference between the models is in the order of 5% for DSC_v or 1% for AP, that does not change much beyond a reference volume of 20mL.

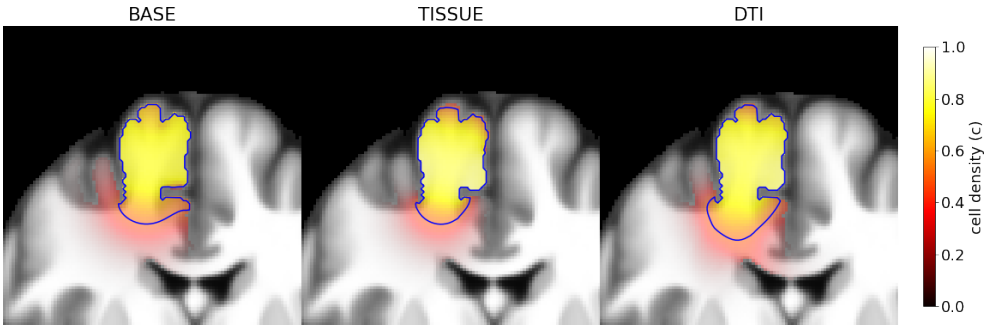


Figure 7.3: Comparison of models in terms cell density distribution $c(x)$ at $t = 900$, where $|S(t)| = 20mL$ approximately, with the isocontour c_v indicated by a blue line (Left: $BASE$, Middle: $TISSUE$, Right: DTI).

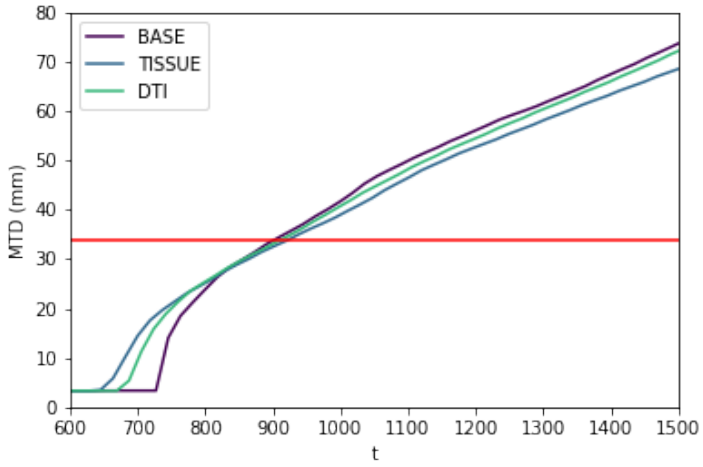


Figure 7.4: Comparison of models in terms of effective growth speed. as mean tumor diameter (MTD) is shown as a function of time for the three growth models.

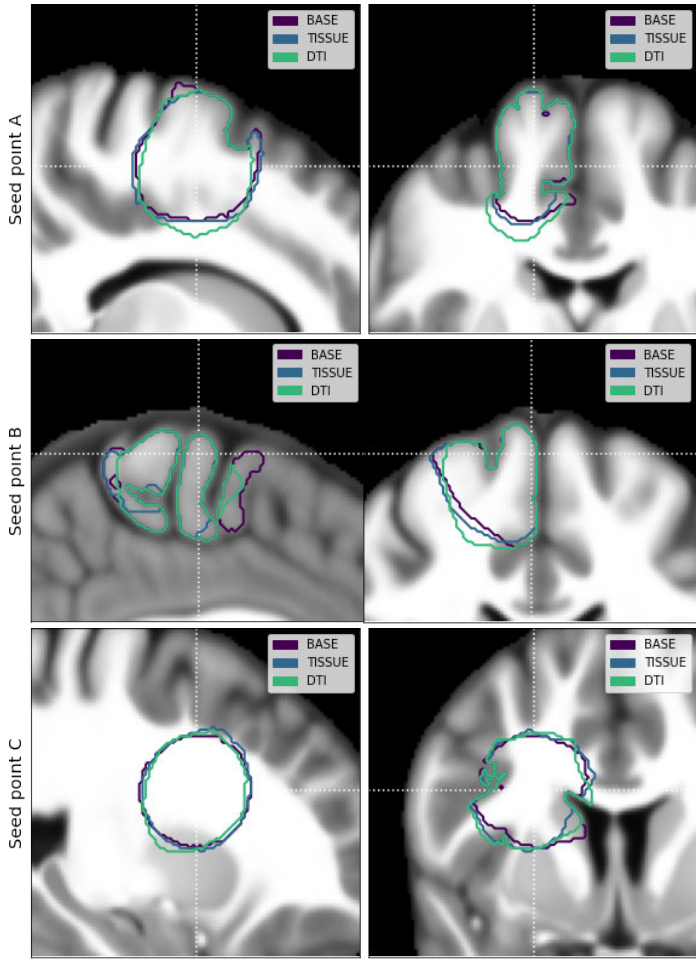


Figure 7.5: Comparison of models in terms of shape of the resulting segmentation obtained in simulations with three different seed locations x_s. Isodensity contours of c_v at 20mL, shown as an outline on the healthy brain atlas (T1w). Intersection of the white dotted lines is the seed location.

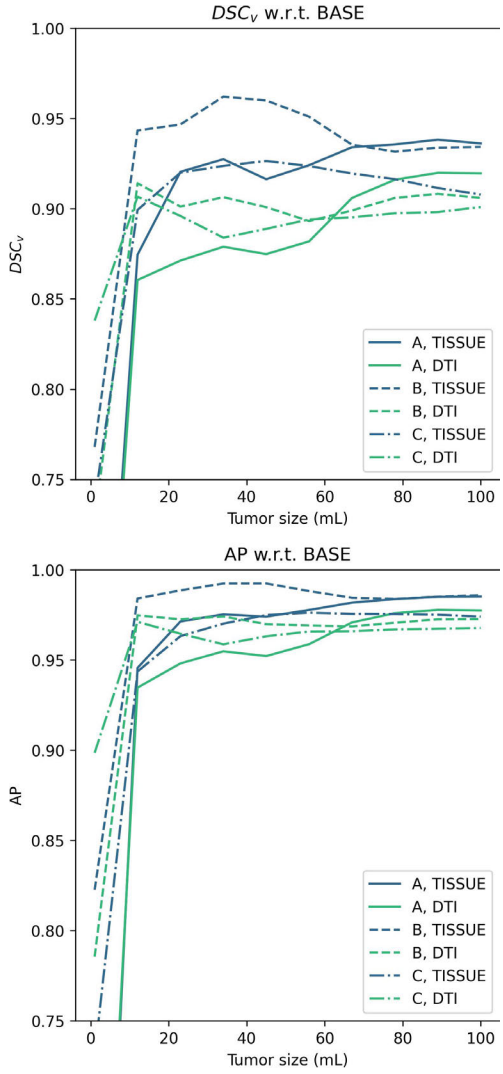


Figure 7.6: Error of the *DTI* and *TISSUE* models with respect to the simulated ground truth from the *BASE* model measured in DSC_v (top) and *AP* (bottom) versus volume, for each of the three seed locations shown above.

7.3.2 Patient data

A total of 14 patients were included. The general characteristics of these patients are listed in Table 7.2. The volumes of S'_0, S'_1 and S'_2 are visualized with the relative scan date in Fig. 7.7. Note that the time difference between the resection and the scan selected for follow-up (t_2) varied from two to ten years. Two patients presented with a small nodular contrast enhancement at time of follow-up. The overview of MR scan parameters is shown in Table 7.3.

The different tumor locations at baseline are visualized as a heatmap in the atlas space in Fig. 7.8, which shows that the tumors included in this dataset are mostly located in or near eloquent areas.

7.3.3 Image registration

In all patients, the non-rigid transformation to atlas space caused a reduction in tumor volume with respect to the affine transformation (10-47%, mean 29%). For the follow-up scans, deformations were caused mostly by the resection cavity and resulting displacement of the brain. The mean change in volume due to the b-spline transformation was +6% (-21/+83 min/max).

7.3.4 Evaluation in patient data

Fig. 7.9 shows summarized measures of the model performance in patient data. The mean DSC_v scores in the initial tumor (S'_0) were 0.72, 0.70 and 0.72 for the *BASE*, *TISSUE* and *DTI* model respectively, and 0.59, 0.60 and 0.63 for

Table 7.2: Clinical details of included patients. Time is relative to the date of resection. CE = Contrast Enhancement.

ID	Sex	Age	1p/19q	Volume (mL)			Time (days)			CE
		y	codel	S'_0	S'_1	S'_2	t_0	t_1	t_2	t_2
01	M	46	No	11	4	11	-72	2	2525	-
02	M	41	Yes	39	23	33	-84	1110	1628	Nodule
03	M	30	No	45	45	62	-74	151	890	-
04	M	27	No	48	2	11	-95	693	2967	-
05	M	45	Yes	16	5	14	-11	158	1774	-
06	F	50	Yes	37	9	27	-7	66	2929	-
07	F	48	Yes	59	0	8	-6	1	1438	-
08	M	46	No	22	2	8	-20	1	1278	-
09	M	33	No	68	41	85	-105	109	772	-
10	F	42	Yes	31	6	30	-116	566	3788	-
11	M	40	Yes	42	7	21	-83	3	2752	Nodule
12	M	42	Yes	54	25	35	-71	1	2482	-
13	M	73	Yes	20	9	14	-68	113	1401	-
14	M	45	Yes	7	6	19	-52	1	726	-

Table 7.3: Range (min, max) of scanner parameters for each sequence, preoperatively (t_0) and at follow-up (t_1 and t_2). SE=Spin Echo, GRE= Gradient Recalled Echo, IR = Inversion Recovery

	Slice thickness (mm)	Repetition time (ms)	Echo time (ms)	Inversion time (ms)	Field strength (T)	Scanning sequence
Preoperative T2w	1.5 - 5.0	5060 - 8892	82 - 140	0.0 - 0.0	1.5 - 3.0	SE
Follow-up T2w	3.0 - 5.0	4330 - 9652	89 - 142	0.0 - 0.0	1.5 - 3.0	SE
Preoperative T1w	1.5	9.1 - 10.5	2.0 - 4.2	300 - 450	1.5 - 3.0	GRE
Follow-up T1w	1.0 - 5.0	8.3 - 720.0	4.2 - 14.0	0 - 450	1.5 - 3.0	SE or GRE
Follow-up T1w+C	1.0 - 1.6	7.1 - 14.0	2.1 - 4.8	400 - 450	1.5 - 3.0	GRE
Follow-up T2w-FLAIR	1.6 - 5.0	4002 - 9502	89 - 143	1335 - 2500	1.5 - 3.0	SE / IR

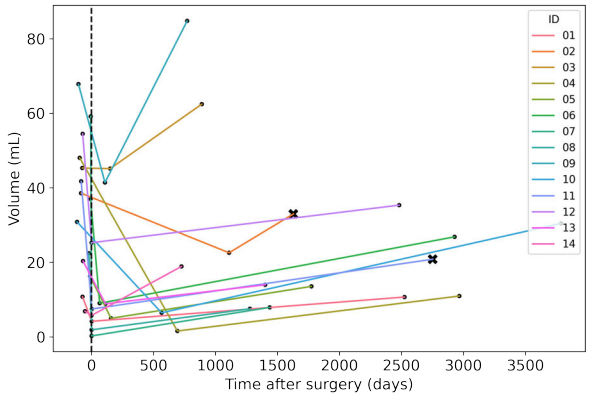


Figure 7.7: Tumor volumes measured at pre-operative imaging (S'_0), at post-operative imaging (S'_1) and at follow-up (S'_2). Volumes from the same patient are connected by a line. The time difference with respect to the resection is indicated on the x-axis. If nodular enhancement was found at follow-up, this is indicated by an enlarged 'x' marker.

the follow-up tumor (S'_2). The mean AP scores in the initial tumor (S'_0) were 0.8, 0.78 and 0.8 for the *BASE*, *TISSUE* and *DTI* model respectively, and 0.59, 0.60 and 0.63 for the follow-p tumor (S'_2).

When considering S'_0, there was no significant difference between the *BASE* and *DTI* models in either metric (DSC_v: p=0.62, AP: p=0.95), while the *TISSUE* model showed significantly lower performance in predicting the initial tumor shape compared to the *DTI* model (DSC_v: p=0.001, AP: p<0.001), but not compared to the *BASE* model (DSC_v: p=0.17, AP: p=0.08). When considering S'_2, the *DTI* model was signficantly better than the *BASE* and *TISSUE* model (p<0.001) in terms of both DSC_v and *AP*. The *TISSUE* model showed a significantly higher performance than the *BASE* model in terms of the AP (p=0.035), but not in terms of the DSC_v (p=0.43). Fig. 7.10 shows a visualization of the model result for patient 02.

The results for different values of p_b are listed in Table 7.4. Although the threshold affects the performance of different models, the general trend was unchanged. The DTI model performed significantly better in terms of AP than BASE and TISSUE for each threshold value at S'_2. At S'_0, the DTI model performed significantly better than TISSUE at each threshold value, but never significantly different from BASE.

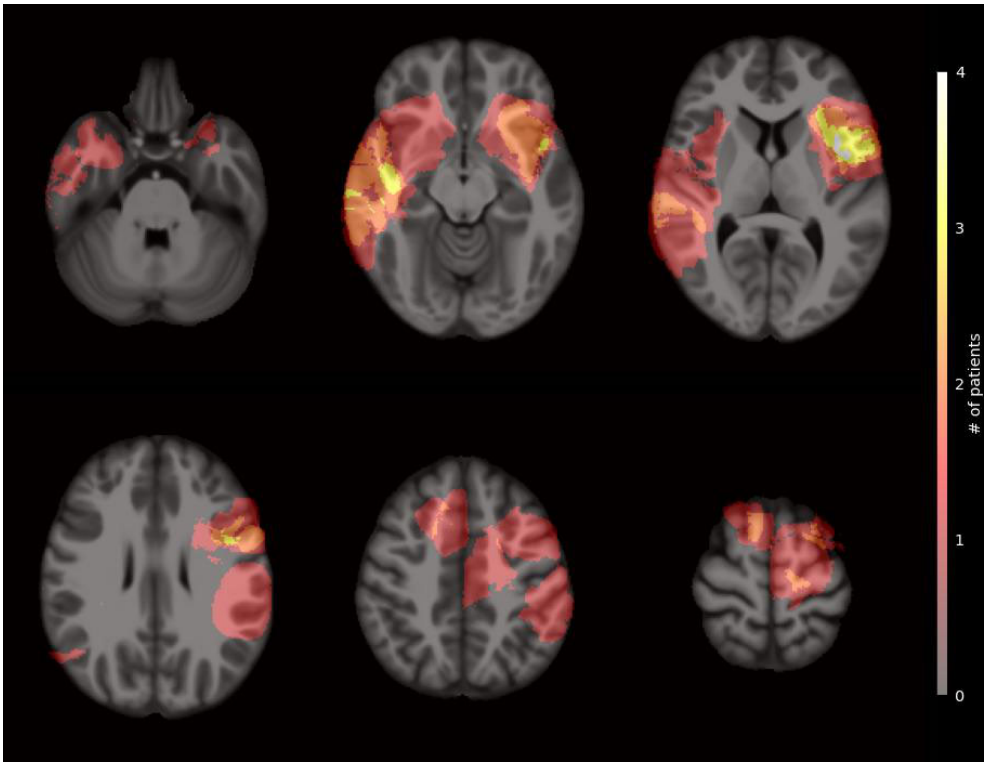


Figure 7.8: Heatmap of the registered baseline segmentations S_0' to the atlas space for all patients combined.

Table 7.4: Performance metrics of models in patient data, measured in *AP* (median [IQR]) for different values of p_b.

