

**UCSF**

**UC San Francisco Electronic Theses and Dissertations**

**Title**

Toward improving the diagnosis of glioma

**Permalink**

<https://escholarship.org/uc/item/16c430mk>

**Author**

Cluceru, Julia

**Publication Date**

2021

Peer reviewed|Thesis/dissertation

Toward improving the diagnosis of glioma

by  
Julia Cluceru

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of  
DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:

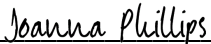


F1E4E52A4E3D4D8...

Janine Lupo

Chair

DocuSigned by:



DocuSigned by:

513E3503F4574E4...

Joanna Phillips

Mark Segal

Committee Members



# Acknowledgments

There are many people that I would like to thank for guiding me and helping me through the 5.5 years I worked to obtain my PhD. First and foremost, I'd like to thank my academic advisor Dr. Janine Lupo. My first year and a half at UCSF was filled with personal and academic growth during which my priorities and interests shifted immensely. Though I was accepted into UCSF to study pharmaceutical sciences, I became set on improving the diagnosis of hard-to-treat cancer. In my neuro-oncology mini-course, I learned from a journal club with Dr. Serah Choi that radiation therapy can cause relatively benign lesions that appear identical to glioma recurrence. I became entranced with this problem, and after speaking with Dr. Lupo about my interest in solving this problem, she generously offered me a rotation in her lab despite my having no background in coding nor bioengineering nor MR physics. Her willingness to take a chance on me completely changed the course of my life and career. She has been incredibly patient as I learned to code, frame a scientific question, and attempt to solve some of the most pressing and difficult questions that face clinical neuro-oncology practice. I am consistently impressed by her immense knowledge of MRI, scientific writing, attention to detail, and kindness - all of which have had an immense impact on the scientist I have become throughout my PhD. Janine, I will always be so thankful.

Additionally, I want to thank Dr. Joanna Phillips for her guidance on my qualifying exam committee as well as my thesis committee. It was Dr. Phillips who originally suggested that I ask Dr. Lupo and Dr. Sarah Nelson about working in diagnosing brain tumor recurrence. Without her, I'm not sure that I would have ever found the right lab to study and pursue this interest. Once I joined Dr. Lupo's lab, I worked for two years on a project involving image-guided tissue samples (Chapter 6) for which her expert pathological evaluation served as the ground truth. She



patiently explained the meaning of numerous pathological markers, deepening my understanding of the biology of the glioma.

I would also like to thank Dr. Mark Segal for his time and expertise. One of my projects was very statistically complex because multiple image-guided tissue samples per patient introduces correlation among data that must be accounted for. His time and feedback about my varied approaches to account for spatially repeated measures was extremely helpful, and for that I am very thankful.

The Multiple Sclerosis Bioscreen tool (that has also gained traction in the UCSF Neuro-Oncology community) is the only project I have worked on that is actually deployed. I am so proud that something I've worked on is actually being used daily by clinicians to help diagnose patients. I want to thank Dr. Jason Crane immensely for his willingness to serve as my mentor for this project. He taught me a lot about best coding practices and how to think deeply about radiomics, imaging features, and deep learning.

Without the mentorship of Dr. Yannet Interian and Dr. Paula Alcaide-Leon, I would not have finished my PhD. Dr. Alcaide-Leon lent the perspective of a clinician to all of my studies and talked with me for hours on end when I felt stuck. Dr. Interian taught me how to use PyTorch and was a crucial resource to understand how to structure machine learning and deep learning projects. Every meeting with either Dr. Alcaide-Leon and Dr. Interian left me with renewed inspiration to continue my research. They were both so generous with their time, encouragement, and mentorship throughout my PhD and I am incredibly thankful.

I would like to thank the late Dr. Sarah Nelson who served as the chair of my qualifying exam committee. I will never forget how she believed in me. Though she retired early on in my

PhD, our whole brain tumor imaging laboratory owes our success to her work. I am proud that my work is part of her legacy.

Dr. Deanna Kroetz was the Pharmaceutical Sciences and Pharmacogenomics department chair throughout the majority of my time at UCSF. She always served as an important, neutral sounding board and was incredibly compassionate during the hardest times. I will always be thankful for those meetings.

I want to thank all of the incredibly kind and awesome people in the Brain Tumor Imaging Group who made my experience throughout these five and a half years so much more enjoyable. Angela Jakary and Marisa Lafontaine for all of their hard work scanning patients and segmenting tumors, Devika Nair for tirelessly collecting tissue samples during surgery and helping me with data collection, and Dr. Tracy Luks for her help with all kinds of issues. Beck Olson was always willing to help me with my computing problems. Dr. Trey Jalbert was the first person I met in the lab, whose optimism was contagious and inspirational; I learned a lot from his example. I was so shy and nervous to be in Dr. Lupo's lab at first, so I will never forget Dr. Adam Autry's first words to me - "Uh, who are *you*?" - which, maybe the tone doesn't come through in this text, but at the time made me laugh a lot and feel right at home. Adam together with Melanie Morrison sat next to me and were so encouraging throughout my PhD journey. Sivakami Avadiappan and Dr. Eason Chen were also so kind and a joy to be around. Dr. Maryam Vareth served as a strong and smart role model for me, who graciously invited me to study at and collaborate with people at the Berkeley Institute for Data Science when I was first getting interested in machine learning. To all of you, I am so thankful and my PhD would not have been the same without you.