	S_0'		
p_b	0.5	0.8	0.9
BASE	0.83 [0.79 - 0.87]	0.80 [0.78 - 0.83]	0.80 [0.77 - 0.81]
TISSUE	0.80 [0.77 - 0.82]	0.79 [0.74 - 0.81]	0.80 [0.73 - 0.82]
DTI	0.83 [0.79 - 0.85]	0.81 [0.77 - 0.84]	0.81 [0.75 - 0.84]
	S_2'		
p_b	0.5	0.8	0.9
BASE	0.59 [0.46 - 0.69]	0.58 [0.50 - 0.70]	0.59 [0.52 - 0.71]
TISSUE	0.63 [0.52 - 0.71]	0.60 [0.54 - 0.72]	0.62 [0.54 - 0.72]
DTI	0.65 [0.55 - 0.72]	0.63 [0.57 - 0.73]	0.63 [0.55 - 0.74]

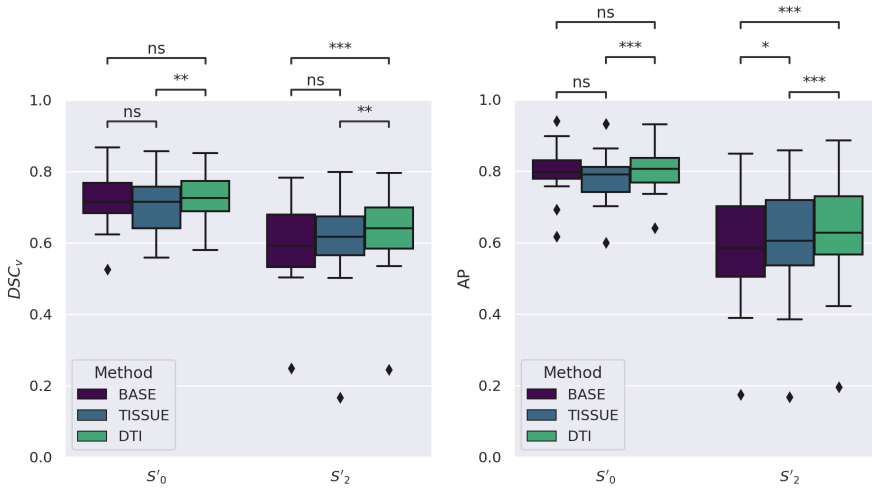


Figure 7.9: Performance metrics of models in patient data, measured in DSC_v (top) and AP (bottom). Significant differences indicated with asterisks (*: p<0.05, **: p<0.01, ***: p<0.001, ns: p>0.05)

7.4 Discussion

In this work we aimed to evaluate the predictive value of tumor growth models in LGG after resection and thereby propose a novel approach that formulates tumor growth as a ranking problem. Using this approach, we investigated whether a diffusion-proliferation model based on a tissue segmentation or DTI measurements improves the prediction with respect to a baseline of homogeneous isotropic diffusion throughout the entire brain. Special attention was given to a careful comparison that is not biased through individual parameter tuning, in order to test the general hypothesis underlying the models. From the results in patient data, we found a significant improvement in the prediction of the recurrent tumor shape when using a DTI-informed anisotropic diffusion model. Although the model informed by structural tissue segmentation improved upon the baseline of homogeneous diffusion, it was less effective than the DTI-informed model. When looking at the initial tumor, the only clear difference was found between the DTI- and tissue-informed model, while neither were significantly better or worse at fitting the intial tumor shape than the baseline. In interpreting these results we have to consider the selection of the seed location, which was at the center of the initial tumor segmentation. This choice may have caused a bias towards the baseline model with homogeneous isotropic growth. The fact that DTI information provides an improved prediction of growth is in line with existing research on the direction of glioma growth [121, 129]. In this work we have shown that this improvement

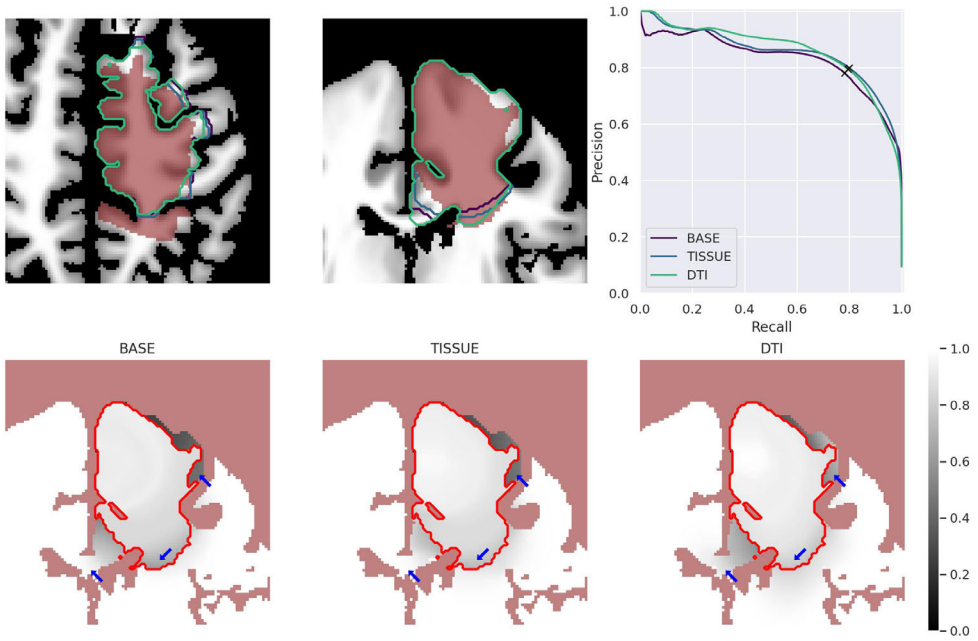


Figure 7.10: Individual model result for patient 02 at S_2'. Top left and top middle: ground-truth segmentation and model results for an axial and coronal slice. The registered ground-truth segmentation is shown in a red overlay; the model rankings thresholded at equal volume are shown as outlines. The voxels omitted from the evaluation, due to being classified as CSF in the atlas or on the follow-up T2w-FLAIR scan, are masked. Top right: precision-recall curves of the three models, with the threshold at equal volume, where precision and recall equal the Dice similarity coefficient (DSC_v), indicated by a black 'x'. Bottom: local precision and recall values for different models, as described in section 7.2.3. Three locations with notable differences between models are indicated by a blue arrow.

is apparent even without any individual parameter tuning, and with respect to a baseline model.

Due to the strict requirements for inclusion (availability of imaging, no treatment), our dataset was relatively small, limiting the detection sensitivity for subtle performance differences between models. Collection of larger datasets, if possible in a prospective setting with harmonized image acquisition protocols, would therefore be desirable, in order to enable verification of our findings and further in-depth comparative studies of tumor growth models using our proposed evaluation framework. Furthermore, for clinical relevance of these methods it is important that they can be applied to patients treated with radiotherapy. If we are able to predict the pattern of tumor growth accurately, this could help to diagnose treatment-associated changes appearing outside the expected pattern of recurrence. However, it is currently impossible to make a reliable distinction between treatment-associated changes and progressive tumor. By excluding patients treated with radiotherapy we can be certain that the ground-truth only consists of tumor growth, and is not polluted by treatment-associated changes. Therefore, in order to use this model for patients treated with radiotherapy in the future, we think it is better to exclude them in this stage of model development.

In terms of the growth model, we may expect further improvement of the prediction when we tune the model parameters to individual patients, especially if we optimize the initial location together with the other parameters. The fitting of growth model parameters, including the point of onset, has been an active research topic for many years [33]. The general consensus is that the problem is ill-posed when using a single observation, and it is therefore impossible to estimate all parameters with any degree of certainty [123]. However, the main motivation for not tuning model parameters in this work is the risk of bias. In terms of model optimization, we know that models with more degrees of freedom will be better able to fit individual cases. As a strict separation of fit and prediction is not possible in personalized growth models, considering that S_0' is at least partially included in S_2', this could lead to a bias towards more complex models. In this work, we used the information embedded in S_0' only to estimate the seed location, and by using the center of gravity the main risk would be a bias towards the baseline model with homogeneous isotropic diffusion. We expect that further improvement can be made by an optimization of the initial location and model parameters, as well as more refined models, but to present a novel model with optimal prediction is not the aim of this work.

Besides the evaluation of tumor growth models in clinical data, this work also proposes a novel problem formulation for tumor growth, which uses the AP as an evaluation metric. The main benefit of the AP over distance- or overlap-based evaluation is that it matches the spatiotemporal nature of the problem and does not rely on a specific cutoff in volume or time. Although the

advantages are qualitative, we think that this problem formulation is a useful step forward. In the patient-wise results, we do notice that the choice of metric can affect the relative performance of models. The benefit of the DTI-based model is more consistent in the AP metric than in the DSC_v metric, possibly due to its effects at the more distant infiltration that is not captured by any of the models when using a volume-based threshold. This is also reflected when testing for significant differences between models, where the AP metric shows a higher level of significance in some of the comparisons. This could indicate that the AP metric is more sensitive to differences between models, and less affected by other factors, although we should be careful in attaching strong conclusions to a difference in p-values. In general, there is no way of objectively comparing one evaluation metric to the other, as the metric should match the (clinical) purpose of the prediction. If the aim is to design, for example, a clinical target volume (CTV) for radiotherapy, a different metric could be preferred. In the design of the CTV the aim would be to keep the irradiated volume to a minimum while covering as much of the (potential) recurrent tumor as possible, so a high recall is required [109]. When using the DSC_v, the cutoff is chosen so that precision and recall are equal, while the AP does not assume a prefferred trade-off between precision and recall.

As optimization plays an important role in this field, both in parameter inversion and statistical models for growth [31, 32, 130], the AP could thus be considered as an alternative target for optimization in future work. Although the biophysical growth models in this work derive the ranking from a moving boundary, linking the ranking explicitly to the predicted time-to-invasion, the same evaluation could be applied to the output of statistical model such as a neural network. In future work, we envision that a loss term based on the AP, combined with a larger longitudinal dataset, could be used to develop solutions based on deep learning.

With most research on growth models being aimed at HGG, it is unclear how well those methods generalize to LGG. From the results in this work, we can conclude that the models informed by structural imaging and DTI do show a clear improvement with respect to a baseline of isotropic growth in this patient group. However, we must also conclude that discrepancies between models in terms of prediction are small compared to the actual error. This raises the question whether the cause of the error is in the model or due to other factors. One important factor in the error is the registration to the atlas space. By using a non-rigid deformation, we tried to capture the deformation caused by mass effect and resection. Although a ground-truth is not available for the registration problem, we can observe from visualizations that the non-rigid step improves the alignment. However, it is clear that a more accurate non-rigid atlas registration would improve the accuracy of the model and its evaluation. A quantitative estimate of the error due to misalignment is difficult to achieve, but research on the topic reports a mean error of up to 3

mm for state-of-the-art deformable models [131].

The dataset used in this research was a selection of patients that underwent surgical resection, but no radio- or chemotherapy. Although it is fair to assume that the diffusive behavior of the tumor is not affected during the surgery, so the model parameters would stay the same, the future growth pattern can be affected by the removal of tumor tissue. Including the resection in the model, e.g. by removing tissue during the simulation, could further improve the predictive performance. The decompression that occurs at resection also complicates the registration of post-operative imaging, leading to registration errors. However, with surgical resection being the recommended primary treatment for LGG [5], this is a complicating factor that cannot be avoided. It is also possible to include treatment effects in a biophysical growth model, including chemo- and radiotherapy [121], but in this work we have chosen to exclude patients who received radiotherapy due to its effect on imaging. For the evaluation of growth models across treatment, it is essential to accurately distinguish treatment effects from tumor growth. A limitation of excluding patients with further treatment is that it results in a selection bias towards patients with a relatively good prognosis. Furthermore, due to the selection of patients referred for awake surgery, there is a bias towards tumors located in or near eloquent areas.

LGG are known to remain stable for long periods of time, or grow at a slow and constant speed, but also to show sudden accelerated growth as a result of malignant transformation.In this work, we have aimed the evaluation exclusively at the spatial growth pattern, therefore allowing for variations in growth speed. However, it is likely that malignant transformation also affects the spatial growth pattern. Although two patients presented with nodular contrast enhancement at progression, indicating potential malignant transformation, we were not able to assess the tumor grade at progression due to the absence of tissue verification. Also, it is possible that patients were included who would be considered grade 4 according to the CNS-WHO 2021 classification [3] due to the homozygous deletion of CDKN2A/B, since this molecular assessment was not available in these patients. In general, it would be interesting to study the effect of differences in grade and changes in malignancy over time on the spatial growth pattern. Growth models can provide an important tool to answer such research questions as they offer a general framework to quantify growth patterns.

7.5 Conclusion

To conclude, in this work we presented a novel formulation of the tumor growth problem that fits the spatiotemporal nature of the prediction. From this formulation, the use of average precision (i.e. area under the precision-

recall curve) as an evaluation metric follows naturally. This metric was used to compare common diffusion-proliferation models in their prediction of LGG growth after surgery. By avoiding individual parameter tuning, we are able to make an unbiased comparison to a baseline model of homogeneous isotropic diffusion. In this comparison, we conclude that there is a significant improvement in the prediction of the recurrent tumor shape when using a DTI-informed anisotropic diffusion model as opposed to an isotropic diffusion model. Through a novel evaluation method and the publication of code and data, we enable a better comparison of growth models.

8

Discussion

We are all living, at most, half of a life, she thought. There was the life you lived, which consisted of the choices you made. And then, there was the other life, the one that was the things you hadn't chosen.

— Gabrielle Zevin, Tomorrow, and Tomorrow, and Tomorrow

This thesis describes methods of MR image analysis that may contribute to the management of adult-type diffuse low-grade glioma, ranging from visual assessment to biophysical modelling and machine learning. Part 1 is dedicated to the technical and clinical aspects of volume measurement, starting with a technical contribution of a deep learning method for the segmentation of glioma with missing sequences (chapter 2), followed by an application of automatic and semi-automatic volume measurement to answer clinical research questions (chapter 3) and the clinical application of volume measurement in the management of non-enhancing low-grade glioma (chapter 4). Part 2 ventures beyond volume measurement to emerging biomarkers in structural MRI. Chapter 5 concerns the T2-FLAIR mismatch sign, which can be assessed visually, in the context of recurrent IDH-mutant Astrocytoma. Chapter 6 describes a method of deep learning for the aligment of longitudinal MRI scans of the same patient. Chapter 7, finally, concerns the biophysical modelling of glioma for the purpose of growth predictions, with special attention to the method of evaluating their predictions. Here, I will recount the main findings, provide some additional context and sketch some perspectives for future research and development.

8.1 Glioma segmentation with missing sequences

Although automated segmentation is used as a research method throughout the first part of this thesis, chapter 2 is the only chapter dealing with methodological research on the topic. It deals with the problem of training a deep learning model that is robust to missing MRI sequences in the context of glioma segmentation, where the four required sequences are T2-weighted (T2w), T2w FLAIR (T2w-FLAIR), pre-contrast T1-weighted (T1w) and post-contrast T1-weighted (T1w+C). The relevance of this problem is illustrated in the following chapter, where a retrospective dataset was used and patients had to be excluded from the automated segmentation due to missing sequences. This is where the fusion networks presented in chapter 2 could have been used to include more patients. However, we did not use them, but rather the cases with missing sequences were left out or annotated using a manual or semi-automatic approach. So were my efforts in chapter 2 a waste of time? Or will we be able to improve upon these methods and develop a model that is used in practice?

To answer these questions, we must first identify why the fusion networks were not used. This was in part due to differences in the data distribution. The fusion networks in chapter 2 were trained to work on pre-operative MRI, while the research in following chapters was on post-operative scans. The publicly available model (HD-GLIO) used in the remainder of the thesis was trained on both pre- and post-operative data, and therefore better suited for the task. If the dataset used to train HD-GLIO were publicly available, however, it could

be used to train a similar, better suited fusion network. Another relevant question is whether, regardless of the technical limitations, missing one of these sequences causes a loss of information or a potential bias in the segmentation. Looking at the results in Table 2.1, using a fusion network would have been an option in the case of a missing T2w-FLAIR, T2w or T1w sequence, or even if multiple of these sequences were missing, depending on the region we are interested in. However, a more in-depth understanding of the role of each sequence is needed to properly interpret the results.

For the whole tumor region we can use either the T2w or T2w-FLAIR sequence, but do not require both. In chapter 3 we did include patients where one of those was missing, and the fusion network could have replaced the manual segmentation in those cases. Does the choice of sequences cause a bias in the results? The T2w-FLAIR provides better contrast for the non-enhancing areas, especially with respect to fluid, but it can be discerned on the T2w or even on the T1w sequence by a trained observer. The exact boundaries will present differently on each sequence, however, and therefore we can expect a slight difference in the resulting segmentation. In a comparison of clinical target volume for treatment planning, Stall et al. found a mean difference of 21% between the delineation on T2w-FLAIR and T2w sequences [132], with the T2w-FLAIR presenting a significantly higher volume in general. From a clinical perspective, there is no reason to prefer the boundaries visible on T2w-FLAIR over T2w, except that they are more easy to distinguish. In chapter 3, I would expect that the differences in volume between patients is large enough to diminish the effect, so using the T2w sequence in some cases likely did not affect the result in a meaningful way. Arguably, a fusion network could have been used to automatically segment these cases. In chapter 4, where longitudinal scans of the same patient are assessed to measure volume change, a change in sequence between longitudinal scans could cause a meaningful bias in the results and I would not recommend to apply a fusion network here.

For the enhancing core, annotated on T1w and T1w+C, the situation is different. Missing either sequence was a reason to exclude patients in chapter 3, even though chapter 2 shows that a missing T1w scan does not cause a large deterioration in performance of the fusion network. The T1w sequence is generally used as a comparison to distinguish T1w hyperintensities (e.g. due to calcification) from contrast enhancement. So either the network has learnt to distinguish T1w hyperintensities from contrast enhancement, or the T1w hyperintensities are rare enough to not cause a large deterioration in performance. Perhaps this also means that we could have included patients with a missing T1w sequence in our clinical studies, thereby accepting a small error in the volume measurement in order to include more patients. However, even if an error is small on a population level, it can have large effects in the final analysis if certain subgroups are affected. For example, a large tumor with ring-like enhancement can be measured safely without T1w MRI, but

in patients with small enhancing nodules the lack of T1w can make a large difference. Whether the T1w sequence is needed would, therefore, depend on the context in which the segmentation is being used.

A popular line of research in this context is to segment the contrast-enhancing region even without a post-contrast T1w sequence. This is the task where the fusion networks of chapter 2 are least effective, but to succeed would have the greatest practical implications. The use of contrast agent comes with an additional burden to the patient primarily due to the fact that it requires injection during each scan session, which can cause adverse reactions in a small percentage of patients [133]. Approaches to predicting the contrast enhancement from non-contrast scans vary in their problem setting, whether they are predicting the presence or volume of contrast [134], or even trying to generate the entire scan [135, 136, 137]. These approaches are more succesful in predicting the enhancing volume than our approach in chapter 2, so clearly there is some improvement to be made in the methodology. Still, the path to actual use of these methods is unclear. My main concern would be the case of malignant transformation in low-grade glioma, where the appearance of a small enhancing nodule has large diagnostic implications. A method that predicts contrast enhancement would only be reliable if it has a high sensitivity to such cases, which is not necessarily reflected by a high performance in terms of overlap, especially if the patient group is predominantly high-grade. We cannot trust an automated method if we do not measure its performance in all relevant subgroups.

In the aforementioned case we are touching on an important element to consider when discussing the use of AI in practice: do we need to trust the model, or can we verify it? When using a model to segment the tumor, something a human can also do, we can verify whether the result is correct. It is a matter of automation rather than prediction. The situation changes completely if the model is used to perform a task that humans cannot reproduce reliably, like predicting an outcome or segmenting a structure that is not completely visible. To use such a model requires trust, because we need to rely on it without verification. It is not yet clear what level of evidence is required before we trust a machine learning method in healthcare. It is reasonably straightforward to train a network to detect the contrast-enhancing region on non-contrast MRI, but it is less straightforward to use such a method in clinical practice or research, and rely on its results for treatment decisions or scientific conclusions. The black-box nature of neural networks does not help, as it is difficult to completely estimate the modes of failure. The next few years will be an interesting time for medical AI, as researchers, clinicians and legislators need to consider the real-world use of AI models that are increasingly difficult to understand or verify.

8.2 Volume measurement in clinical practice

In this section, I would like to take a step back and consider the clinical and economical context for methods of volume measurement in glioma. The current standard of practice for the assessment of glioma growth is a visual assessment combined with manual measurement of diameters. It is up to the radiologist to decide which measurements are relevant and can be performed accurately, guided by the RANO recommendations [138]. In order to provide benefit for patient care, an automated volume measurement should be relevant, accurate and cost-effective.

8.2.1 Cost-effectiveness

It is not yet clear whether the quantification of glioma volume provides measurable benefits for patient care [139], but the rationale for volumetric measurement is strong. Glioma are rarely perfect spheres, so by measuring diameters we achieve an approximation of the volume. If a radiologist would like to have a more exact measurement, if the diagnosis is not immediately obvious and growth is not easily measurable, a volume measurement can provide more certainty.

From a cost perspective, even if there is a clinical need for volume measurement the manual delineation is not cost-effective in clinical practice. In a research setting the balance between cost and benefit may be different, making a manual delineation worthwile. Automatic segmentation is mainly a solution to make volume measurement cheaper and therefore cost-effective in both clinic and research, although increased accuracy could hypothetically be achieved. The current methods do come at a cost, considering the need for powerful hardware, but with sufficient scale those costs are negligible compared to the cost of an MRI exam and assessment. The main cost is due to the effort and time of the radiologist to check the segmentation and incorporate the measurement in their assessment. This cost is still relatively high in the case of EASE (chapter 4), but it should be possible to reduce time spent to achieve a volume measurement through better system design. Ideally, it would take a radiologist only a few clicks and scrolls to achieve a volumetric assessment, and this would be incorporated in their existing workflow. To build an efficient and user-friendly system is a challenge best tackled by industry, and so the next step for this technology is to move from a research question to a business case. Translational research such as presented in chapter 4 can help to pave the way for clinical adoption.

8.2.2 Accuracy

In methodological research on glioma segmentation we are used to measure accuracy in terms of the Dice coefficient with respect to to a manual segmentation (chapter 2). When a ground-truth is not available, such as in chapter 4, we require a radiologist to rate the accuracy and accept or reject the segmentation. In both cases, it is not clear when a method is accurate enough to be used in the clinic. In the case of EASE I would argue that the clinical protocol ensures that the quality is sufficient and the limited accuracy of the segmentation method is mainly an issue for cost-effectiveness. The failed segmentations take time from the radiologist without providing any benefit. However, even with the stamp of approval given by the radiologist, we may question the accuracy of the results. Important to keep in mind in this case is the actual goal of the measurement, which is to detect and quantify subtle changes in tumor volume. To quantify subtle changes in any signal it is essential to limit the noise. In the limited time that EASE was evaluated we had one case where the radiologist rejected the volume measurements due to longitudinal inconsistencies, so it is not unlikely that this would happen more often. It is not clear whether the cause of this error is in the variation in contrast on the T2w-FLAIR scan, or due to other factors influencing the segmentation, but it is clear that this forms a risk for the quality of the diagnosis. This makes the protocol for the use of EASE, which ensures a consensus diagnosis in case of uncertainty, an important quality assurance in this stage of the technological development.

Variation in contrast between T2w-FLAIR scans form a clear risk to the quality of longitudinal segmentations, but also to longitudinal assessment by a radiologist. Subtle changes in the lesion are difficult to identify if the scanning protocol causes large changes in contrast. One way to tackle this challenge is to further homogenize scanning protocols. An improvement could be made by using quantitative MRI (qMRI), which provides a map of T1 and T2 values rather than a weighted image [134]. Those maps are potentially more robust to scanner differences than weighted images, and any weighted image can be reconstructed from the maps to provide the contrasts that radiologists use in their assessment. This comes at the cost of an increased scan time or decreased resolution, so further research into more efficient qMRI methods and their benefit to clinical interpretation is needed before quantitative MRI will be incorporated in clinical practice.

To make further improvements to segmentation quality it is essential to evaluate the segmentation in the context of non-enhancing low-grade glioma, and incorporate the longitudinal consistency in the evaluation. For that purpose it would make sense to evaluate segmentation results in terms of growth rate rather than overlap, and make sure the raters have access to longitudinal information. The inter-rater agreement is known to be poor in this context [12], so employing multiple expert raters might be needed for a proper evaluation.

It is especially interesting to know whether the estimate of growth rate can be reliable enough so that even small changes in the volume or growth rate can be used in clinical decision-making.

8.2.3 Relevance

To address the last criterion for the benefit of volume measurement is the relevance. In many cases a radiologist would not be interested in the exact volume of the lesion due to the fact that it is not needed for the clinical decision-making. If substantial growth is obvious, if an enhancing nodule has appeared or if there is a new lesion, the diagnosis of progression can already be clear. This is not a problem and does not need to be solved, because the radiologist can choose to perform the measurement as they see fit.

Another challenge to the relevance of EASE is treatment-induced abnormalities. EASE measures the lesion, but does not separate tumor-induced from treatment-induced abnormalities. To distinguish the two is still a major challenge even for radiologists, and not something that artificial intelligence is currently able to solve reliably. This distinction is therefore best left to the radiologist, who is able to combine the growth measurements provided by EASE with treatment information and advanced MRI.

It is also important to consider that a lesion may be heterogeneous, especially in later stages of the disease. One part of the lesion may be stable for a long time, possibly suspected treatment effect, while another region is showing signs of progression. In that case a radiologist can choose to consider only a part of the lesion in their diagnosis. Currently this is not possible with EASE. The technology would be even more relevant if it enables a partial measurement, either by automatically identifying distinct subregions or by enabling user interaction to perform a partial measurement. This would also enable a more nuanced diagnosis that considers different scenarios, and enable the radiologist to incorporate additional information such as advanced MRI (see sec. 8.3.3). This may not be routinely needed, but it would be beneficial to provide a neuro-radiologist with more tools especially for those patients where the interpretation of the MRI is not straightforward. An effective and user-friendly method of correcting the segmentation would further improve the accuracy and therefore cost-effectiveness by reducing the frequency with which the segmentation has to be rejected.

8.2.4 Conclusion and recommendations

Of course, the aspects of cost-effectiveness, accuracy and relevance are intertwined. Figure 8.1 outlines their relation and the main recommendations. In short, to provide a volume measurement that is cost-effective, accurate

and relevant, I would recommend the following directions for research and development:

1. Improve system design for increased efficiency and user-friendly interaction.
2. Homogenize the implementation of the T2w-FLAIR sequence and consider replacing it with quantitative MRI.
3. Incorporate longitudinal information into the method and evaluate on longitudinal changes.
4. Research effective user interactions to correct the segmentation and enable the measurement of specific parts of the lesion.

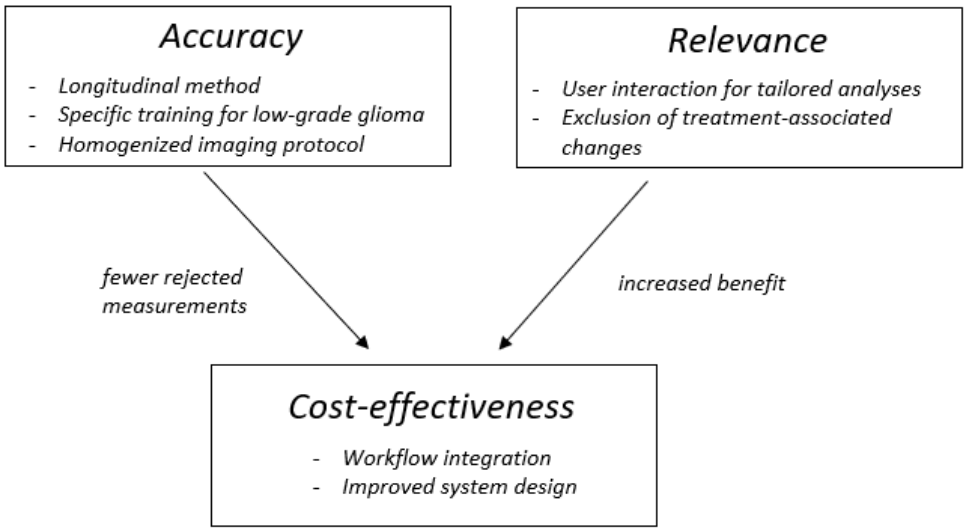


Figure 8.1: Overview of recommendations for the improved cost-effectiveness, accuracy and relevance of automated volume measurement. Improved accuracy and relevance also improve the cost-effectiveness.

8.3 Emerging biomarkers

8.3.1 T2-FLAIR mismatch

In Chapter 5 this thesis moved away from quantification and automated measurement to visual assessment. This chapter is the least technical in nature, but the closest to the topic of malignant progression in low-grade glioma. I investigated the T2-FLAIR mismatch mainly out of curiosity, because it is

such a striking feature that was often seen in the patients in GLASS-NL. There is no indication in existing literature that the mismatch is clinically meaningful beyond the initial diagnosis, but so far no research has been done in a patient group as large as GLASS-NL or including recurrent lesions. The mismatch makes these lesions appear so differently that I would intuitively expect that it has clinical relevance.

In this chapter the limitations of retrospective research become very clear, as the inclusion criteria of GLASS-NL make it likely that the results are highly influenced by selection bias. By including only patients with two resections there is definitely a bias towards patients in a relatively good condition, and possibly a reduced statistical effect of anything we measure at initial diagnosis. It would be interesting to study the T2-FLAIR mismatch in a general cohort of astrocytoma, regardless of treatment, in a longitudinal setting. My hypothesis would be that patients who show the mismatch sign, at initial presentation or recurrence, are more often (re-)resected due to (radiological) characteristics of the tumor: they often appear more well-delineated and grow in an expansive way, making them potentially easier to resect. If this holds true, it would be all the more important to verify whether the benefit to overall survival of the T2-FLAIR mismatch sign still holds.

8.3.2 Expansive and infiltrative growth

Chapter 5 also includes the visual assessment of growth patterns. Although this is presented as a sidenote to the T2-FLAIR mismatch sign, the analysis of longitudinal changes in terms of expansive or invasive growth may hold potential for future research. In one recent example of its use, Landers et al. [140] showed that the infiltrative or expansive growth pattern is different between oligodendrogliomas and astrocytomas, specifically with respect to the frontal aslant tract. Mass effect is often mentioned as a defining imaging characteristic in the assessment of glioma lesions, but it is not well studied as a prognostic marker in itself. To accurately distinguish expansive or invasive growth is difficult, because it depends on the timing of the reference scan and the interpretation of the potentially heterogeneous lesion. In chapter 5 we did not investigate the inter-rater agreement of the annotation of longitudinal growth patterns, so the reproducibility of this marker is unknown.

I expect there is more to gain with an automated quantification of growth through registration and segmentation, where the expansion of a lesion could be measured rather than visually assessed. The clinical relevance of such an assessment is largerly unknown, but there are many potential applications. One such case is that of slow-growing tumors, where a subtle expansion might be easier to measure than an increase in absolute volume. Another case is the distinction between treatment effect and tumor infiltration, especially concerning T2w-hyperintensities. The fact that a lesion is generating mass effect

could be an indicator that it is recurrent tumor, but if this mass effect is subtle it may be difficult to judge visually especially if the lesion is heterogeneous.

8.3.3 Advanced imaging

When discussing emerging methods in glioma imaging there is one group of methods that has not been discussed in this thesis, which is advanced MRI. The T2w-FLAIR and DTI are the most modern MRI methods used in this thesis, but MRI offers a plethora of other measurements such as perfusion (dynamic susceptibility contrast (DSC), dynamic contrast enhanced (DCE), or arterial spin labeling (ASL)), chemical exchange saturation transfer (CEST) and MR spectroscopy. It is beyond the scope of this discussion, and beyond my personal expertise, to discuss the potential of these methods for the diagnosis and management of glioma. However, a scientific discussion of emerging biomarkers in glioma is not complete without mentioning new imaging methods. They provide additional insight and, of special interest in the context of this thesis, may be used as early markers for malignant transformation [141, 142] and to distinguish tumor progression from treatment effect [143, 144, 145], but have not been validated extensively and adopted into the standard clinical workflow.

Longitudinal image analysis could aid in the validation of advanced imaging, as tumor development can serve as an intermediate outcome measure. If advanced MRI is able to highlight areas of the tumor that are more active than others, we would expect that heterogeneity to also be expressed in the pattern of growth. It should be possible to measure tumor growth locally and specifically, using accurate segmentation and longitudinal registration, and relate differences in tumor growth speed to imaging biomarkers on a spatial level. Although targeted biopsies and overall survival are preferable as hard evidence of malignancy, the local tumor growth can be computed non-invasively in large patient groups and with shorter follow-up. In the future, a longitudinal series of scans could be translated to a 'growth map' of expansive and invasive growth, changes in presentation such as the appearance of contrast enhancement, necrosis or cysts, that can be correlated to imaging markers at the start of the series. With enough data we could start to think of the proper statistical methods to make predictions of growth as well. Instead of predicting the probability of malignant progression of the tumor as a whole, we could predict the malignant transformation at a specific location. If the tumor is heterogeneous, then being specific about the location would increase our statistical power. More importantly, an accurate and local prediction of growth or malignant transformation would be valuable for local treatment.

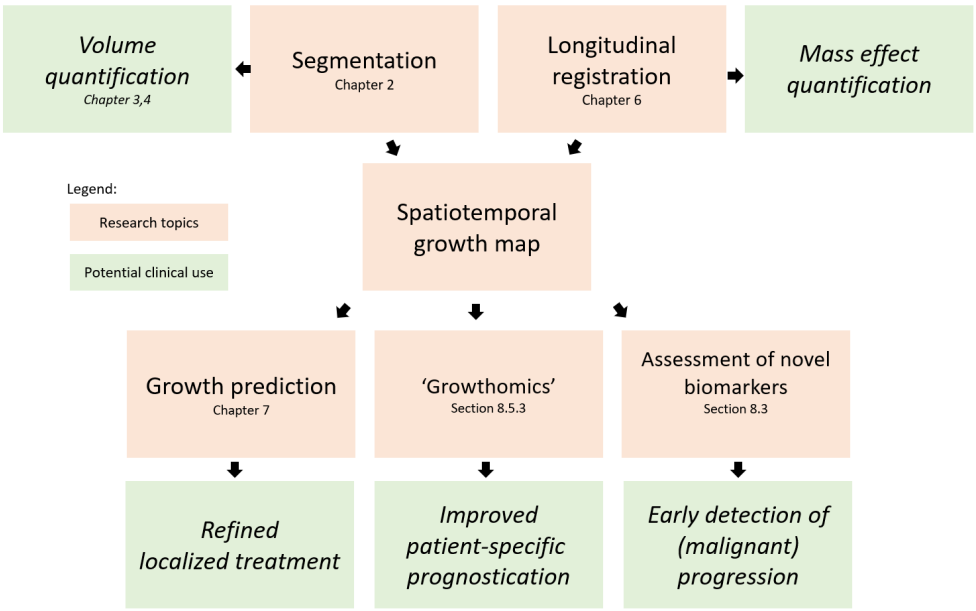


Figure 8.2: Overview of methods discussed in this thesis, including future developments described in the discussion, and their potential clinical application.

8.4 Longitudinal registration

To assess growth automatically and quantitatively requires an accurate longitudinal registration method that can cope with changing pathology, which is the topic of chapter 6. In fact, longitudinal registration is an essential step for the analysis of tumor growth in general. With tumor segmentation we can measure precisely the global change in volume, but registering the images allows us localize those changes. This allows us to answer questions like: Is the tumor growing in a specific place, or is it expanding in all directions? Is it only increasing as a mass, or is it infiltrating formerly healthy tissue? A human observer can usually answer such questions intuitively by performing an implicit registration in their mind, but we need an algorithm to perform quantitative analysis in large numbers of patients. It is the first step towards what I would call 'growth-omics' (sec. 8.5.4).

Longitudinal registration is a key technique for the quantitative analysis of glioma growth, and can be a bottleneck for many future developments as shown in Figure 8.2. Therefore, it is good to see that it is receiving increased attention with the recent BratsReg challenge [146]. This challenge uses landmarks as a means of ground truth, annotated by experts on three MR scan sessions of a glioma patient over the course of treatment. It is an exceptionally difficult problem, especially if the changes between images are large due to a surgical resection or large amount of growth, so it will be interesting to see what

types of methods will be most succesful in solving it. Solutions based on deep learning, such as chapter 6, are increasingly common in this field. As opposed to traditional registration methods, which iteratively optimize the transformation from scratch on each new set of images, these methods are trained beforehand on a large dataset to predict the optimal transformation in a single forward pass. This amortized optimization gives a large benefit in terms of inference speed, but has also been found to give relatively accurate results [147, 148].