Dr. Simon Gao and Neha Anegondi were my mentors during my internship at Genentech. I had just three months time working with them, but that experience was invaluable to my professional development and confidence. I want to thank them for affording me the opportunity to work there both over the summer, and I am extremely excited to begin my career with them as a Clinical Imaging Scientist at Genentech in just a few months time.

Beau Norgeot and Jasleen Sodhi were and are my best friends in the Pharmaceutical Sciences and Pharmacogenomics graduate program. They are both so incredibly successful and smart and encouraging. If you're reading this: I love you guys and thank you so much for the late nights, tasty beverages, and encouragement throughout my time at UCSF. I couldn't have done it without you. To every single one of the rest of my friends, you guys kept me sane and happy throughout it all. I love you and thank you for being there for me!

To my mother, Margaret Louise Marks Cluceru, this dissertation is dedicated to you. I wish you were here to see me graduate from my PhD, I think you would have been really proud. I miss you every day, but it is your strength that I carry to accomplish something like this. In addition to my mom, I want to thank my brother, Alexander Higle, my sister, Heather Higle, and my father Valentin Cluceru for shaping me into who I am. Mom's illness and death made this PhD incredibly difficult for me but having such a supportive and loving family helped get me through.

Last but not least, I want to thank my partner Matthew Faluotico. Words can't even describe how essential his support and unconditional love has been for the completion of this degree. It was him who comforted and encouraged me and made me laugh through the most difficult moments but was also there to celebrate every achievement. I love you and thank you more than you could ever know.

## Contributions

Chapter 7.1 is adapted from the following publication:

Cluceru, J.; Nelson, S. J.; Wen, Q.; Phillips, J. J.; Shai, A.; Molinaro, A. M.; Alcaide-Leon, P.; Olson, M. P.; Nair, D.; LaFontaine, M.; Chunduru, P.; Villanueva-Meyer, J. E.; Cha, S.; Chang, S. M.; Berger, M. S.; Lupo, J. M. Recurrent Tumor and Treatment-Induced Effects Have Different MR Signatures in Contrast Enhancing and Non-Enhancing Lesions of High-Grade Gliomas. *Neuro Oncol.* **2020**, 22, 1516–1526.

<https://pubmed.ncbi.nlm.nih.gov/32319527/>

# Abstract

## *Toward improving the diagnosis of glioma*

Julia Cluceru

Glioma is a heterogeneous and incurable neoplastic mass derived from the glial cells in the brain. The clinical course of a glial tumor can range from slow growing to highly aggressive and it is driven by the genetic and epigenetic alterations within the neoplastic cells' DNA. In order to diagnose both that a patient has a glioma and what kind of glioma it may be, it is necessary to acquire magnetic resonance images (MRI) of the brain. MRI not only provides unparalleled soft tissue contrast in the brain compared with other tomographic imaging techniques (e.g. CT), it is also incredibly flexible: in addition to structural anatomy of the lesion, it can probe the physiology (e.g. diffusion, perfusion) and metabolism of the brain. Together, these data guide neuroradiologists and neurooncologists to provide the optimal treatment plan for a patient with glioma.

Even with the most talented clinicians and the highest-quality MR acquisitions, there still exist issues along the trajectory of diagnosing a glioma. In this dissertation, I harnessed the incredibly rich biological information within the pixels of MRI with modern advances in data science and computer vision to address three of the most urgent problems along the diagnostic pipeline: 1) Can we identify a patient's glioma subtype prior to surgical intervention?; 2) Can we create an automatic MR dashboard for clinicians to monitor patient disease over time?; and 3) When a patient appears to recur, is it a true glioma or the effect of radiation therapy? For each of these questions, I found that rigorous application of statistical learning and deep learning together with MR imaging can improve these problems, even with limited patient data.

# Table of Contents

<b>1. Introduction</b> .....	<b>1</b>
<b>2. Brain tumors</b> .....	<b>6</b>
2.1 Epidemiology and the evolving classification of brain tumors .....	6
2.2 The challenges of glioma treatment .....	8
2.2.1 Treating brain tumors across the blood-brain barrier .....	8
2.2.2 Treatment for Newly Diagnosed Patients .....	9
2.2.3 Treatment of Recurrent Glioma .....	14
<b>3. Magnetic resonance imaging of glioma</b> .....	<b>16</b>
3.1 Conventional anatomic imaging sequences .....	16
3.1.1 Introduction to MRI .....	16
3.1.2 From signal to image .....	24
3.1.3 The role of conventional anatomic imaging in brain tumor diagnosis .....	26
3.2 Diffusion MRI .....	28
3.2.1 Introduction to diffusion-weighted and diffusion-tensor imaging .....	28
3.2.2 The role of diffusion-weighted and diffusion tensor imaging in glioma .....	31
3.3 Perfusion MRI .....	34
3.3.1 Introduction to perfusion MRI .....	34
3.3.2 The role of perfusion-weighted imaging in brain tumor diagnosis .....	38
3.4 MR Spectroscopy .....	40
3.4.1 Introduction to MR Spectroscopy .....	40

3.4.2 The role of MR spectroscopy in brain tumor diagnosis.....	44
<b>4. Introduction to machine learning and deep learning for neuroimaging.....</b>	<b>49</b>
4.1 Introduction to machine learning.....	49
4.2 Classic machine learning techniques.....	50
4.2.1 Linear models.....	50
4.2.2 Decision trees and random forests.....	52
4.3 Spatially repeated measures.....	55
4.4 Introduction to deep learning and convolutional neural network classifiers.....	57
4.4.1 Introduction to artificial neural networks.....	57
4.4.2 Introduction to convolutional neural networks.....	60
<b>5. Presurgical identification of genetic alterations in glioma.....</b>	<b>66</b>
5.1 Introduction.....	66
5.2 Methods.....	69
5.2.1 Patient Characteristics and Study Design.....	69
5.2.2 Assessment of genetic alterations.....	71
5.2.3 Image Acquisition and Processing.....	71
5.2.4 Baseline models from clinical metrics.....	73
5.2.5 Hyperparameter search.....	75
5.2.6 Model comparison.....	75
5.2.7 Model generalization to an independent test set.....	76
5.3 Results.....	79
5.3.1 Characteristics of the Study Sample.....	79
5.3.2 Clinical baseline modeling results.....	80

5.3.3 Model development results .....	81
5.3.4 Model comparison .....	83
5.4 Discussion .....	88
<b>6. Automatic classification of MR image contrast .....</b>	<b>95</b>
6.1 Introduction .....	95
6.2 Materials and methods .....	101
6.2.1 MR Sequences .....	101
6.2.2 MR exams and cohorts .....	102
6.2.3. Labels.....	102
6.2.4 Train, validation, and test splits.....	103
6.2.4 Rule-based approach.....	104
6.2.5 Metadata models.....	104
6.2.6 Image processing and imaging model development.....	105
6.2 Results.....	107
6.2.1 Rule-based approach.....	107
6.2.2 Modeling results .....	107
6.2.3 t-SNE results .....	114
6.3 Discussion.....	117
<b>7. Distinguishing recurrent tumor from treatment-induced effects .....</b>	<b>122</b>
7.1 Recurrent tumor and treatment-induced effects have different MR signatures in contrast- enhancing and non-enhancing lesions of high-grade gliomas .....	123
7.1.1 Introduction .....	123
7.1.2 Methods .....	126



7.1.3 Results.....	134
7.1.4 Discussion.....	139
7.2 Using anatomic and diffusion MRI with deep convolutional neural networks to distinguish treatment-induced injury from recurrent glioblastoma.....	145
7.2.1 Introduction.....	145
7.2.2 Methods.....	146
7.2.3 Results and Discussion.....	150
7.2.4 Conclusion.....	153
<b>8. Conclusions and future directions.....</b>	<b>154</b>
8.1 Conclusions.....	154
8.2 Future directions.....	156
<b>References.....</b>	<b>158</b>

# List of Figures

Figure 2.1. Evolving WHO categorization of glioma places greater emphasis on genetic subtype.....	8
Figure 2.2. Mechanism of Temozolomide.....	11
Figure 2.3. The benefit of PCV treatment for 1p19q codeleted patients. ....	12
Figure 2.4. The similarity of treatment-induced lesions and recurrent, high-grade glioma on conventional MR imaging.....	15
Figure 3.1. Excitation and magnetization of a proton depicted in (a) the lab frame and (b) the rotating frame.....	18
Figure 3.2. Example of magnetization differences between tissues A and B when repetition time (TR) is short. ....	21
Figure 3.3. Example tissue recovery times of (A) longitudinal and (B) transverse magnetization after an RF pulse.....	21
Figure 3.4. Effects of acquisition parameters on image contrast.....	23
Figure 3.5. Localizing signal in MRI using (A) a slice selection gradient; (B) frequency encoding gradient; and (C) a phase encoding gradient.....	24
Figure 3.6. Relationship of signal acquired in k-space (time domain) to an MR image (spatial domain) via the inverse Fourier transformation.....	25
Figure 3.7. Schematic of basic pulse sequence sensitive to the diffusion of water molecules within tissues.....	29
Figure 3.8. The creation of diffusion-weighted images and ADC maps from three directions.....	29
Figure 3.9. Arterial Spin Labeling (ASL).....	36
Figure 3.10. Diagram of the two-compartment pharmacokinetic model used to calculate $K^{\text{trans}}$ in DCE imaging. ....	36
Figure 3.11. Changes in $T2^*$ and $\Delta R2^*$ signal for DSC metric calculations.....	37
Figure 3.12. The NMR spectrum of lactate.....	42

Figure 3.13. Chemical structures of metabolites relevant for quantifying in glioma at long TE. ....	45
Figure 3.14. Examples of differential metabolite concentrations in regions with disparate pathological burden. ....	46
Figure 4.2. Diagram of a Random Forest algorithm.....	54
Figure 4.3. Structural comparison between a biologic neuron and an artificial neuron.....	58
Figure 4.4. Diagram of a simple two-layer feed-forward artificial neural network.....	59
Figure 4.5. Convolution operation of a 4x4 input image with a 3x3 shaded kernel to create a 2x2 green output.....	61
Figure 4.6. Example of a max pooling and average pooling operations in a convolutional neural network. ....	62
Figure 4.7. Architecture of LeNet-5. ....	62
Figure 4.8. Architecture of AlexNet. ....	63
Figure 4.9. Residual “skip” connections that improve training accuracy of very deep convolutional neural networks. ....	64
Figure 5.1. Inclusion and exclusion criteria that led to the final numbers used in the study. ....	70
Figure 5.2. Study design detailing the hyperparameter search and model comparison phase. ....	70
Figure 5.3. Schematic of image processing strategy. ....	74
Figure 5.4. Acceptable and unacceptable training/validation loss curves.....	78
Figure 5.5. 3-class model development tradeoff. ....	78
Figure 5.6. Main results from the hyperparameter search and the model comparison phase.....	82
Figure 5.7. Additional insights from the hyperparameter search phase.....	82
Figure 5.8. Detailed patient classification of each of the final models.....	85
Figure 5.9. GradCAM analysis.....	87
Figure 6.1 The application of contrast classification of MRI in clinical practice. ....	100
Figure 6.2. Examples of images in each category. ....	101
Figure 6.3. RF feature importance graphs on validation and test sets (MSC + GR together).....	110