Regardless of the method of optimization, the cost function is likely the most essential part of the registration. Especially because, unlike segmentation or classification, a ground truth for registration is difficult to capture. One way to compare to a human expert is to use landmark points, annotated in multiple images, but a ground truth for the complete registered volume is generally not available. In the case of growing glioma, the additional challenge is that there are missing correspondences due to the growing tumor and treatment. Especially a surgical resection can cause dramatic deformations that are not easily captured in the registration due to the accompanying change in appearance. The challenge of missing correspondences is generally approached by excluding the tumor volume from the cost function, forcing the algorithm to focus on the remaining healthy tissue, but it is especially those areas of tumor growth are the main areas of interest. Another essential way to guide the registration is to limit the deformation to a scale that is biologically plausible, e.g. by setting the scale of the spline functions or adding a regularization term to the cost function. We could take these ways of thinking one step further and specify exactly what deformations and missing correspondences are allowed depending on the biological context. For example, the areas of CSF may expand or collapse without limit, but they can never turn into healthy tissue or tumor. Healthy tissue, on the other hand, may show only a limited degree of compression but may be replaced by tumor tissue or, in case of a surgical resection, by CSF. Tumor growth models can inform in even more detail what changes in the tumor can be expected [111]. Such a biophysical view on the registration problem requires a very accurate segmentation of the tumor and brain tissue, so this is an example where image analysis methods are interdependent.

In the context of growing glioma, I expect an intensity-based optimization will not be sufficient to achieve good results, even with cost-function masking. Methods that explicitly encode the biophysical limitations of glioma growth and brain deformation [111] might be able to fill the gap left by the missing correspondences. Alternatively, or even additionally, explicit supervision by expert annotations could be able to guide learning-based registration methods. In any case, the availability of public longitudinal data makes a large difference. In addition to the BratsReg challenge [146], which is aimed mostly at high-grade tumor with large deformations, I hope that low-grade glioma and small

but gradual growth will receive just as much attention in future developments.

8.5 Growth prediction

Predicting the future growth of the tumor is a logical next step and can serve several purposes. The premise of growth modelling is that a tumor grows in a predictable way, and that we can capture its growth in a system of partial differential equations. After a few years of studying glioma and these growth models, I can conclude that this is only partially correct. It is undoubtedly true that the growth of glioma is predictable in the sense that it is not random. However, the hypothesis underlying the GLASS project is that glioma evolve over time in a process of mutation that is stochastic in nature. Although even random processes become predictable when they occur at scale, the tumor growth models used in chapter 7 and the majority of the existing literature are rather strict in their assumptions. One of the core assumptions of the growth model is that the properties of the tumor are constant over time. Another important assumption of the personalized, image-driven model is that the observed abnormalities on MRI form an outline, or iso-density contour, of the actual tumor cells. In low-grade glioma this may be a valid approximation, but I cannot conclude that this is always true for the regions of edema in high-grade glioma. To my knowledge, this assumption has never been verified with histopathological samples even though it has been attempted [149]. So to claim that this model is valid and can be used to support clinical decision-making seems exceptional. Exceptional claims require exceptional evidence and should be under exceptional scrutiny, which in many ways is the main purpose of chapter 7. In this section I will elaborate on some fundamental challenges in growth modelling and provide some future outlooks.

To make a perfect growth model is impossible, but to make a reasonable one might be easier than we realise. There are some general qualities of tumor growth that the diffusion-proliferation model definitely captures. The most essential being that it expands with a constant speed, which is shown to be the case in general [32, 36]. The second general quality is that the tumor cells do not cross the brain boundaries. This is also simple to model, although challenging in practice due to the fact that a good anatomical model for the complete cortical surface is not easily available. In chapter 7 I made an effort to include the boundaries of the brain correctly using the tissue segmentation, but there were still some areas where the simulated tumor could grow across the sulci. Therefore, models that capture the sulci implicitly, by means of a slower growth in the grey matter or directed growth along white matter tracts, are at an advantage simply for better capturing the brain anatomy. Arguably, we do not need a complex model to predict a constant isotropic growth. So the real question is not whether tumor growth models provide a reasonable prediction,

but whether this prediction is good enough to warrant the complexity of the model. This is especially important if we want to draw the reverse conclusion: that a good performance of the model proves the validity of its assumptions. A strong baseline model, preferably one that is easy to implement and publicly available, is absolutely essential to draw conclusions of this kind.

8.5.1 Choice of model

Recently the most interesting advance in tumor growth prediction is the advance of deep learning. Will the diffusion-proliferation model be replaced by an assumption-free data-driven deep learning model? The main strength of the diffusion-proliferation model is that it models the cell density explicitly, and therefore predicts something that is profoundly more interesting than just the visible tumor outline. This is a good reason to continue investigating and developing this model, but also a reason to be more critical of how we interpret its predictions. A disadvantage of the diffusion-proliferation model is the computational cost of running the simulations, which we can mitigate by approximating the solutions [119] or even solve them using deep learning [150].

Another disadvantage of the diffusion-proliferation model is that it provides no measure of uncertainty. Considering that a perfect prediction is likely not achievable, it would be useful to have some measure of the uncertainty. This, combined with the fact that we have gathered more and more longitudinal imaging of glioma, is a motivation to move towards a probabilistic rather than deterministic model completely. There are efforts in this direction [31, 32, 34], but there is to my knowledge no existing method that truly provides a data-driven probabilistic model. However, it is unclear what method should be used to translate a set of longitudinal scans of glioma patients to a statistical model. It is not even clear what the outcome of such a model should be, but I would envision a voxel-wise probability map outlining the likelihood for each voxel to see glioma recurrence. As I suggest briefly in chapter 7, I think the ranking-based approach could be very helpful in this regard. If I were tasked with developing a growth model, and were provided with the right training data - that is, correctly segmented and registered - I would investigate a deep learning model with the average precision loss to predict tumor growth. Note, however, that such a deep learning model would not necessarily provide us with well-calibrated probabilities or an estimation of cell densities.

What type of method is preferable also depends on the context of the implementation. The prediction of tumor growth, although it is an interesting problem, is not the end in itself. The prediction is an intermediate result of the model, that can ultimately be used to inform treatment decisions, can lead to better results in a downstream task or increase our understanding of glioma in general. In a general sense, I think we can define three down-stream purposes of growth modelling: to estimate the extent of tumor infiltration, to serve as

a reference for tumor shape, or to characterize growth. The latter I will call 'growthomics'. I will discuss each of the three in more detail, although I realize that I cannot provide an exhaustive overview of all potential applications. Table 8.1 provides a summarized overview of the purposes and qualities of growth models and how they compare between biophysical (diffusion-proliferation) models and deep learning.

8.5.2 Estimating extent of infiltration

Estimating the extent and degree of glioma infiltration is one of the most profound problems in glioma imaging. There is a potential immediate clinical use by informing treatment decisions and planning, and especially radiotherapy planning is a clear candidate. The current clinical practice is to irradiate a fixed margin around the visible tumor, the 'clinical target volume', knowing that invisible glioma infiltration can be present there [151]. A more detailed estimation of the infiltration could refine this clinical target volume [152]. Note that this application relies explicitly on the estimation of cell densities, which makes that biophysical models are the only logical candidate. However, in practice the location of recurrence with respect to the clinical target volume, referred to as the pattern of failure, is often used as a proxy to evaluate the treatment planning [153, 154]. If this implicit assumption that the pattern of failure coincides with the extent of tumor infiltration at time of radiation treatment is valid, then any method that predicts the recurrence pattern would be a viable solution. In other words, we should not disregard deep learning in

Table 8.1: Overview of model characteristics and output. Note that this is a crude simplification of the methodologies of the field, especially when it comes to deep learning, so different specific methods will have different characteristics. In this case, 'Deep Learning' refers to a method that predicts the tumor outline directly from input imaging as in Petersen et al. [31]. A '+' indicates that the model has the capability, a '-' means it does not. The '∼' indicates it is more or less fit for the purpose.

	Biophysical model	**Deep Learning**
Predicts future tumor outline	+	+
Estimates extent of infiltration	+	-
Can learn from data	-	+
Provides uncertainty estimate	-	∼
Extracts relevant features (growthomics, sec 8.5.4)	+ (model parameters)	-
Provides shape prior (sec. 8.5.3)	∼	∼

this case.

The question then remains whether growth modelling is the solution to estimating extent of infiltration. Even if a model is very effective at fitting and predicting the tumor outline, it is quite a leap of faith to assume that the estimated cell densities are also correct. With the advent of more advanced imaging techniques (see section 8.3.3) there is hope for a more direct measurement of tumor activity [155]. Growth modelling can still play a role in this problem, but I expect it would be part of the solution. A combination of advanced imaging, growth modelling and potentially even data-driven techniques such as machine learning can be a way forward [108], because each taps into a different source of information. Imaging is useful because it is direct, but potentially noisy. Data-driven methods such as machine learning can benefit from examples, although data on exact cell densities is difficult to obtain. And finally, growth modelling can be used to encode anatomical information as a sort of spatial prior probability, which I will further explore in the next section.

8.5.3 A reference for tumor shape

A growth model can be used as a reference for the expected shape or location of the recurrent tumor. This can aid, for example, methods of segmentation or registration by limiting the outcomes to reasonable tumor shapes [110]. If we take this reasoning one step further, we could imagine the model to function as a prior probability for tumor shape and growth. I could imagine that this notion of an expected tumor shape could aid to distinguish treatment effect from tumor growth by their pattern of appearance. In a way this is a reversal of tumor growth prediction. Instead of asking the model which is the most likely (future) shape of the tumor, we give the model a shape and ask: how likely is this shape?

For biophysical models this is actually a relatively hard question, because their deterministic nature makes that they only provide a single truth. Generating alternative shapes could be done by varying parameters and doing repeated simulations, but the notion of uncertainty does not come naturally to these models. For data-driven models like deep learning the question of 'Is this a likely outcome?' comes more naturally. I will not elaborate further into which specific model design could best serve this purpose, but I think it is safe to say that it is best to learn likelihoods from actual data than from assumptions.

8.5.4 Growthomics

Intuitively, I think we can all understand that tumor growth patterns can be informative, as already discussed briefly in section 8.3. Currently we have no good definition of 'growth pattern' for glioma, and we lack the tools to quantify

growth in a meaningful way beyond the change in volume. If radiomics is the extraction of meaningful parameters from imaging, then I would define 'growthomics' as the extraction of meaningful parameters from longitudinal imaging of glioma. Is the tumor growing on all sides, or in a specific direction? Is it expansive or infiltrative? These factors can be relevant to personalize treatment and management of glioma [112, 113]. Especially in the context of malignant transformation, where we would expect a change in growth dynamics, this would be an interesting avenue to explore.

The diffusion-proliferation model, in a way, is a form of growthomics as it defines tumor growth in terms of model parameters. We could imagine an unsupervised machine learning approach as well, for example in the family of autoencoders, that extract meaningful modes of variation from large high-dimensional datasets. However, due to the inherent diversity in appearance of glioma caused by location and surrounding anatomy, it is not straightforward to design a method that would extract meaningful features. Perhaps this (as of yet nonexistent) field of growthomics would benefit from handcrafted features in the same way that radiomics was developed.

8.5.5 Conclusion

To conclude this part of the discussion, growth modelling is an exciting problem with many potential clinical applications. We are currently at a cross-roads in the methodology where deep learning provides new data-driven ways of modelling growth in a field where biophysical knowledge-driven models were dominant. Both classes of methods have their advantages and disadvantages, and depending on the intended use of the model we might prioritize different aspects of the model. The most promising path forward is likely a combination of learning-based methods and biophysical knowledge. On the other hand, I think we should also not disregard classical statistics, as it could be possible to formulate tumor growth prediction in a way that we can apply statistical methods to it. Regardless of the method, it is essential to consider what the evaluation criteria are to define a succesful prediction. In chapter 7 I make the case for a ranking-based approach, which I think has potential for the training of deep learning in this context. As a last remark I would stress that the input data, consisting of the segmented and co-registered longitudinal imaging, is perhaps the most essential and most difficult aspect of glioma modelling. Aggregating sufficient longitudinal data and developing accurate spatiotemporal maps of the tumor and tissue is a first step, and the entire field would benefit if such data is publicly available.

8.6 Deep learning

Deep learning is quickly becoming an essential and extremely versatile tool in all fields of image analysis. With the flexibility that deep learning offers in terms of architecture, input data and supervision, we can barely consider it a single image analysis method. It is a combination of a universal approximator with a very effective optimization method that can solve a wide variety of different problems, from classification to registration and even growth prediction. It is not a magic bullet though. The most essential parts of the research, that of defining the problem correctly and gathering a representative dataset, are still the same irrespective of the tools being used.

Although most of the points discussed before are independent of the underlying technique, it is important to view this discussion in the context of general developments in artificial intelligence. This is a fast-moving field where the solutions in medical applications often follow closely behind the developments in computer vision and natural language processing, although medical data offers some specific challenges and opportunities. This section describes, in broad terms, some recent developments in computer vision that affect medical image analysis specifically.

8.6.1 Methodological advances

The use of convolutional neural networks, and specifically the U-Net, was just picking up steam when my research started. As I am writing this discussion, it is already starting to look outdated as vision transformers (ViT's)[156] are replacing CNN's as the dominant technique in computer vision. The main advantage of these models over CNN's is that they are better able to learn distant relationships. Where CNN's are limited in this sense by the receptive field of the convolutional layers, the self-attention mechanism employed in ViT's makes global connections early in the network. The ViT model is already becoming dominant in classification challenges [157], while the recently most succesful method in segmentation, Swin-UNETR, is a hybrid between the ViT and U-Net [158]. However, a recent publication has made seemingly groundbreaking advances in segmentation using the ViT encoder and a light-weight mask decoder that bears no resemblance to the U-Net anymore [159]. From this we can certainly conclude that the field is evolving fast in terms of model design. From a general perspective the methodology does not change much when using a ViT instead of a CNN, but there are some trends that I would like to highlight, that will impact the way research is conducted in this field.

First of all, we can observe that the computational resources required to compete in deep learning tasks is increasing. The recent 'Segment Anything' model [159] was trained on 256 A100 GPU's, which represents an investment

of millions of dollars on GPU's alone. Even with such an investment, a single training round lasts several days. This increase in cost means that models of this scale are practically out of reach for small research groups. However, one positive result is that model design is becoming increasingly modular, with large pre-trained foundation models [160] being used as a backbone for multiple different tasks. This means that instead of training a model specifically for each task and each dataset, a single pre-trained model is trained on a large and general dataset. This model can be re-used by attaching smaller subnetworks, e.g. a task-specific decoder, and finetuning on a more specific dataset. This makes the entire process of developing an AI solution more efficient, both in terms of data and resources.

Many of these foundation models are publicly available and open source, which means that smaller groups can profit from the technological advances made at large institutes and corporations. The more pessimistic view is that it is only a matter of time before the corporations developing foundation models start imposing financial barriers on their use, and stop publishing the details of their work. With the expected importance of artificial intelligence in our future society, there is an argument to be made for investment in large publicly available foundation models. For the medical field especially, which is of large public interest and encompasses a large variety of data and tasks, the development foundation models would be a logical next step. It would require a consolidation of resources and data, combined with a targeted investment in large-scale computational power. I would see this as a resource that belongs in the public domain. Commercial companies play an essential role in the development of products for the medical domain, but a publicly available foundation model could boost also the development of commercial products that build upon it.

With an increased effectiveness and generalization of models, the design of a specific deep learning solution will become an engineering effort rather than a research topic. If model development is consolidated and becomes less attractive as a research topic, this also means that more attention can be directed towards validation and reproduction, which is essential for clinical adoption but not always valued as a rewarding scientific endeavour. Research efforts can be directed towards alignment of model outcomes with the interest of patients and clinicians, and better evaluation of models in clinically relevant context. Another positive development could be that as some problems become trivial to solve, the field will move to new challenges. One such challenge is the analysis of tumor growth, which is described in this thesis in chapter 7 and discussed in section 8.5.

8.6.2 Future role in medical imaging

Deep learning definitely simplifies the field of image analysis, and with the promise of large foundation models, which reduce the requirements for task-specific data, the development of new solutions will become easier than ever. In its current form, machine learning is useful in problems where we know approximately what to look for and we are sure of the outcome that we need, but we are unsure of the exact influence of different features and their interaction. Machine learning can find those small nuances over large sets of data (e.g. diagnostics), or simply automate a task that is easy but tedious (e.g. segmentation). In longitudinal assessment we have the problem that the ground truth is often difficult to define (e.g. registration) or measure (e.g. extent of tumor infiltration). On a more fundamental level, we are not so much in need of accurate predictions as much as methods that illuminate and aggregate information. Remember that a prediction is not a treatment decision. Nevertheless, there are many ways in which deep learning can assist in longitudinal assessment as we have seen in this thesis.

Of course, one day we may develop an artificial intelligence that beats humans at any task, and we will be able to replace diagnostic experts by computers. The same is true for computer scientists. Until that day, it is my strong belief we should be working towards methods that enable the radiologist to use their intelligence rather than try to bypass them. In the field of medical image analysis the interaction with the user is often overlooked. Perhaps engineers do not like to think about user interaction because it is unpredictable, or perhaps it is too difficult to recruit radiologists for user studies. We like to have a performance measure to optimize or a p-value to compute. However, my interaction with clinical experts during this research, and my own adventures in more clinical topics, have made me realize that the way we present information and allow interactions with data can make all the difference. Furthermore, it made me realize that a deep understanding of the clinical context is essential when developing technical solutions. I have often heard the claim that clinicians, and especially radiologists, should understand artificial intelligence in order to stay relevant. It should not be too much to ask computer scientists to also understand clinicians.