Figure 6.4. t-SNE of the 6 logits derived from the final layer of the convolutional neural network. ....	116
Figure 7.1 Overview of tissue samples. ....	127
Figure 7.2. Within-patient imaging differences between treatment-induced injury and recurrent HGG..	134
Figure 7.3. Boxplots representing distributions of values in recurrent HGG samples and treatment-induced injury samples. ....	135
Figure 7.4. Summarizing the centrally restricted diffusion sign. ....	146
Figure 7.5. Image processing pipeline. ....	148
Figure 7.6. Modeling approaches. ....	149
Figure 7.7. ResNet-34 CNN architecture and 3-direction averaging technique utilized. ....	149
Figure 7.8. ROC curve for each fold using the probabilities derived from the sigmoid value of the average logit. ....	150
Figure 7.9. Examples of misclassified patients. ....	151
Figure 7.10. Increasing validation loss while overfitting to the training fold. ....	152

# List of Tables

Table 3.1. Effects of acquisition parameters on image contrast.....	23
Table 5.1. Hyperparameter search space. ....	77
Table 5.2. Patient demographics and differences between UCSF and TCGA data.....	79
Table 5.3. Results of all six models compared.....	80
Table 5.4. Final hyperparameters of each of the deep learning models in the model comparison phase. ...	84
Table 6.1. Distribution of MR exams and series used for this study.....	102
Table 6.2 Metadata-only random forest feature importances.....	111
Table 6.3 Combined random forest feature importances.....	112
Table 6.4 Final model comparison for the MSC cohort. ....	113
Table 6.5 Final model comparison for the GR cohort. ....	113
Table 6.6 Misclassification analysis of MSC cohort.....	113
Table 6.7 Misclassification analysis of GR cohort.....	113
Table 7.1. Clinical demographics of the patient population.....	127
Table 7.2. Number of samples included for each test.....	136
Table 7.3. Generalized estimating equation (GEE) results from imaging values associated with pathology. ....	137
Table 7.4. Threshold and logistic regression analysis results. ....	139

# 1. Introduction

Gliomas are rare but deadly. They have relatively few treatment options, due in part to their clinical heterogeneity, in part to their genetic heterogeneity, and in part due to their location in the brain, one of the most functionally important organs in our bodies. In turn, diagnosing patients early and accurately is critical to maximizing longevity and quality of life.

A journey through the trajectory of a patient diagnosed with glioma can add perspective to the kinds of questions that still remain unanswered by clinicians and scientists. Typical patients present with some combination of seizures, headaches, nausea, or disequilibrium. A clinician will administer an MRI, from which a trained neuroradiologist will have to make an assessment of whether the patient is experiencing a glioma or a lesion that can look extremely similar, like a metastasis or a CNS lymphoma[1].

Once the diagnosis of glioma is established, the following decision is whether the patient needs immediate surgical resection or if alternative treatment options are appropriate. The need for surgical resection can depend on many factors, and institutions handle this decision differently. In general, the decision of whether to perform surgery or utilize other treatment options depends greatly on the prognosis of the patient. The prognosis is determined by what *kind* of glioma that patient has. Throughout history, defining the “kind” of glioma has evolved. Most recently, the collection of genetic alterations that the tumor harbors is the most important factor for understanding a patient's trajectory and response to different therapies. This is discussed in more detail in Chapter 1, but it follows that presurgically identifying a patient's genetic alterations has important implications in treatment management.

Following this determination, many patients will go on to receive neurosurgery, from which tissue can be collected and the genetic variations can be directly analyzed. Once confirmed, the genetic alterations will inform the decision to administer radiation therapy and/or chemotherapy. During and after this treatment period, patients receive routine MRI to monitor response to treatment as well as recurrence of their tumor.

It is the job of a skilled neuroradiologist to monitor changes in patients' brains over time. However, this can prove to be a more daunting task than it seems: each MRI scan consists of at the very least 4 different kinds of MR image contrasts of soft tissue. The trouble with all of these different MRI acquisition strategies is that there is no consistent way of labeling or acquiring them across institutions. In other words, it is often up to the institution or even the specific MR technician to describe the kind of MR acquired in the series description field. These MR images are stored in hospital databases, so the radiologist who wishes to monitor a patient's changes over time must go into the database, find the modality of interest at timepoints A, B, and C etc., find the same slice of the MR volume at each timepoint, and subsequently assess the tumor size changes through time either qualitatively or using crude quantitative tools, like measuring the change in diameter. However, because of software limitations and the immense heterogeneity in MR volume labeling, this process can be extraordinarily tedious, taking up valuable clinician time. In an ideal world, there would exist a dashboard that automatically presents aligned slices of the chosen MR modality together with important clinical metrics (e.g. tumor volume) over timepoints A, B, and C.

Once the proper scans are pulled up and aligned, the neuroradiologist is faced with determining whether a patient is responding to treatment. If a lesion is not apparent or is obviously shrinking, the answer is straightforward. However, if a new enhancing lesion appears,

the answer may not be so clear. True, it may be the case that there is a recurrent tumor that needs to be treated again; however, it may be the case this is actually not a recurrent tumor at all, but rather the effect of radiation therapy causing inflammation and necrosis that mimics the appearance of a real recurrence. In turn, a clinician can use alternative MRI modalities or even PET scans to attempt to understand the underlying biology driving this new lesion, but this remains an active area of research and is not yet integrated into clinical practice. With the advent of new immuno- and antiangiogenic therapies, the growth or shrinkage of lesions is even more complex, as antiangiogenic therapies may lessen lesion appearance without actually mitigating the aggressiveness of the tumor.

The work presented in this dissertation aims at many pain points described within this process. The following structure is used to take the reader through the essential background for understanding the research that has gone toward improving the diagnosis of glioma.

### *Chapter 2: Introduction to Brain Tumors*

This chapter aims to provide context for the reader to understand the prevalence of intracranial tumors as well as some of the relevant basic biology and genetic mutations that define glioma. Even further, it describes how the biology of tumors can influence the treatment decisions that are made in patient management. It dives into the complicated nature of treating glioma, including the challenges presented by neurosurgery itself and the delivery of therapeutics across the blood brain barrier.



### *Chapter 3: Introduction to MRI for imaging gliomas*

This chapter provides the necessary background for understanding the science of magnetic resonance imaging. It describes the phenomenon of nuclear magnetic resonance as well as signal generation and acquisition. It dives into the generation of conventional T1 and T2 weighted MR images, as well as diffusion, perfusion, and spectroscopic MR. For each kind of MR, a discussion of its relevance for the diagnosis of brain tumors is presented.

### *Chapter 4: Introduction to machine learning*

This chapter is meant to introduce the reader to the fundamentals of machine learning while also capturing the power and relevance it has to medical imaging. It begins with an introduction to linear and logistic regression, which are used to illustrate concepts used later in deep neural networks (e.g. backpropagation). This section covers all statistical and machine learning concepts relevant for research presented in chapters 5 through 7; such as random forest, spatially repeated measures, and convolutional neural networks.

### *Chapter 5: Presurgical identification of genetic alterations in glioma*

In this study, concepts from the first three chapters are combined to tackle one of the problems outlined in the introduction: presurgical identification of molecular subtype in glioma. Though many prior investigations use deep learning to predict IDH mutation status from MR imaging, few have incorporated 1p19q chromosomal arm codeletion, a defining feature of glioma subtype. In addition, the optimal framework and modalities for prediction of both IDH and 1p19q status are investigated.

### *Chapter 6: Automatic annotation of MR Image contrasts for Neuro-Imaging*

To perform analyses of disease progression, it is necessary to evaluate the change in disease over time on MR images of similar contrast. The goal of this study is to create an algorithm that can reliably classify brain exams by MR image contrast as part of an automatic MR annotation and delivery system to both clinicians and researchers. We use two modeling strategies (metadata with random forest and imaging data with deep learning) to compare the performance of different approaches on a poorly-annotated, heterogenous multiple sclerosis dataset. We then demonstrate generalizability of our algorithm to a distinct cohort of brain tumors with increased disease burden.

### *Chapter 7: Differentiating the effects of treatment from recurrent tumor*

One of the most important questions for clinicians monitoring patients for recurrence is whether a newly enhancing lesion on an MRI is a real recurrent tumor, or whether it is the effect of treatment; often, they are identical on conventional MRI. In this study, we aim to identify physiologic and metabolic MR imaging parameters that are associated with true recurrent tumor or treatment effect. We utilize image-guided tissue samples acquired with neuronavigation tools to match specific regions of MRI to the pathological outcome, thereby overcoming heterogeneity within patients.

### *Chapter 8: Conclusions and Future Directions*

In Chapter 8, we summarize the findings of this dissertation and discuss future directions of the work presented.

## 2. Brain tumors

### 2.1 Epidemiology and the evolving classification of brain tumors

Intracranial cancers can originate from cells in the brain (“primary”) or from tumors elsewhere in the body (“secondary” or “metastasis”). Approximately half of newly diagnosed brain tumors are primary, which impact 23.79 people per 100,000 in the United States per year[2]. Of these, 70.3% are benign, while 29.7% are malignant. All in all, this translates to an estimated and expected 24,970 primary malignant brain tumors diagnosed in the year 2021.

Within the subset of primary malignant brain tumors, glioma - or tumors arising from the glial cells in the brain - comprise 81%[3]. Glioma are extremely clinically heterogeneous: the five year survival rate ranges between 7.2% and 83.4%, with different symptomatic burdens and responses to therapy. In light of this enormous variation, it is an important clinical objective to accurately diagnose glioma patients with favorable and unfavorable prognoses more specifically and accurately.