8.7 Conclusion

To conclude, this thesis provides insights in the use of structural MRI and image analysis in the management of glioma. Volume measurement is an important method that is close to clinical adoption, requiring mostly an effort in validation and practical implementation. In the future, the collection and analysis of longitudinal datasets will open up new possibilities for image analysis. The longitudinal analysis of structural MRI could improve the use of growth as an

intermediate outcome measure. The characterization of growth in spatial terms is a promising avenue that requires more developments in methodology, with longitudinal image registration being an important challenge. The detection of malignant transformation is of specific interest in low-grade glioma. From a technical point of view, I would encourage a holistic view on the analysis of longitudinal data that combines structural segmentation, registration and growth modelling as each of these methodologies are strongly related. Although machine learning is an essential tool, we should not forget to consider model-based analysis and emerging biomarkers that are more easily interpretable. In general, the interaction between clinical experts and engineers is essential to focus our research efforts.

Bibliography

[1] Centers for Disease Control and Prevention, "An update on cancer deaths in the united states," *Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, Division of Cancer Prevention and Control*, 2022.

[2] Q.-W. Wang, Y.-W. Wang, Z.-L. Wang, Z.-S. Bao, T. Jiang, Z. Wang, and G. You, "Clinical and molecular characterization of incidentally discovered lower-grade gliomas with enrichment of aerobic respiration," *OncoTargets and Therapy*, vol. 13, pp. 9533–9542, Sep. 2020.

[3] 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*, vol. 23, p. 1231, 8 Aug. 2021.

[4] E. S. Murphy, C. M. Leyrer, M. Parsons, J. H. Suh, S. T. Chao, J. S. Yu, R. Kotecha, X. Jia, D. M. Peereboom, R. A. Prayson, G. H. Stevens, G. H. Barnett, M. A. Vogelbaum, and M. S. Ahluwalia, "Risk factors for malignant transformation of low-grade glioma," *International Journal of Radiation Oncology, Biology, Physics*, vol. 100, pp. 965–971, 4 Mar. 2018.

[5] M. Weller *et al.*, "EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood," *Nature Reviews Clinical Oncology 2020*, vol. 18, pp. 170–186, 3 Dec. 2020.

[6] M. J. van den Bent, J. S. Wefel, D. Schiff, M. J. Taphoorn, K. Jaeckle, L. Junck, T. Armstrong, A. Choucair, A. D. Waldman, T. Gorlia, M. Chamberlain, B. G. Baumert, M. A. Vogelbaum, D. R. Macdonald, D. A. Reardon, P. Y. Wen, S. M. Chang, and A. H. Jacobs, "Response assessment in neuro-oncology (a report of the RANO group): Assessment of outcome in trials of diffuse low-grade gliomas," *The Lancet Oncology*, vol. 12, pp. 583–593, 6 Jun. 2011.

[7] M. C. de Wit, H. G. de Bruin, W. Eijkenboom, P. A. Sillevis Smitt, and M. J. van den Bent, "Immediate post-radiotherapy changes in malignant glioma can mimic tumor progression," *Neurology*, vol. 63, pp. 535–537, 3 Aug. 2004.

[8] A. J. Kumar, N. E. Leeds, G. N. Fuller, P. van Tassel, M. H. Maor, R. E. Sawaya, and V. A. Levin, "Malignant gliomas: MR imaging spectrum of radiation therapy- and chemotherapy-induced necrosis of the brain after treatment," *Radiology*, vol. 217, pp. 377–384, 2 2000.

[9] A. M. Norris, B. M. Carrington, and N. J. Slevin, "Late radiation change in the CNS: MR imaging following gadolinium enhancement," *Clinical radiology*, vol. 52, pp. 356–362, 5 1997.

[10] M. Katsura, J. Sato, M. Akahane, T. Furuta, H. Mori, and O. Abe, "Recognizing radiation-induced changes in the central nervous system: Where to look and what to look for," *Radiographics*, vol. 41, pp. 224–248, 1 Jan. 2021.

[11] R. J. Young, A. Gupta, A. D. Shah, J. J. Graber, Z. Zhang, W. Shi, A. I. Holodny, and A. M. Omuro, "Potential utility of conventional MRI signs in diagnosing pseudoprogression in glioblastoma," *Neurology*, vol. 76, p. 1918, 22 May 2011.

[12] M. Visser, D. M. Müller, R. J. van Duijn, M. Smits, N. Verburg, E. J. Hendriks, R. J. Nabuurs, J. C. Bot, R. S. Eijgelaar, M. Witte, M. B. van Herk, F. Barkhof, P. C. de Witt Hamer, and J. C. de Munck, "Inter-rater agreement in glioma segmentations on longitudinal MRI," *NeuroImage: Clinical*, vol. 22, p. 101 727, Jan. 2019.

[13] T. T. Batchelor *et al.*, "AZD2171, a pan-VEGF receptor tyrosine kinase inhibitor, normalizes tumor vasculature and alleviates edema in glioblastoma patients," *Cancer cell*, vol. 11, pp. 83–95, 1 2007.

[14] L. C. H. Da Cruz, I. Rodriguez, R. C. Domingues, E. L. Gasparetto, and A. G. Sorensen, "Pseudoprogression and pseudoresponse: Imaging challenges in the assessment of posttreatment glioma," *American journal of neuroradiology*, vol. 32, pp. 1978–1985, 11 Dec. 2011.

[15] D. R. Macdonald, T. L. Cascino, S. C. Schold, and J. G. Cairncross, "Response criteria for phase II studies of supratentorial malignant glioma," *Journal of Clinical Oncology*, vol. 8, pp. 1277–1280, 7 Sep. 2016.

[16] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," *Journal of Digital Imaging*, vol. 32, pp. 582–596, 4 Aug. 2019.

[17] 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*, vol. 18, pp. 203–211, 2 Feb. 2021.

[18] S. Bakas *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," vol. 124, Nov. 2018. arXiv: 1811.02629.

[19] P. Kickingereder *et al.*, "Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: A multicentre, retrospective study," *The Lancet Oncology*, vol. 20, pp. 728–740, 5 May 2019.

[20] F. Kofler, C. Berger, D. Waldmannstetter, J. Lipkova, I. Ezhov, G. Tetteh, J. Kirschke, C. Zimmer, B. Wiestler, and B. H. Menze, "BraTS toolkit: Translating BraTS brain tumor segmentation algorithms into clinical and scientific practice," *Frontiers in Neuroscience*, vol. 14, p. 125, Apr. 2020.

[21] H. Peng, A. Orlichenko, R. J. Dawe, G. Agam, S. Zhang, and K. Arfanakis, "Development of a human brain diffusion tensor template," *NeuroImage*, vol. 46, pp. 967–980, 4 Jul. 2009.

[22] C. Davatzikos, R. Verma, and D. Shen, "Statistical atlases," *Handbook of Biomedical Imaging: Methodologies and Clinical Research*, pp. 125–145, Jan. 2015.

[23] L. Sun, L. Zhang, and D. Zhang, "Multi-atlas based methods in brain MR image segmentation," *Chinese Medical Sciences Journal*, vol. 34, pp. 110–119, 2 Jun. 2019.

[24] M. BachCuadra, V. Duay, and J. P. Thiran, "Atlas-based segmentation," *Handbook of Biomedical Imaging: Methodologies and Clinical Research*, pp. 221–244, Jan. 2015.

[25] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Transactions on Medical Imaging*, vol. 32, pp. 1153–1190, 7 2013.

[26] S. R. van der Voort, F. Incekara, M. M. Wijnenga, G. Kapas, M. Gardeniers, J. W. Schouten, M. P. Starmans, R. N. Tewarie, G. J. Lycklama, P. J. French, H. J. Dubbink, M. J. van den Bent, A. J. Vincent, W. J. Niessen, S. Klein, and M. Smits, "Predicting the 1p/19q codeletion status of presumed low-grade glioma with an externally validated machine learning algorithm," *Clinical Cancer Research*, vol. 25, pp. 7455–7462, 24 Dec. 2019.

[27] B. S. Jang, S. H. Jeon, I. H. Kim, and I. A. Kim, "Prediction of pseudoprogression versus progression using machine learning algorithm in glioblastoma," *Scientific Reports*, vol. 8, pp. 1–9, 1 Aug. 2018.

[28] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 7553 May 2015.

[29] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "3D deep learning for glioma classification and survival prediction from multimodal MRI," *Medical Image Analysis*, vol. 35, pp. 18–31, 2017.

[30] M. de Bruijne, "Machine learning approaches in medical image analysis: From detection to diagnosis," *Medical Image Analysis*, vol. 33, pp. 94–97, Oct. 2016.

[31] J. Petersen, P. F. Jäger, F. Isensee, S. A. A. Kohl, U. Neuberger, W. Wick, J. Debus, S. Heiland, M. Bendszus, P. Kickingereder, and K. H. Maier-Hein, "Deep probabilistic modeling of glioma growth," Jul. 2019.

[32] I. Ezhov, J. Lipkova, S. Shit, F. Kofler, N. Collomb, B. Lemasson, E. Barbier, and B. Menze, "Neural parameters estimation for brain tumor growth modeling," *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pp. 787–795, Jul. 2019.

[33] A. Mang, S. Bakas, S. Subramanian, C. Davatzikos, and G. Biros, "Integrated biophysical modeling and image analysis: Application to neuro-oncology," *Annual Reviews*, vol. 22, pp. 309–341, Jun. 2020.

[34] J. Lipkova, P. Angelikopoulos, S. Wu, E. Alberts, B. Wiestler, C. Diehl, C. Preibisch, T. Pyka, S. E. Combs, P. Hadjidoukas, K. V. Leemput, P. Koumoutsakos, J. Lowengrub, and B. Menze, "Personalized radiotherapy design for glioblastoma: Integrating mathematical tumor models, multimodal scans, and bayesian inference," *IEEE Transactions on Medical Imaging*, vol. 38, pp. 1875–1884, 8 Feb. 2019.

[35] A. Swan, T. Hillen, J. C. Bowman, and A. D. Murtha, "A patient-specific anisotropic diffusion model for brain tumour spread," *Bulletin of Mathematical Biology*, vol. 80, pp. 1259–1291, 5 May 2018.

[36] J. Jacobs, R. C. Rockne, A. J. Hawkins-Daarud, P. R. Jackson, S. K. Johnston, P. Kinahan, and K. R. Swanson, "Improved model prediction of glioma growth utilizing tissue-specific boundary effects," *Mathematical Biosciences*, vol. 312, pp. 59–66, Jun. 2019.

[37] B. H. Menze, A. Jakab, S. Bauer, *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[38] H. Van Nguyen, K. Zhou, and R. Vemulapalli, "Cross-domain synthesis of medical images using efficient location-sensitive deep network," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 677–684.

[39] J. Jerez, I. Molina, P. García-Laencina, E. Alba, N. Ribelles, M. Martín, and L. Franco, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artificial Intelligence in Medicine*, vol. 50, no. 2, pp. 105–115, 2010.

[40] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio, "HeMIS: Hetero-modal image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, Cham: Springer International Publishing, 2016, pp. 469–477.

[41] G. van Tulder and M. de Bruijne, "Representation learning for cross-modality classification," in *Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging*, Cham: Springer International Publishing, 2017, pp. 126–136.

[42] Ö. Çiçek, A. Abdulkadir, S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds., Cham: Springer International Publishing, 2016, pp. 424–432.

[43] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "No New-Net," *ArXiv e-prints*, Sep. 2018. arXiv: `1809.10483 [cs.CV]`.

[44] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, p. 170 117, 2017.

[45] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, Dec. 2014.

[46] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[47] P. Y. Wen *et al.*, "Glioblastoma in adults: a Society for Neuro-Oncology (SNO) and European Society of Neuro-Oncology (EANO) consensus review on current management and future directions," *Neuro-Oncology*, vol. 22, no. 8, pp. 1073–1113, Apr. 2020.

[48] Y. M. Li, D. Suki, K. Hess, and R. Sawaya, "The influence of maximum safe resection of glioblastoma on survival in 1229 patients: Can we do better than gross-total resection?" *Journal of Neurosurgery*, vol. 124, pp. 977–988, 4 Apr. 2016.

[49] K. L. Chaichana, I. Jusue-Torres, R. Navarro-Ramirez, S. M. Raza, M. Pascual-Gallego, A. Ibrahim, M. Hernandez-Hermann, L. Gomez, X. Ye, J. D. Weingart, A. Olivi, J. Blakeley, G. L. Gallia, M. Lim, H. Brem, and A. Quinones-Hinojosa, "Establishing percent resection and residual volume thresholds affecting survival and recurrence for patients with newly diagnosed intracranial glioblastoma," *Neuro-Oncology*, vol. 16, p. 113, 1 Jan. 2014.

[50] B. M. Ellingson *et al.*, "Validation of postoperative residual contrast-enhancing tumor volume as an independent prognostic factor for overall survival in newly diagnosed glioblastoma," *Neuro-Oncology*, vol. 20, no. 9, pp. 1240–1250, Apr. 2018.

[51] M. M. Grabowski, P. F. Recinos, A. S. Nowacki, J. L. Schroeder, L. Angelov, G. H. Barnett, and M. A. Vogelbaum, "Residual tumor volume versus extent of resection: Predictors of survival after surgery for glioblastoma: Clinical article," *Journal of Neurosurgery*, vol. 121, pp. 1115–1123, 5 Nov. 2014.

[52] P. Karschnia *et al.*, "Prognostic validation of a new classification system for extent of resection in glioblastoma: A report of the rano resect group," *Neuro-oncology*, noac193, Aug. 2022.

[53] A. Lasocki and F. Gaillard, "Non-contrast-enhancing tumor: A new frontier in glioblastoma research," *American Journal of Neuroradiology*, vol. 40, no. 5, pp. 758–765, 2019.

[54] A. Kotrotsou, A. Elakkad, J. Sun, G. A. Thomas, D. Yang, S. Abrol, W. Wei, J. S. Weinberg, A. S. Bakhtiari, M. F. Kircher, M. M. Luedi, J. F. D. Groot, R. Sawaya, A. J. Kumar, P. O. Zinn, and R. R. Colen, "Multi-center study finds postoperative residual non-enhancing component of glioblastoma as a new determinant of patient outcome," *Journal of Neuro-Oncology*, vol. 139, pp. 125–133, 2018.

[55] A. M. Molinaro *et al.*, "Association of maximal extent of resection of contrast-enhanced and non–contrast-enhanced tumor with survival within molecular subgroups of patients with newly diagnosed glioblastoma," *JAMA Oncology*, vol. 6, no. 4, pp. 495–503, Apr. 2020.

[56] F. Incekara, M. Smits, S. R. van der Voort, H. J. Dubbink, P. N. Atmodimedjo, J. M. Kros, A. J. Vincent, and M. van den Bent, "The association between the extent of glioblastoma resection and survival in light of MGMT promoter methylation in 326 patients with newly diagnosed IDH-Wildtype glioblastoma," *Frontiers in Oncology*, vol. 10, p. 1087, Jul. 2020.

[57] F. Gessler, J. D. Bernstock, A. Braczynski, S. Lescher, P. Baumgarten, P. N. Harter, M. Mittelbronn, T. Wu, V. Seifert, and C. Senft, "Surgery for glioblastoma in light of molecular markers: Impact of resection and mgmt promoter methylation in newly diagnosed IDH-1 wild-type glioblastomas," *Neurosurgery*, vol. 84, p. 190, 1 Jan. 2019.

[58] D. N. Louis, A. Perry, G. Reifenberger, A. von Deimling, D. Figarella-Branger, W. K. Cavenee, H. Ohgaki, O. D. Wiestler, P. Kleihues, and D. W. Ellison, "The 2016 world health organization classification of tumors of the central nervous system: A summary," *Acta Neuropathologica*, vol. 131, pp. 803–820, 6 Jun. 2016.

[59] L. B. Nabors, K. L. Fink, T. Mikkelsen, D. Grujicic, R. Tarnawski, D. H. Nam, M. Mazurkiewicz, M. Salacz, L. Ashby, V. Zagonel, R. Depenni, J. R. Perry, C. Hicking, M. Picard, M. E. Hegi, B. Lhermitte, and D. A. Reardon, "Two cilengitide regimens in combination with standard treatment for patients with newly diagnosed glioblastoma and unmethylated mgmt gene promoter: Results of the open-label, controlled, randomized phase II CORE study," *Neuro-Oncology*, vol. 17, pp. 708–717, 5 May 2015.

[60] R. Stupp *et al.*, "Cilengitide combined with standard treatment for patients with newly diagnosed glioblastoma with methylated mgmt promoter (CENTRIC EORTC 26071-22072 study): A multicentre, randomised, open-label, phase 3 trial," *The Lancet Oncology*, vol. 15, pp. 1100–1108, 10 Sep. 2014.

[61] S. R. van der Voort, M. Smits, and S. Klein, "DeepDicomSort: An automatic sorting algorithm for brain magnetic resonance imaging data," *Neuroinformatics*, vol. 19, pp. 159–184, 1 Jan. 2021.

[62] X. Li, P. S. Morgan, J. Ashburner, J. Smith, and C. Rorden, "The first step for neuroimaging data analysis: DICOM to NIfTI conversion," *Journal of Neuroscience Methods*, vol. 264, pp. 47–56, May 2016.

[63] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: A toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2010.

[64] D. Shamonin, "Fast parallel image registration on CPU and GPU for diagnostic classification of alzheimer's disease," *Frontiers in Neuroinformatics*, vol. 7, 2013.

[65] F. Isensee, M. Schell, I. Pflueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H.-P. Schlemmer, S. Heiland, W. Wick, M. Bendszus, K. H. Maier-Hein, and P. Kickingereder, "Automated brain extraction of multisequence MRI using artificial neural networks," *Human Brain Mapping*, vol. 40, pp. 4952–4964, 17 Dec. 2019.

[66] 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*, vol. 29, pp. 1310–1320, 6 Jun. 2010.

[67] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *NeuroImage*, vol. 31, pp. 1116–1128, 3 Jul. 2006.

[68] M. M. J. Wijnenga, P. J. French, H. J. Dubbink, W. N. M. Dinjens, P. N. Atmodimedjo, J. M. Kros, M. Smits, R. Gahrmann, G.-J. Rutten, J. B. Verheul, R. Fleischeuer, C. M. F. Dirven, A. J. P. E. Vincent, and M. J. van den Bent, "The impact of surgery in molecularly defined low-grade glioma: An integrated clinical, radiological, and molecular analysis," *Neuro-Oncology*, vol. 20, pp. 103–112, 1 Jan. 2018.

[69] P. Karschnia *et al.*, "Prognostic validation of a new classification system for extent of resection in glioblastoma: A report of the RANO resect group," *Neuro-Oncology*, Aug. 2022.

[70] M. E. Hegi, A.-C. Diserens, T. Gorlia, M.-F. Hamou, N. de Tribolet, M. Weller, J. M. Kros, J. A. Hainfellner, W. Mason, L. Mariani, J. E. Bromberg, P. Hau, R. O. Mirimanoff, J. G. Cairncross, R. C. Janzer, and R. Stupp, "MGMT gene silencing and benefit from temozolomide in glioblastoma," *New England Journal of Medicine*, vol. 352, pp. 997–1003, 10 Mar. 2005.

[71] M. M. Binabaj, A. Bahrami, S. ShahidSales, M. Joodi, M. J. Mashhad, S. M. Hassanian, K. Anvari, and A. Avan, "The prognostic value of MGMT promoter methylation in glioblastoma: A meta-analysis of clinical trials," *Journal of Cellular Physiology*, vol. 233, pp. 378–386, 1 Jan. 2018.

[72] G. Reifenberger, B. Hentschel, J. Felsberg, G. Schackert, M. Simon, O. Schnell, M. Westphal, W. Wick, T. Pietsch, M. Loeffler, and M. Weller, "Predictive impact of MGMT promoter methylation in glioblastoma of the elderly," *International Journal of Cancer*, vol. 131, pp. 1342–1350, 6 Sep. 2012.

[73] P. Yang, W. Zhang, Y. Wang, X. Peng, B. Chen, X. Qiu, G. Li, S. Li, C. Wu, K. Yao, W. Li, W. Yan, J. Li, Y. You, C. C. Chen, and T. Jiang, "IDH mutation and MGMT promoter methylation in glioblastoma: Results of a prospective registry," *Oncotarget*, vol. 6, pp. 40 896–40 906, 38 Dec. 2015.

[74] K. Yamashita, A. Hiwatashi, O. Togao, K. Kikuchi, R. Hatae, K. Yoshimoto, M. Mizoguchi, S. Suzuki, T. Yoshiura, and H. Honda, "MR imaging–based analysis of glioblastoma multiforme: Estimation of IDH1 mutation status," *American Journal of Neuroradiology*, vol. 37, pp. 58–65, 1 Jan. 2016.

[75] A. Pirzkall, C. McGue, S. Saraswathy, S. Cha, R. Liu, S. Vandenberg, K. R. Lamborn, M. S. Berger, S. M. Chang, and S. J. Nelson, "Tumor regrowth between surgery and initiation of adjuvant therapy in patients with newly diagnosed glioblastoma," *Neuro-Oncology*, vol. 11, pp. 842–852, 6 Dec. 2009.

[76] M. Weller *et al.*, "EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood," *Nature Reviews Clinical Oncology*, vol. 18, pp. 170–186, 3 Mar. 2021.

[77] A. Lasocki, F. Gaillard, M. Tacey, K. Drummond, and S. Stuckey, "Incidence and prognostic significance of non-enhancing cortical signal abnormality in glioblastoma," *Journal of Medical Imaging and Radiation Oncology*, vol. 60, pp. 66–73, 1 Feb. 2016.

[78] 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*, vol. 23, pp. 1231–1251, 8 Aug. 2021.

[79] E. Mandonnet, J.-Y. Delattre, M.-L. Tanguy, K. R. Swanson, A. F. Carpentier, H. Duffau, P. Cornu, R. V. Effenterre, E. C. Alvord, and L. Capelle, "Continuous growth of mean tumor diameter in a subset of grade II gliomas," *Annals of Neurology*, vol. 53, pp. 524–528, 4 Apr. 2003.

[80] J. Rees, H. Watt, H. R. Jäger, C. Benton, D. Tozer, P. Tofts, and A. Waldman, "Volumes and growth rates of untreated adult low-grade gliomas indicate risk of early malignant transformation," *European Journal of Radiology*, vol. 72, pp. 54–64, 1 Oct. 2009.

[81] G. B. Caseiras, O. Ciccarelli, D. R. Altmann, C. E. Benton, D. J. Tozer, P. S. Tofts, T. A. Yousry, J. Rees, A. D. Waldman, and H. R. Jäger, "Low-grade gliomas: Six-month tumor growth predicts patient outcome better than admission tumor volume, relative cerebral blood volume, and apparent diffusion coefficient," *Radiology*, vol. 253, pp. 505–512, 2 Nov. 2009.

[82] H. Duffau and L. Taillandier, "New concepts in the management of diffuse low-grade glioma: Proposal of a multistage and individualized therapeutic approach," *Neuro-Oncology*, vol. 17, pp. 332–342, 3 Mar. 2015.

[83] A. S. Jakola, K. G. Moen, O. Solheim, and K. A. Kvistad, *"no growth" on serial MRI scans of a low grade glioma?* Dec. 2013.

[84] D. Merkel, "Docker: Lightweight linux containers for consistent development and deployment," 2014.

[85] D. S. Marcus, T. R. Olsen, M. Ramaratnam, and R. L. Buckner, "The extensible neuroimaging archive toolkit: An informatics platform for managing, exploring, and sharing neuroimaging data," *Neuroinformatics*, vol. 5, pp. 11–33, 1 Mar. 2007.

[86] H. C. Achterberg, M. Koek, and W. J. Niessen, "Fastr: A workflow engine for advanced data flows in medical image analysis," *Frontiers in ICT*, vol. 3, p. 24, AUG Aug. 2016.

[87] S. Klein, M. Staring, K. Murphy, M. Viergever, and J. Pluim, "Elastix: A toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, pp. 196–205, 1 Jan. 2010.

[88] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in mri data," *IEEE Transactions on Medical Imaging*, vol. 17, pp. 87–97, 1 1998.

[89] K. Aldape *et al.*, "Glioma through the looking GLASS: Molecular evolution of diffuse gliomas and the glioma longitudinal analysis consortium," *Neuro-Oncology*, vol. 20, pp. 873–884, 7 Jun. 2018.

[90] S. H. Patel, L. M. Poisson, D. J. Brat, Y. Zhou, L. Cooper, M. Snuderl, C. Thomas, A. M. Franceschi, B. Griffith, A. E. Flanders, J. G. Golfinos, A. S. Chi, and R. Jain, "T2–flair mismatch, an imaging biomarker for idh and 1p/19q status in lower-grade gliomas: A tcga/tcia project," *Clinical Cancer Research*, vol. 23, pp. 6078–6086, 20 Oct. 2017.

[91] A. Corell, S. F. Vega, N. Hoefling, L. Carstam, A. Smits, T. O. Bontell, I. M. Björkman-Burtscher, H. Carén, and A. S. Jakola, "The clinical significance of the t2-flair mismatch sign in grade ii and iii gliomas: A population-based study," *BMC Cancer*, vol. 20, pp. 1–10, 1 May 2020.

[92] M. P. Broen, M. Smits, M. M. Wijnenga, H. J. Dubbink, M. H. Anten, O. E. Schijns, J. Beckervordersandforth, A. A. Postma, and M. J. V. den Bent, "The t2-flair mismatch sign as an imaging marker for non-enhancing idh-mutant, 1p/19q-intact lower-grade glioma: A validation study," *Neuro-Oncology*, vol. 20, pp. 1393–1399, 10 Sep. 2018.

[93] S. Deguchi, T. Oishi, K. Mitsuya, Y. Kakuda, M. Endo, T. Sugino, and N. Hayashi, "Clinicopathological analysis of t2-flair mismatch sign in lower-grade gliomas," *Scientific Reports 2020 10:1*, vol. 10, pp. 1–6, 1 Jun. 2020.

[94] S. Yamashita, H. Takeshima, Y. Kadota, M. Azuma, T. Fukushima, N. Ogasawara, T. Kawano, M. Tamura, J. Muta, K. Saito, G. Takeishi, A. Mizuguchi, T. Watanabe, H. Ohta, and K. Yokogami, "T2-fluid-attenuated inversion recovery mismatch sign in lower grade gliomas: Correlation with pathological and molecular findings," *Brain tumor pathology*, vol. 39, pp. 88–98, 2 Apr. 2022.

[95] R. Sawaya, M. Hammoud, D. Schoppa, K. R. Hess, S. Z. Wu, W. M. Shi, and D. M. Wildrick, "Neurosurgical outcomes in a modern series of 400 craniotomies for treatment of parenchymal tumors," *Neurosurgery*, vol. 42, pp. 1044–1056, 5 May 1998.

[96] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3d active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *NeuroImage*, vol. 31, pp. 1116–1128, 3 Jul. 2006.

[97] M. Nicolasjilwan *et al.*, "Addition of mr imaging features and genetic biomarkers strengthens glioblastoma survival prediction in tcga patients," *Journal of Neuroradiology*, vol. 42, pp. 212–221, 4 Jul. 2015.

[98] R. Jain, D. R. Johnson, S. H. Patel, M. Castillo, M. Smits, M. J. van den Bent, A. S. Chi, D. P. Cahill, and C. Author, ""real world" use of a highly reliable imaging sign: "t2-flair mismatch" for identification of idh mutant astrocytomas," *Neuro-Oncology*, vol. 22, pp. 936–943, 7 2020.

[99] M. Kinoshita, H. Arita, M. Takahashi, T. Uda, J. Fukai, K. Ishibashi, N. Kijima, R. Hirayama, M. Sakai, A. Arisawa, H. Takahashi, K. Nakanishi, N. Kagawa, K. Ichimura, Y. Kanemura, Y. Narita, and H. Kishima, "Impact of inversion time for flair acquisition on the t2-flair mismatch detectability for idh-mutant, non-codel astrocytomas," *Frontiers in Oncology*, vol. 10, p. 3050, Jan. 2021.

[100] M. Polfliet, S. Klein, W. Huizinga, M. M. Paulides, W. J. Niessen, and J. Vandemeulebroucke, "Intrasubject multimodal groupwise registration with the conditional template entropy," *Medical Image Analysis*, vol. 46, pp. 15–25, May 2018.

[101] B. Li, W. J. Niessen, S. Klein, M. A. Ikram, M. W. Vernooij, and E. E. Bron, "Learning unbiased group-wise registration (lugr) and joint segmentation: Evaluation on longitudinal diffusion mri," B. A. Landman and I. Išgum, Eds., SPIE, Feb. 2021, p. 14.

[102] W. R. Vallentgoed, J. M. Niers, M. J. van den Bent, M. C. M. Kouwenhoven, J. M. Kros, I. Martin, H. F. van Thuijl, M. A. van de Wiel, P. Wesseling, and P. J. French, "Methylation analysis of matched primary and recurrent idhmt astrocytoma; an update from the glioma longitudinal analysis nl (glass-nl) consortium," *Neuro-Oncology*, vol. 23, pp. ii7–ii8, Sep. 2021.

[103] Y. Zhang, X. Wu, H. M. Gach, H. Li, and D. Yang, "Groupregnet: A groupwise one-shot deep learning-based 4d image registration method," *Physics in Medicine & Biology*, vol. 66, p. 045 030, 4 Feb. 2021.

[104] W. Huizinga, D. Poot, J.-M. Guyader, R. Klaassen, B. Coolen, M. van Kranenburg, R. van Geuns, A. Uitterdijk, M. Polfliet, J. Vandemeulebroucke, A. Leemans, W. Niessen, and S. Klein, "Pca-based groupwise image registration for quantitative mri," *Medical Image Analysis*, vol. 29, pp. 65–78, Apr. 2016.

[105] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, and S. Ourselin, "Fast free-form deformation using graphics processing units," *Computer Methods and Programs in Biomedicine*, vol. 98, pp. 278–284, 3 Jun. 2010.

[106] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ants similarity metric performance in brain image registration," *NeuroImage*, vol. 54, pp. 2033–2044, 3 Feb. 2011.

[107] N. Lamprinou, N. Nikolikos, and E. Z. Psarakis, "Groupwise image alignment via self quotient images," *Sensors*, vol. 20, p. 2325, 8 Apr. 2020.

[108] N. Gaw, A. Hawkins-Daarud, L. S. Hu, H. Yoon, L. Wang, Y. Xu, P. R. Jackson, K. W. Singleton, L. C. Baxter, J. Eschbacher, A. Gonzales, A. Nespodzany, K. Smith, P. Nakaji, J. R. Mitchell, T. Wu, K. R. Swanson, and J. Li, "Integration of machine learning and mechanistic models accurately predicts variation in cell density of glioblastoma using multiparametric MRI," *Scientific Reports*, vol. 9, no. 1, pp. 1–9, 2019.

[109] A. K. Trip, M. B. Jensen, J. F. Kallehauge, and S. Lukacova, "Individualizing the radiotherapy target volume for glioblastoma using DTI-MRI: A phase 0 study on coverage of recurrences," *Acta Oncologica*, vol. 58, pp. 1532–1535, 10 Jul. 2019.

[110] S. Bakas, K. Zeng, A. Sotiras, S. Rathore, H. Akbari, B. Gaonkar, M. Rozycki, S. Pati, and C. Davatzikos, "GLISTRboost: Combining multimodal MRI segmentation, registration, and biophysical tumor growth modeling with gradient boosting machines for glioma segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic*

Brain Injuries, ser. Lecture Notes in Computer Science, vol. 9556, 2016, pp. 144–155.

[111] D. Kwon, M. Niethammer, H. Akbari, M. Bilello, and K. M. Pohl, "PORTR: Pre-operative and post-recurrence brain tumor registration," *IEEE Trans Med Imaging*, vol. 33, no. 3, pp. 651–667, 2014.

[112] F. Raman, E. Scribner, O. Saut, C. Wenger, T. Colin, and H. M. Fathallah-Shaykh, "Computational trials: Unraveling motility phenotypes, progression patterns, and treatment options for glioblastoma multiforme," *PLOS ONE*, vol. 11, no. 1, e0146617, Jan. 2016.

[113] E. Mandonnet, J. Pallud, O. Clatz, L. Taillandier, E. Konukoglu, H. Duffau, and L. Capelle, "Computational modeling of the WHO grade II glioma dynamics: Principles and applications to management paradigm," *Neurosurgical review*, vol. 31, no. 3, pp. 263–269, 2008.

[114] M. L. Neal, A. D. Trister, S. Ahn, A. Baldock, C. A. Bridge, L. Guyman, J. Lange, R. Sodt, T. Cloke, A. Lai, T. F. Cloughesy, M. M. Mrugala, J. K. Rockhill, R. C. Rockne, and K. R. Swanson, "Response classification based on a minimal model of glioblastoma growth is prognostic for clinical outcomes and distinguishes progression from pseudoprogression," *Cancer Research*, vol. 73, no. 10, pp. 2976–2986, May 2013.

[115] K. M. Field, M. A. Rosenthal, M. Khasraw, K. Sawkins, and A. K. Nowak, "Evolving management of low grade glioma: No consensus amongst treating clinicians," *Journal of Clinical Neuroscience*, vol. 23, pp. 81–87, Jan. 2016.

[116] L. S. Hu, A. Hawkins-Daarud, L. Wang, J. Li, and K. R. Swanson, "Imaging of intratumoral heterogeneity in high-grade glioma," *Cancer Letters*, vol. 477, pp. 97–106, 2020.

[117] D. L. Silbergeld and M. R. Chicoine, "Isolation and characterization of human malignant glioma cells from histologically normal brain," *Journal of Neurosurgery*, vol. 86, no. 3, pp. 525–531, 1997.

[118] S. Jbabdi, E. Mandonnet, H. Duffau, L. Capelle, K. R. Swanson, M. Pélégrini-Issac, R. Guillevin, and H. Benali, "Simulation of anisotropic growth of low-grade gliomas using diffusion tensor imaging," *Magnetic Resonance in Medicine*, vol. 54, no. 3, pp. 616–624, 2005.

[119] E. Konukoglu, O. Clatz, B. H. Menze, B. Stieltjes, M. A. Weber, E. Mandonnet, H. Delingette, and N. Ayache, "Image guided personalization of reaction-diffusion type tumor growth models using modified anisotropic eikonal equations," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 77–95, 2010.

[120] A. Elazab, C. Wang, S. J. S. Gardezi, H. Bai, Q. Hu, T. Wang, C. Chang, and B. Lei, "GP-GAN: Brain tumor growth prediction using stacked 3d generative adversarial networks from longitudinal mr images," *Neural Networks*, vol. 132, pp. 321–332, 2020.

[121] A. Elazab, H. Bai, Y. M. Abdulazeem, T. Abdelhamid, S. Zhou, K. K. Wong, and Q. Hu, "Post-Surgery Glioma Growth Modeling from Magnetic Resonance Images for Patients with Treatment," *Scientific Reports*, vol. 7, no. 1, 2017.

[122] K. A. van Garderen, S. R. van der Voort, M. M. J. Wijnenga, F. Incekara, G. Kapsas, R. Gahrmann, A. Alafandi, M. Smits, and S. Klein, "Evaluating glioma growth predictions as a forward ranking problem," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, ser. Lecture Notes in Computer Science, vol. 12962, 2021, pp. 100–111.