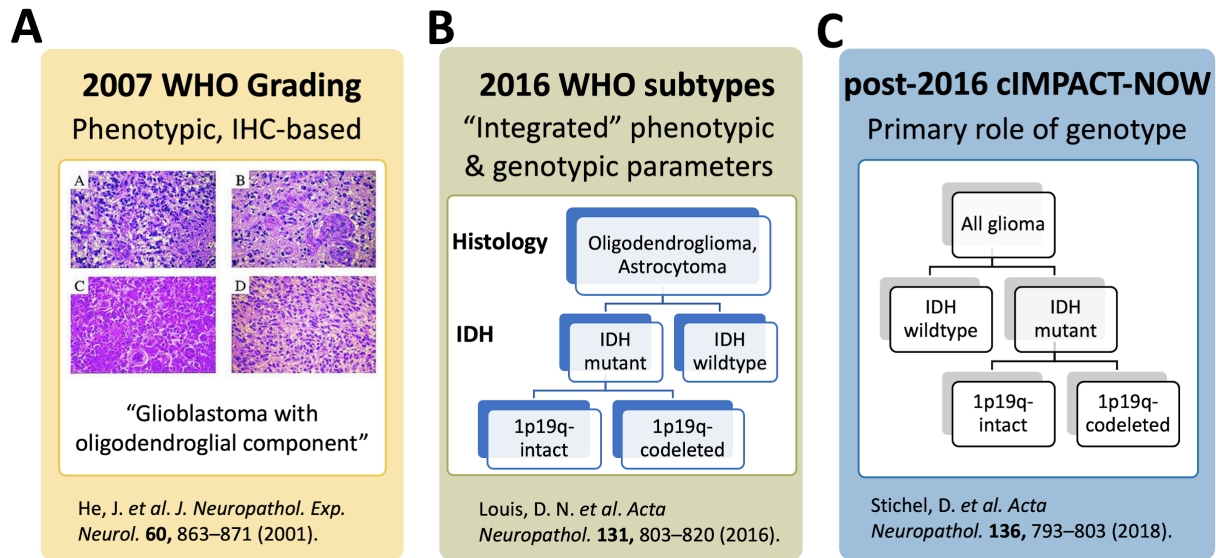
Historically, scientists and clinicians have relied on phenotypic pathological assessment of the tumor progenitor cell type (Figure 2.1a)[4] as well as other markers of proliferation, microvascularization, and necrosis. Oligodendrogliomas, or glioma arising from oligodendrocytes, tended to be lower in grade and slower-growing, with favorable survival outcome. Astrocytomas, or tumors arising from astrocytes, were categorized into prognostic grades using classic phenotypic histochemistry assessment: diffuse astrocytoma with “cytological atypia” as grade II; anaplastic astrocytoma with “anaplasia and mitotic activity” as grade III; and “microvascular proliferation and/or necrosis” as grade IV, otherwise known as glioblastoma.

In 2016, this categorization was drastically overhauled to comprise an integrated genotypic-phenotypic assessment (Figure 2.1b). The categorization still first considered pathological characterization of whether the sample was glioblastoma or not, and was then followed by establishment of the presence or absence of an IDH mutation. If the tumor was determined to have an IDH mutation, the sample would then be considered for the joint deletion of the 1p and 19q chromosomal arms.

In the past year, the cIMPACT-NOW consortium that informs the WHO has placed even greater emphasis on genetic classification of glioma (Figure 2.1C). In brief, the recommendations suggest reorganizing diffuse or anaplastic IDH wildtype tumors with EGFR amplification, whole chromosome 7 gain and 10 deletion, and/or TERT promoter mutation into the same category as “glioblastoma, grade IV, IDH wildtype,” despite the absence of histological markers such as microvascular proliferation or necrosis[5,6]. Similarly, numerous retrospective studies have concluded that histologic grading criteria may not stratify risk for patients with IDH-mutant astrocytomas[7–9]. cIMPACT-NOW suggests that tumors previously categorized as “glioblastoma, grade IV, IDH mutant” should be recognized as distinct from glioblastoma. In addition to histological characteristics of aggressiveness, CDKN2A/B homozygous deletion should be a WHO grade 4 criterion for IDH-mutant astrocytomas, and instead denoted as “IDH mutant, astrocytoma, Grade 4”.

To summarize, the evolving classification of glioma can be characterized by an increased emphasis on genetic alterations for proper diagnosis and prognosis. In the near future, we expect to categorize patients as IDH-wildtype (IDHwt), IDH-mutant 1p19q intact (IDHmut-intact), or IDH-mutant 1p19q-codeleted (IDHmut-codel). In addition, there exists more genetic alterations (e.g. CDKN2A/B homozygous deletion) that can more precisely denote the aggressiveness and

clinical course of a particular tumor, but remain to be implemented in routine clinical practice. These different categories are able to inform patient management in more detail than ever before.



**Figure 2.1. Evolving WHO categorization of glioma places greater emphasis on genetic subtype.**

(A) An example of histopathological evaluation that classified tumors based on the inferred tumor cell origin. (B) Half of the flowchart of the 2016 WHO integrated diagnosis, exemplifying how the classification first relied on histopathological evaluation of tumor cell of origin followed by IDH mutation status and 1p19q codeletion. The other half classifying glioblastoma is not shown. (C) The new proposed classification scheme, where all glioma regardless of tumor cell of origin are classified first by IDH mutation status followed by 1p19q codeletion status.