[123] A. Gholami, A. Mang, and G. Biros, "An inverse problem formulation for parameter estimation of a reaction-diffusion model of low grade gliomas," *Journal of mathematical biology*, vol. 72, no. 1-2, pp. 409–433, 2016.

[124] M. S. Alnaes, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wells, "The FEniCS Project Version 1.5," vol. 3, no. 100, pp. 9–23, 2015.

[125] S. Zhang and K. Arfanakis, "Evaluation of standardized and study-specific diffusion tensor imaging templates of the adult human brain: Template characteristics, spatial normalization accuracy, and detection of small inter-group FA differences," *NeuroImage*, vol. 172, pp. 40–50, 2018.

[126] X. Qi and K. Arfanakis, "Regionconnect: Rapidly extracting standardized brain connectivity information in voxel-wise neuroimaging studies," *NeuroImage*, vol. 225, 2021.

[127] D. P. Shamonin, E. E. Bron, B. P. Lelieveldt, M. Smits, S. Klein, and M. Staring, "Fast parallel image registration on CPU and GPU for diagnostic classification of alzheimer's disease," *Frontiers in Neuroinformatics*, vol. 7, p. 50, 2014.

[128] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm," *IEEE Transactions on Medical Imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[129] M. Esmaeili, A. L. Stensjøen, E. M. Berntsen, O. Solheim, and I. Reinertsen, "The direction of tumour growth in glioblastoma patients," *Scientific Reports*, vol. 8, no. 1, pp. 1–6, 2018.

[130] J. C. Peeken, M. Molina-Romero, C. Diehl, B. H. Menze, C. Straube, B. Meyer, C. Zimmer, B. Wiestler, and S. E. Combs, "Deep learning derived tumor infiltration maps for personalized target definition in glioblastoma radiotherapy," *Radiotherapy and Oncology*, vol. 138, pp. 166–172, 2019.

[131] X. Han, S. Bakas, R. Kwitt, S. Aylward, H. Akbari, M. Bilello, C. Davatzikos, and M. Niethammer, "Patient-specific registration of pre-operative and post-recurrence brain tumor MRI scans," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, ser. Lecture Notes in Computer Science, vol. 11383, 2019, p. 105.

[132] B. Stall, L. Zach, H. Ning, J. Ondos, B. Arora, U. Shankavaram, R. W. Miller, D. Citrin, and K. Camphausen, "Comparison of T2 and FLAIR imaging for target delineation in high grade gliomas," *Radiation Oncology*, vol. 5, p. 5, Jan. 2010.

[133] V. M. Runge, "Safety of the gadolinium-based contrast agents for magnetic resonance imaging, focusing in part on their accumulation in the brain and especially the dentate nucleus," *Investigative radiology*, vol. 51, pp. 273–279, 5 2016.

[134] L. Nunez-Gonzalez, K. A. van Garderen, M. Smits, J. Jaspers, A. M. Romero, D. H. Poot, and J. A. Hernandez-Tamames, "Pre-contrast MAGiC in treated gliomas: A pilot study of quantitative MRI," *Scientific Reports*, vol. 12, p. 21 820, 1 Dec. 2022.

[135] J. Kleesiek, J. N. Morshuis, F. Isensee, K. Deike-Hofmann, D. Paech, P. Kickingereder, U. Köthe, C. Rother, M. Forsting, W. Wick, M. Bendszus, H. P. Schlemmer, and A. Radbruch, "Can virtual contrast enhancement in brain MRI replace gadolinium?: A feasibility study," *Investigative Radiology*, vol. 54, pp. 653–660, 10 Oct. 2019.

[136] C. J. Preetha *et al.*, "Deep-learning-based synthesis of post-contrast T1-weighted MRI for tumour response assessment in neuro-oncology: A multicentre, retrospective cohort study," *The Lancet Digital Health*, vol. 3, e784–e794, 12 Dec. 2021.

[137] C. Liu, N. Zhu, H. Sun, J. Zhang, X. Feng, S. Gjerswold-Selleck, D. Sikka, X. Zhu, X. Liu, T. Nuriel, H. J. Wei, C. C. Wu, J. T. Vaughan, A. F. Laine, F. A. Provenzano, S. A. Small, and J. Guo, "Deep learning of MRI contrast enhancement for mapping cerebral blood volume from single-modal non-contrast scans of aging and alzheimer's disease brains," *Frontiers in Aging Neuroscience*, vol. 14, p. 893, Aug. 2022.

[138] 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. V. D. Bent, and S. M.

Chang, "Updated response assessment criteria for high-grade gliomas: Response assessment in neuro-oncology working group," *Journal of Clinical Oncology*, vol. 28, pp. 1963–1972, 11 Apr. 2010.

[139] R. Gahrmann, M. van den Bent, B. van der Holt, R. M. Vernhout, W. Taal, M. Vos, J. C. de Groot, L. V. Beerepoot, J. Buter, Z. H. Flach, M. Hanse, B. Jasperse, and M. Smits, "Comparison of 2d (RANO) and volumetric methods for assessment of recurrent glioblastoma treated with bevacizumab-a report from the BELOB trial," *Neuro-Oncology*, vol. 19, pp. 853–861, 6 Jun. 2017.

[140] M. J. F. Landers, H. B. Brouwers, G. J. Kortman, I. Boukrab, W. D. Baene, and G. J. M. Rutten, "Oligodendrogliomas tend to infiltrate the frontal aslant tract, whereas astrocytomas tend to displace it," *Neuroradiology*, vol. 65, pp. 1127–1131, 7 Jul. 2023.

[141] C. Hlaihel, L. Guilloton, J. Guyotat, N. Streichenberger, J. Honnorat, and F. Cotton, "Predictive value of multimodality MRI using conventional, perfusion, and spectroscopy MR in anaplastic transformation of low-grade oligodendrogliomas," *Journal of neuro-oncology*, vol. 97, pp. 73–80, 1 2010.

[142] N. Danchaivijitr, A. D. Waldman, D. J. Tozer, C. E. Benton, G. B. Caseiras, P. S. Tofts, J. H. Rees, and H. R. Jäger, "Low-grade gliomas: Do changes in rCBV measurements at longitudinal perfusion-weighted mr imaging predict malignant transformation?" *Radiology*, vol. 247, pp. 170–178, 1 Apr. 2008.

[143] S. Wang, M. Martinez-Lage, Y. Sakai, S. Chawla, S. G. Kim, M. Alonso-Basanta, R. A. Lustig, S. Brem, S. Mohan, R. L. Wolf, A. Desai, and H. Poptani, "Differentiating tumor progression from pseudoprogression in patients with glioblastomas using diffusion tensor imaging and dynamic susceptibility contrast MRI," *AJNR. American journal of neuroradiology*, vol. 37, pp. 28–36, 1 Jan. 2016.

[144] D. Aquino, A. Gioppo, G. Finocchiaro, M. G. Bruzzone, and V. Cuccarini, "MRI in glioma immunotherapy: Evidence, pitfalls, and perspectives," *Journal of Immunology Research*, vol. 2017, 2017.

[145] N. Verma, M. C. Cowperthwaite, M. G. Burnett, and M. K. Markey, "Differentiating tumor recurrence from treatment necrosis: A review of neuro-oncologic imaging strategies," *Neuro-oncology*, vol. 15, pp. 515–534, 5 May 2013.

[146] B. Baheti, D. Waldmannstetter, S. Chakrabarty, H. Akbari, M. Bilello, B. Wiestler, J. Schwarting, E. Calabrese, J. Rudie, S. Abidi, M. Mousa, J. Villanueva-Meyer, D. S. Marcus, C. Davatzikos, A. Sotiras, B. Menze,

and S. Bakas, "The brain tumor sequence registration challenge: Establishing correspondence between pre-operative and follow-up MRI scans of diffuse glioma patients," Dec. 2021. arXiv: 2112.06979v1.

[147] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "VoxelMorph: A learning framework for deformable medical image registration," *IEEE Transactions on Medical Imaging*, vol. 38, pp. 1788–1800, 8 Aug. 2019.

[148] X. Chen, A. Diaz-Pinto, N. Ravikumar, and A. Frangi, "Deep learning in medical image registration," *Progress in Biomedical Engineering*, Dec. 2020.

[149] C. Martens, L. Lebrun, C. Decaestecker, T. Vandamme, Y.-R. V. Eycke, A. Rovai, T. Metens, O. Debeir, S. Goldman, I. Salmon, and G. van Simaeys, "Initial condition assessment for reaction-diffusion glioma growth models: A translational MRI-Histology (in)validation study," *Tomography*, vol. 7, 4 Feb. 2021.

[150] I. Ezhov, T. Mot, S. Shit, J. Lipkova, J. C. Paetzold, F. Kofler, C. Pellegrini, M. Kollovieh, F. Navarro, H. Li, M. Metz, B. Wiestler, and B. Menze, "Geometry-aware neural solver for fast bayesian calibration of brain tumor models," *IEEE Transactions on Medical Imaging*, vol. 41, pp. 1269–1278, 5 May 2022.

[151] P. Farace, M. G. Giri, G. Meliadò, D. Amelio, L. Widesott, G. K. Ricciardi, S. Dall'Oglio, A. Rizzotti, A. Sbarbati, A. Beltramello, S. Maluta, and M. Amichetti, "Clinical target volume delineation in glioblastomas: Pre-operative versus post-operative/pre-radiotherapy MRI," *The British Journal of Radiology*, vol. 84, pp. 271–278, 999 Mar. 2011.

[152] Y. Huang, H. Ding, M. Luo, Z. Li, S. Li, C. Xie, and Y. Zhong, "A new approach to delineating clinical target volume for radiotherapy of glioblastoma: A phase II trial," *Frontiers in Oncology*, vol. 12, Oct. 2022.

[153] G. Minniti, P. Tini, M. Giraffa, L. Capone, G. Raza, I. Russo, E. Cinelli, P. C. Gentile, A. Bozzao, S. Paolini, and V. Esposito, "Feasibility of clinical target volume reduction for glioblastoma treated with standard chemoradiation based on patterns of failure analysis," *Radiotherapy and Oncology*, vol. 181, p. 109 435, Apr. 2023.

[154] G. Minniti, D. Amelio, M. Amichetti, M. Salvati, R. Muni, A. Bozzao, G. Lanzetta, S. Scarpino, A. Arcella, and R. M. Enrici, "Patterns of failure and comparison of different target volume delineations in patients with glioblastoma treated with conformal radiotherapy plus concomitant and adjuvant temozolomide," *Radiotherapy and Oncology*, vol. 97, pp. 377–381, 3 Dec. 2010.

[155] H. Hyare, S. Thust, and J. Rees, "Advanced MRI techniques in the monitoring of treatment of gliomas," *Current treatment options in neurology*, vol. 19, 3 Mar. 2017.

[156] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR 2021 - 9th International Conference on Learning Representations*, Oct. 2020.

[157] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *Medical Image Analysis*, vol. 88, p. 102 802, Aug. 2023.

[158] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," Jan. 2022.

[159] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," Apr. 2023.

[160] R. Bommasani *et al.*, "On the opportunities and risks of foundation models," Aug. 2022. arXiv: 2108.07258.

Summary

Every year approximately one thousand people in the Netherlands are diagnosed with diffuse glioma, a type of infiltrative brain tumor that originates from the glial cells. There is no curative treatment available for adults diagnosed with a diffuse glioma, although surgical resection, radiotherapy and chemotherapy are used to improve prognosis and decrease symptoms. Low-grade glioma can remain stable for long periods of time before, inevitably, malignant progression occurs. The radiological assessment of glioma through magnetic resonance imaging (MRI) plays an important role in the management of glioma. In this thesis I explore the role of quantitative measurements, emerging imaging markers and predictive modelling in the management of glioma. These methods can aid the radiologist to predict the timing, location and severity of tumor progression, to ultimately improve the quality of life for glioma patients.

Chapter 1 introduces the main topics of this thesis. Specifically, it describes the categorization of glioma in types and grades, the general practices in disease management, the general presentation of glioma on MR imaging and the most important methodologies in image analysis.

In **chapter 2** I describe a method for the segmentation of glioma with missing imaging modalities. This methodology takes into account missing modalities in the design and training of neural networks, to ensure that they are capable of providing the best possible prediction even when multiple images are not available. The proposed network combines three modifications to the standard 3D UNet architecture: a training scheme with dropout of modalities, a multi-pathway architecture with fusion layer in the final stage, and the separate pre-training of these pathways.

Chapter 3 concerns the the association between post-operative tumor burden and overall survival in patients with newly diagnosed glioblastoma. This is evaluated in a cohort treated with radio-/chemotherapy with temozolomide after resection, and the analysis is adjusted for the prognostic value of O6-methylguanine DNA methyltransferase (MGMT) promoter methylation. The tumor burden is assessed using automated segmentation followed by manual correction where needed. We found that pre-radiotherapy contrast enhancing volume was strongly associated with overall survival in patients receiving radio-/chemotherapy for newly diagnosed glioblastoma stratified by MGMT promoter methylation status.

Chapter 4 describes a clinical implementation of automated tumor volume measurement (EASE) for low-grade glioma. Besides the technical implementation of the algorithm, this chapter describes a clinical protocol for the use of its results in the diagnosis of progression. Additionally, to ensure patient

safety and quality of care, protocols were established for the usage of volume measurements in clinical diagnosis and for future updates to the algorithm. It was applied to a total of 55 patients, and in 36 of those the radiologist was able to make a volume-based diagnosis using three successful consecutive measurements from EASE. In all cases the volume-based diagnosis was in line with the conventional visual diagnosis.

Chapter 5 concerns the T2-FLAIR mismatch sign, which is defined by signal loss of the T2-weighted hyperintense area with FLAIR (Fluid-Attenuated Inversion Recovery) on MRI. It is a highly specific diagnostic marker for IDH-mutant astrocytoma, and is postulated to be caused by intercellular microcystic change in the tumor tissue. The aim of this chapter is to determine whether the T2-FLAIR mismatch phenomenon has any prognostic value beyond initial non-invasive molecular diagnosis. The relation between the T2-FLAIR mismatch sign and tumor grade, microcystic change, overall survival and other clinical parameters was investigated both at first and second resection in the GLASS-NL cohort. We found that T2-FLAIR mismatch in IDH-mutant astrocytomas is correlated with microcystic change in the tumor tissue, favorable prognosis and grade 2 tumors at time of second resection.

Chapter 6 describes a method for the longitudinal registration of MR imaging with glioma. By performing a group-wise deep learning-based registration, we align consecutive images of the same patient. This enables a more detailed analysis of the changes over time, such as whether the glioma grows expansively or invasively.

Chapter 7 concerns tumor growth models, which have the potential to model and predict the spatiotemporal evolution of glioma in individual patients. In this chapter, we propose to formulate the problem of tumor growth as a ranking problem, as opposed to a segmentation problem, and use the average precision (AP) as a performance metric. Using the AP metric, we evaluate diffusion-proliferation models informed by structural MRI and DTI, after tumor resection. We conclude there is a significant improvement in the prediction of the recurrent tumor shape when using a DTI-informed anisotropic diffusion model with respect to istropic diffusion, and that the AP is a suitable metric to evaluate these models.

Finally, in **chapter 8** I discuss the findings in this thesis in a broader context and describe how each of the methodologies described in this thesis play an important role in future developments. Furthermore, I provide some future outlooks on the topic of longitudinal analysis of glioma, concluding that there is much potential for MR image analysis methods to aid radiologists in the management of low-grade glioma.

Samenvatting

Elk jaar krijgen ongeveer duizend mensen in Nederland de diagnose diffuus glioom, een type infiltrerende hersentumor dat ontstaat uit gliacellen. Er is geen genezing mogelijk voor volwassenen met een diffuus glioom, maar chirurgische resectie, radiotherapie en chemotherapie kunnen de prognose verbeteren en symptomen verlichten. Laaggradige gliomen kunnen langdurig stabiel zijn voordat een maligne progressie onvermijdelijk optreedt. De radiologische beoordeling van gliomen door middel van magnetic resonance imaging (MRI) speelt een belangrijke rol in het behandelproces van gliomen. In dit proefschrift onderzoek ik de rol van kwantitatieve metingen, nieuwe biomarkers en voorspellende modellen bij de zorg rond glioom. Deze methoden kunnen radiologen helpen om het tijdstip, de locatie en de ernst van de tumorgroei te voorspellen, met als uiteindelijk doel de kwaliteit van leven van glioompatiënten te verbeteren.

Hoofdstuk 1 introduceert de belangrijkste onderwerpen van dit proefschrift. Specifiek wordt de typering van gliomen in typen en graden beschreven, evenals de gangbare behandelpraktijken, de algemene presentatie van gliomen op MRI en de belangrijkste methodes voor beeldanalyse.

In **hoofdstuk 2** beschrijf ik een methode voor de segmentatie van gliomen met ontbrekende MRI sequenties. Deze methodologie houdt rekening met ontbrekende sequenties bij het ontwerpen en trainen van neurale netwerken, om ervoor te zorgen dat ze de best mogelijke segmentatie leveren, zelfs wanneer meerdere beelden niet beschikbaar zijn. Het voorgestelde netwerk combineert drie aanpassingen aan de standaard 3D UNet-architectuur: een trainingsmethode met uitval van sequenties, een architectuur met meerdere paden, waar fusie plaatsvindt in de laatste fase en de afzonderlijke training van deze paden.

Hoofdstuk 3 behandelt de relatie tussen de postoperatieve tumorrest en de algehele overleving bij patiënten bij nieuw gediagnosticeerd glioblastoom. Dit wordt geëvalueerd in een cohort dat behandeld is met radio-/chemotherapie met temozolomide na resectie, en de analyse is gecorrigeerd voor de prognostische waarde van O6-methylguanine DNA methyltransferase (MGMT) promotor-methylatie. Het tumorvolume wordt beoordeeld met geautomatiseerde segmentatie, gevolgd door handmatige correctie indien nodig. We ontdekten dat het contrast-aankleurend volume vóór radiotherapie sterk geassocieerd was met de algehele overleving bij patiënten die radio-/chemotherapie kregen, gescheiden naar de MGMT-promotor-methylatiestatus.

Hoofdstuk 4 beschrijft de klinische implementatie van geautomatiseerde meting van tumorvolume (EASE) voor laaggradig glioom. Naast de technische implementatie van het algoritme beschrijft dit hoofdstuk een klinisch protocol voor het gebruik van de resultaten bij de diagnose van progressie. Daarbij

zijn protocollen opgesteld om de veiligheid van de patiënt en de kwaliteit van de zorg te waarborgen bij het gebruik van volumemetingen in de klinische diagnose en voor toekomstige verbeteringen van het algoritme. Het werd toegepast op in totaal 55 patiënten, en bij 36 van hen kon de radioloog een diagnose stellen op basis van drie succesvolle opeenvolgende volumemetingen. In alle gevallen kwam de diagnose op basis van het volume overeen met de conventionele visuele diagnose.

Hoofdstuk 5 behandelt het T2-FLAIR mismatch-teken, gedefinieerd door signaalverlies van het T2-gewogen hyperintense gebied met FLAIR (Fluid-Attenuated Inversion Recovery) op MRI. Het is een zeer specifieke diagnostische marker voor IDH-gemuteerd astrocytoom, en er zijn aanwijzingen dat dit wordt veroorzaakt door intercellulaire microcysteuze veranderingen in het tumorweefsel. Het doel van dit hoofdstuk is om te bepalen of dit fenomeen prognostische waarde heeft buiten de initiële niet-invasieve diagnose. De relatie tussen het T2-FLAIR mismatch-teken en tumorgraad, microcysteuze veranderingen, algehele overleving en andere klinische parameters is onderzocht, zowel bij de eerste als bij tweede resectie in het GLASS-NL-cohort. We ontdekten dat het T2-FLAIR mismatch-fenomeen geassocieerd is met microcysteuze veranderingen, een gunstige prognose en graad 2 op het moment van tweede resectie.

Hoofdstuk 6 beschrijft een methode voor de longitudinale registratie van MRI met gliomen. Door groepsgewijze registratie op basis van deep learning uit te voeren, worden opeenvolgende beelden van dezelfde patiënt opgelijnd. Dit maakt een gedetailleerdere analyse van de veranderingen in de loop van de tijd mogelijk, zoals of het glioom expansief of invasief groeit.

Hoofdstuk 7 beschrijft modellen voor tumorgroei, die de ontwikkeling van gliomen over tijd kunnen modelleren en voorspellen. In dit hoofdstuk stellen we voor om het probleem van tumorgroei te formuleren als een rangschikkingsprobleem, in tegenstelling tot een segmentatieprobleem, en gebruiken we de gemiddelde precisie (Average Precision, AP) als uitkomstmaat. Met behulp van de AP evalueren we diffusie-proliferatie modellen die zijn gebaseerd op structurele MRI en diffusie-tensor imaging (DTI), na resectie. We concluderen dat er een significante verbetering is in de voorspelling van de vorm van de recidief tumor wanneer een een diffusiemodel wordt gebruikt op basis van anisotrope diffusie, ten opzichte van isotrope diffusie, en dat de AP een geschikte maat is om deze modellen te beoordelen.

Ten slotte bespreek ik in **hoofdstuk 8** de bevindingen in dit proefschrift in een bredere context, en beschrijf ik hoe de beschreven methodes een belangrijke rol speelt in toekomstige ontwikkelingen. Bovendien geef ik enkele toekomstperspectieven op het gebied van longitudinale analyse van gliomen, waarbij ik tot de conclusie kom dat er mogelijkheden zijn voor methoden voor MR-beeldanalyse om radiologen in de toekomst te helpen bij de zorg rond laaggradig glioom.

Acknowledgements

That was it. The work is finished. There are no more drafts and revisions, but only the final version. Maybe I have done a fraction of the research than I had planned to do, but I have learned more than I could have imagined. I learned that research is better when you are not doing it alone. So with the last words I write in this thesis I would like to thank all the people who helped me along the way, and shared parts of the journey with me.

First of all, I have to thank the almost legendary duo of **Marion** and **Stefan** for the excellent guidance and. It is a privilege to be one of the alumni of this dream team. You both set an amazing example, which I hope to keep following in the future. Though I am grateful you never expected me to follow your example in terms of working hours. **Marion**, you have been an amazing mentor and inspiration. It was great to be a part of a research group that is so diverse, in terms of background and disciplines. I came to the group without any knowledge of glioma or MRI, but I was taught by the best. **Stefan**, with your technical guidance and eye to detail you could always make my work that much better. Your excitement and curiosity are infectious, and I always enjoyed our brainstorms together. I hope a bit of that passion has rubbed off on me.

Then there are the many colleagues whose company I enjoyed along the way. I was lucky to have many of them, so apologize in advance that I cannot name them all. First and foremost I want to thank my paranymphs, who were by my side for much of this journey and will be at the very end. **Sebastian**, as my predecessor you helped me on my feet. You taught me by example to do things in the correct way, even if that was not always the easy way. Your efforts to leave behind useful software and data were never wasted, and neither were your efforts to make the office a fun place. **Fatemeh**, my friend, your presence always lights up the room and I could always count on you to boost my confidence. Thank you for sharing your chocolate and gossip. I am sure you will be very succesful in whatever challenge comes next, and I hope I will be there to hear the stories.

I would like to thank all my colleagues at BIGR. **Martijn**, it was always a joy work together, or share a coffee break. We could always complain together about all the things that could be better, and I am sure you will make it happen. **Wietske**, thank you for the cake, the gezelligheid and wise words. **Kim**, thank you for sharing your limitless energy and enthusiasm with us. **Bo**, thank you for sharing ideas and plans for shared projects with me, even if we only finished a fraction. **Hakim**, it was a pleasure to teach with you during the crazy times of COVID and afterwards. Thank you **Esther**, for letting

me tag along in the neuro meetings. Thank you to **Marcel**, **Adriaan** and **Mart** for your help in getting EASE up and running. Special thanks also to **Jose**, **Robin**, **Arno**, **Hoel**, **Douwe**, **Richard**, **Antonio**, **Danilo**, **Luisa**, and everyone else who made life at BIGR lively, even through the pandemic. The list is honestly too long, which is amazing, so I apologize for not thanking everyone by name.

I also want to thank my colleagues at the neuro-onco-imaging group. It was great to see people join over time with such different backgrounds and expertise. Special thanks to **Wouter** for answering my quick radiological questions, **Ahmad** for sharing a hundred nice chats and annotating approximately a million MRI's, **Esther**, **Fatih**, **Sophie**, **Ilanah**, **Krishnapriya**, **Patrick** and everyone else for the many fun meetings and coffee breaks.

Then, I was lucky to also venture outside the radiology department. Thank you to everyone in **GLASS-NL** for helping to create this fascinating project. I want to thank especially **Wies** for doing most of the hard work in putting this cohort together, **Martin** for patiently correcting me on everything related to glioma, and to **Pim** for keeping me as imaging person on board when you were discussing the molecular side of things. Thank you also to **Max** for introducing me to the world of glioma histopathology and indulging me in my quest for microcysts.

And thank you to all other students, fellows, and other colleagues who crossed my path, and to the **Eye-EPI** group for the warm welcome after finishing this research.

Dan is er naast werk natuurlijk ook nog een leven. Dank aan al mijn vrienden en familie die me gesteund hebben de afgelopen jaren. In het bijzonder natuurlijk mijn ouders, **Endry** en **Harm-Jan**, die mij er al sinds de basisschool aan herinneren dat studeren leuk is, maar ook weer niet zo belangrijk, en mijn grote broers **Frank** en **Marcel**, en natuurlijk **Kathelijn**. Ook wil ik mijn schoonfamilie, **Willeke**, **Erik**, **Yannick**, **Annemarie**, **Robin** en **Jamie**, bedanken. Jullie noemen het dan wel de koude kant, maar zo heeft het zeker nooit gevoeld.

Dan is er nog mijn allergrootste steun en toeverlaat, **Raoul**, die altijd voor mij klaar staat en samen met mij elk avontuur, inclusief deze, aan wil gaan. Dankjewel voor alles.

Last but not least, I feel that I need to acknowledge the **patients** who allowed me to peer into their brains. I hope my work has made a difference.

Publications

Journal Papers

Karin A. van Garderen, S. R. van der Voort, A. Versteeg, M. Koek, A. Gutierrez, M. van Straten, M. Rentmeester, S. Klein, and M. Smits, "EASE: Clinical implementation of automated tumor segmentation and volume quantification for adult low-grade glioma," *Frontiers in Medicine*, p. 1791, Oct. 2021.

F. Arzanforoosh, P. L. Croal, **Karin A. van Garderen**, M. Smits, M. A. Chappell, and E. A. H. Warnert, "Effect of applying leakage correction on rCBV measurement derived from DSC-MRI in enhancing and nonenhancing glioma," *Frontiers in Oncology*, vol. 3, Mar. 2021.

L. Nunez-Gonzalez, **Karin A. van Garderen**, M. Smits, J. Jaspers, A. M. Romero, D. H. J. Poot, and J. A. Hernandez-Tamames, "Pre-contrast MAGiC in treated gliomas: A pilot study of quantitative MRI," *Scientific Reports*, vol. 12, p. 21 820, 1 Dec. 2022.

A. Alafandi, **Karin A. van Garderen**, S. Klein, S. R. van der Voort, D. Rizopoulos, L. Nabors, R. Stupp, M. Weller, T. Gorlia, J. C. Tonn, and M. Smits, "Association of pre-radiotherapy tumor burden and overall survival in newly diagnosed glioblastoma adjusted for MGMT promoter methylation status," *European Journal of Cancer*, vol. 188, pp. 122–130, Jul. 2023.

Karin A. van Garderen, S. R. van der Voort, M. M. Wijnenga, F. Incekara, A. Alafandi, G. Kapsas, R. Gahrmann, J. W. Schouten, H. J. Dubbink, A. J. Vincent, M. van den Bent, P. J. French, M. Smits, and S. Klein, "Evaluating the predictive value of glioma growth models for low-grade glioma after tumor resection," *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, pp. 253–263, 2024.

Karin A. van Garderen*, W. R. Vallentgoed*, A. Lavrova, J. M. Niers, W. W. de Leng, Y. Hoogstrate, I. de Heer, B. Ylstra, E. van Dijk, S. Klein, K. Draaisma, P. A. Robe, R. G. Verhaak, B. A. Westerman, P. J. French, M. J. van den Bent, M. C. Kouwenhoven, J. M. Kros, P. Wesseling, and M.

Smits, "Longitudinal characteristics of T2-FLAIR mismatch in IDH-mutant astrocytoma: Relation to grade, histopathology and overall survival in the GLASS-NL cohort," *Neuro-Oncology Advances*, vol. 5, no. 1, vdad149,

Conference Papers

Karin A. van Garderen†, M. Smits, and S. Klein, "Multi-modal segmentation with missing MR sequences using pre-trained fusion networks," in *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, Presented at MICCAI MIL3ID Workshop, vol. 11795 LNCS, Springer, 2019, pp. 165–172.

K. B. de Raad†, **Karin A. van Garderen**, M. Smits, S. R. van der Voort, F. Incekara, E. H. Oei, J. Hirvasniemi, S. Klein, and M. P. Starmans, "The effect of preprocessing on convolutional neural networks for medical image segmentation," in *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 655–658.

Karin A. van Garderen†, S. R. van der Voort, M. M. J. Wijnenga, F. Incekara, G. Kapsas, R. Gahrmann, A. Alafandi, M. Smits, and S. Klein, "Evaluating glioma growth predictions as a forward ranking problem," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Presented at MICCAI BrainLesion Workshop, Springer, 2022, pp. 100–111.

Conference Abstracts

Karin A. van Garderen†, S. R. van der Voort, F. Incekara, M. Smits, and S. Klein, "Towards continuous learning for glioma segmentation with elastic weight consolidation," Abstract presented at MIDL 2019, Sep. 2019.

C. C. Hammecher†, **Karin A. van Garderen**, M. Smits, P. Wesseling, B. Westerman, P. French, M. Kouwenhoven, R. Verhaak, F. Vos, E. Bron, and B. Li, "Deep learning-based groupwise registration for longitudinal MRI analysis in glioma," *Abstract presented at ISMRM 2023*. arXiv: 2306.10611.

* indicates equal contributions
† indicates presenting author

PhD portfolio

Courses	Year	ECTS
BROK *Erasmus MC, The Netherlands*	2018	1
Advanced Neuro Imaging: Diffusion, Perfusion, Spectroscopy *ESMRMB School of MRI, Leuven, Belgium*	2018	2
Scientific Integrity *Erasmus MC, The Netherlands*	2019	0.3
Markers and Prediction Research *Erasmus MC, The Netherlands*	2019	0.7
Cohort studies *Erasmus MC, The Netherlands*	2019	0.7
Joint Models for Longitudinal and Survival Data *Erasmus MC, The Netherlands*	2019	0.7
MLSS: Machine Learning Summer School *Skoltech, Moscow, Russia*	2019	4
Inverse Problems in Imaging *MasterMath, Utrecht University, The Netherlands*	2020	5
Total		14.4

International and local research meetings	**Year**	**ECTS**
MIDL Conference *London, United Kingdom*	2019	1
ESMRMB Conference *Rotterdam, the Netherlands*	2019	1
MICCAI Conference *Shenzhen, China*	2019	1.5
ESMRMB Conference *Virtual*	2020	1
MICCAI Conference *Virtual*	2020	1.5
WBIR Conference *Virtual*	2020	1
MICCAI Conference *Virtual*	2021	1.5
GliMR annual meeting: Bridging Clinic and Research *Virtual*	2021	1
GliMR networking event *Brno, Czech Republic*	2021	1
Biomedical imaging group seminars (biweekly) *Erasmus MC, The Netherlands*	2018 – 2022	1
Medical Informatics research lunch meeting (biweekly) *Erasmus MC, The Netherlands*	2018 – 2020	1
Radiomics research meetings (monthly) *Erasmus MC, The Netherlands*	2018 – 2022	0.5
Neuro-Oncology meeting (monthly) *Erasmus MC, The Netherlands*	2018 – 2022	0.5
Total		14

Student supervision	**Year**	**ECTS**
Supervision Master thesis - Alice Dudle *Primary Liver Tumor Classification on MRI using Deep Learning*	2018	0.5
Supervision internship - Abdullah Thabit *Active Learning for Glioma Segmentation*	2019	0.5
Supervision Bachelor thesis - Diederik Moorlag *Knee Joint Segmentation Based on MRI scans*	2019	0.5
Supervision Master thesis - Koen de Raad *The Effect of Preprocessing on Convolutional Neural Networks for Medical Image Segmentation*	2020	0.5
Supervision internship Edgar van der Meer *Automated Segmentation of Tissue Structures in the Brain using Deep Learning*	2019	0.5
Supervision internship Zayaan Khan *Data Sharing in Multi-center Studies*	2019	0.5
Supervision internship Giovanni Hitharie *Pre-processing a low-grade Glioma Dataset for Deep learning*	2019	0.5
Supervision Master thesis Hao Ni *Deep learning for 4D Longitudinal Segmentation of MRI Brain Tissues and Glioma*	2020	0.5
Supervision Master thesis Teun Tanis *AUTOMONAI: Towards automatic tuning of medical image segmentation networks*	2021	0.5
Supervision Master thesis Coen van Gruijthuijsen *Automated Machine Learning in Medical Image Segmentation*	2020	0.5
Total		5

Teaching activities	**Year**
Image Processing *Bachelor Clinical Technology*	2020
Machine Learning *Master Clinical Technology*	2020 – 2022
Advanced Image Processing *Master Technical Medicine*	2022

Grants & Awards	**Year**
Pilot grant SURFsara	2018

Committees	**Year**
Neuro-oncology Imaging meeting *Organizing committee member*	2018 - 2020
GPU cluster Radiology *Administrator*	2020 - 2022
Erasmus MC Brain Tumor Center Retreat *Organizing committee member*	2022

About the author

Karin Alida van Garderen was born in Freyung, Germany on the 14th of January 1993. At the age of three, she moved to Bilthoven, the Netherlands, where she spent most of her childhood. She completed secondary school at the Christelijk Lyceum Zeist in 2011, after which she started a Bachelor of Applied Physics at Delft University of Technology. After successfully completing the bachelor and a bridging program, she transitioned to computer science and completed a Master of Computer Science in 2018, with a specialization in Data Science and Engineering. She completed her master thesis during an internship at ASML, Eindhoven, where she investigated the use of interactive machine learning for the prediction of manufacturing errors in lithography. Upon completion, she received the predicate "Cum Laude".

In July 2018, Karin started a PhD at Erasmus MC on the topic of longitudinal assessement of low-grade glioma. She was supervised by prof. dr. Marion Smits and dr. Stefan Klein. As a data scientist, she first investigated the role of machine learning in the automated assessment of MRI in glioma management, but as the project developed and the clinical research questions became more clear, the focus of her research shifted to the clinical aspects of longitudinal assessment. During her time as a PhD student she also participated in teaching activities and helped to develop and teach a course in machine learning for students of technical medicine. After completion of her PhD trajectory she continued her work on the intersection between technical and clinical research at Erasmus MC, but now in the department of Ophthalmology.