## 2.2 The challenges of glioma treatment

### 2.2.1 Treating brain tumors across the blood-brain barrier

The blood-brain barrier (BBB) is a vascular system unique to the brain, composed of endothelial cells strung together with tight junctions, pericytes to regulate them, and astrocytic endfeet[10]. In addition to the tight junctions between endothelial cells limiting transportation between cells, endothelial cells also exhibit reduced numbers of surface pores compared with

endothelial cells in other locations, as well as limited intracellular trafficking mechanisms[11,12]. The BBB is surrounded by basal lamina, an extracellular matrix (ECM) formed predominantly of glycoproteins that can be proteolytically cleaved to influence BBB function in health and disease[13]. These features work together in concert to tightly regulate brain homeostasis in healthy brains; however, these same features are what limit the delivery of therapeutics.

The tenets of the healthy BBB change when a neoplastic mass expands in the brain. This new mass of cells needs to be supplied with oxygen in order to continue its metabolic processes and replication. As the region becomes hypoxic, the cells upregulate vascular endothelial growth factor (VEGF), which as the name suggests, promotes blood vessel recruitment and development in a process called angiogenesis. This ad-hoc blood vessel formation will often lack the same integrity as the intact BBB, forming what is known as the blood-tumor-barrier (BTB). It is characterized by aberrant pericyte distribution, loss of astrocytic feet, and is generally considered “leakier”[10]. At first glance, this is advantageous for drug delivery; however, like most aspects of glioma, clinical data suggests heterogeneity with regard to the integrity of the BBB in all glioma, with regions of both intact and compromised BBB[14], adding complexity and unpredictability to the drug delivery process.

### 2.2.2 Treatment for Newly Diagnosed Patients

Treatment options for patients diagnosed with glioma are limited. First, by infiltrating one of the body’s most crucial organs, they can present in areas that even the most skilled neurosurgeons may have trouble resecting without impacting normal brain function. Next, the blood brain barrier will impede systemic therapeutic delivery because at best it is heterogeneously porous. Finally, the rarity of glioma means that there is little incentive from the

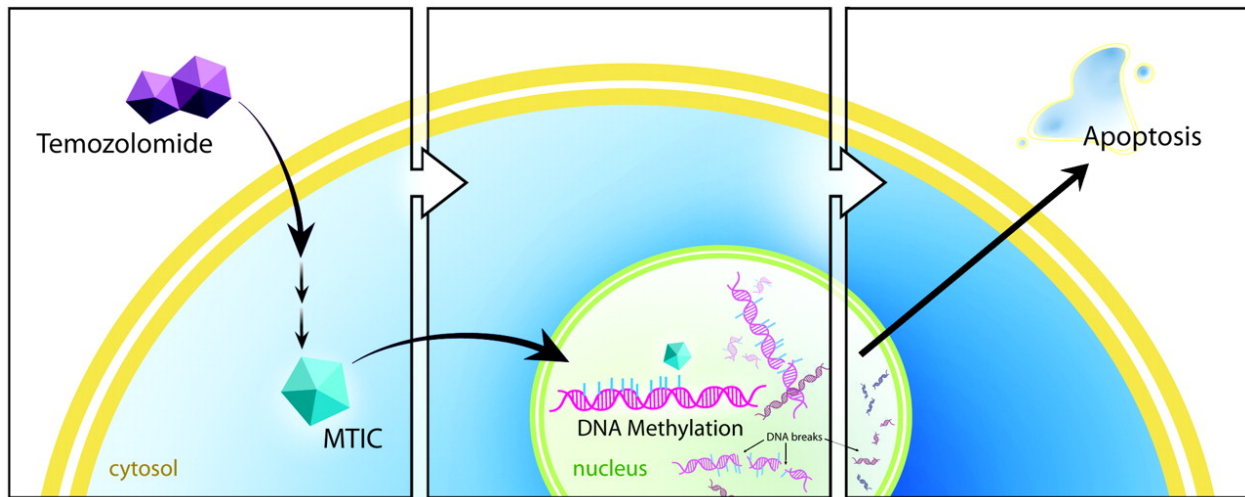
pharmaceutical industry to fund research due to low potential for profit. In light of these compounded challenges, it seems remarkable that scientists have found treatments that prolong the life of glioma patients.

The first line of treatment is surgical resection. It is recommended that all patients regardless of grade receive the greatest degree of surgical resection that can be undertaken safely[15]. However, removing the tumor is often complicated by the diffuse boundaries and the great importance of maintaining normal brain activity. In most cases, radiation therapy (RT) is subsequently employed to target the tumor cells that remain. Historically, doses of 50 to 60 Gy in 1.8- to 2-Gy fractions over 5 to 6 weeks was the mainstay of treatment after surgical resection; however, the utility of administering this dose immediately following resection has been brought into question for Grade 2 slow-growing glioma[15,16].

After surgery, chemotherapy in conjunction with RT can be beneficial in the treatment of patients with certain glioma. In 2005, temozolomide (TMZ) administered together with radiation therapy was found to increase survival of glioblastoma patients to a median of 14.6 months from a median 12.1 months with radiation therapy alone[17]. It was the first chemotherapy shown to produce any overall survival benefit in glioblastoma patients, and it induced minimal side effects. It was quickly adopted as the standard of care.

Temozolomide is lipophilic in nature, which allows it to cross the blood brain barrier. A high-level overview of its mechanism of action is demonstrated in Figure 2.2. Once converted to MTIC, it adds methyl groups to DNA or RNA at N7 and O6 sites on guanine and the N3 on adenine. During subsequent DNA replication, a thymine will be inserted opposite the methylguanine in place of a cytosine, inducing apoptosis[18]. However, the mismatch-repair (MMR) enzyme O6-methylguanine–DNA methyltransferase (MGMT) can reverse the effect of

the mismatched thymine by reinserting cytosine, invalidating the mechanism of the drug. So why does this drug work at all?



**Figure 2.2. Mechanism of Temozolomide.**

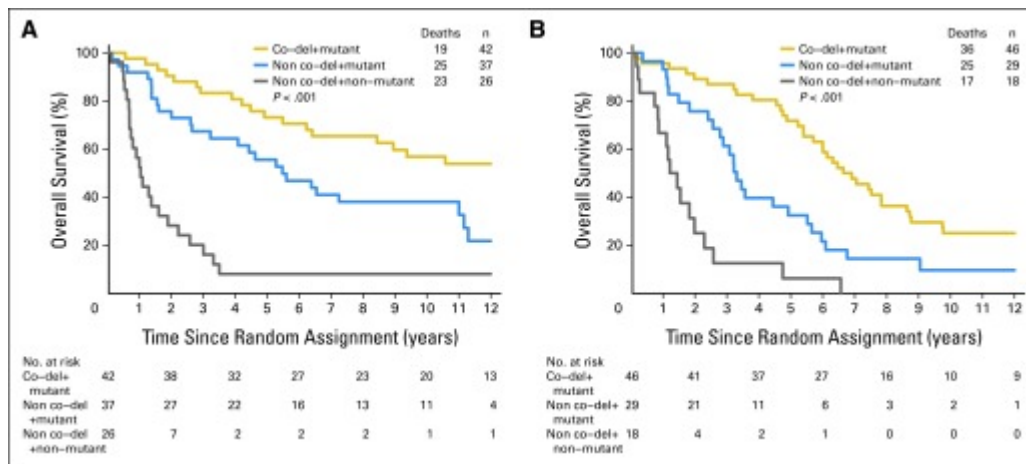
Lipophilic temozolomide can cross the cellular membrane where it is converted intracellularly into MTIC, which methylates DNA. If MGMT is inactive, cellular repair mechanisms cannot adjust, resulting in DNA nicks and ultimately apoptosis. Figure adapted from Wesolowski et al. 2010[19].

Interestingly, approximately half of patients diagnosed with glioblastoma exhibit methylation on the MGMT gene promoter. This methylation silences the *MGMT* gene, disrupting the mismatch repair mechanism and thereby sensitizing the patient to the effects of TMZ. Therefore, in these methylated patients, the TMZ chemotherapy is most effective. The results of a 2005 clinical trial demonstrate that in the population of patients that are not MGMT methylated, the median survival was 12.7 months among those assigned to TMZ and radiotherapy (RT) and 11.8 months among those assigned to RT alone. In contrast, MGMT methylated patients assigned to the TMZ+RT group had a median overall survival of 21.7 months compared with 15.3 months using RT alone[20], a significantly prolonged survival.

In addition to TMZ, procarbazine, lomustine, and vincristine (PCV) chemotherapy has been demonstrated to improve overall survival in patients with lower-grade, IDH mutated



tumors[21,22]. The overall survival benefit is even more dramatic in the subset of these patients that experience 1p and 19q codeletion in their tumors (Figure 2.3), improving median overall survival from 6.8 to 14.7 years. Though it is still unclear whether lower-grade IDHmut-1p19q codeleted tumors should be treated with PCV or TMZ, many practitioners have opted to use TMZ due to its favorable toxicity profile[23,24]. One notable caveat is that TMZ treatment in lower-grade patients can cause a “hypermuted” phenotype when the tumor recurs, which is associated with high-grade transformation, distant recurrence, and shortened survival[25]. Greater understanding of the genetic profiles of patients whose molecular mechanisms potentiate hypermutation could allow clinicians to selectively administer TMZ to patients at low risk for hypermutation.



**Figure 2.3. The benefit of PCV treatment for 1p19q codeleted patients.**

Kaplan-Meier estimates of overall survival for patients whose tumors were IDH mutated and 1p/19q codeleted (co-del; gold), mutated (mut) and noncodeleted (blue), and nonmutated and noncodeleted (gray) after (A) procarbazine, lomustine, and vincristine (PCV) plus radiotherapy (RT) and (B) RT alone. Median survivals after (A) PCV plus RT were 14.7, 5.5, and 1.0 years, respectively. Median survivals after (B) RT alone were 6.8, 3.3, and 1.3 years, respectively. Figure and caption adapted from Cairncross et al. 2014[22].

In 2015, the addition of tumor treating fields (TTF) was approved by the FDA as an optional adjuvant therapy to TMZ following RT for the standard of care in Grade 4 Glioblastoma.

In brief, TTFs are alternating electrical fields designed to interfere with normal mitosis that are delivered through a wearable helmet. The results of one clinical trial show marked increases in 2-year survival rate, which was 43% with TTF vs. 29% without TTF, along with increased median overall survival to 19.6 months with TTF device treatment vs. 16.6 months without TTF device treatment (HR = 0.744, p = 0.0038)[26]. Despite evidence of the safety and efficacy of the TTF device, skepticism remains persistent among neuro-oncologists[27]. Its usage remains infrequent (3-12% of patients with newly diagnosed GBM) and hinges on the compliance of the patient: patients with a compliance rate of >90% showed a prolonged median survival of 24.9 months and a five-year survival rate of 29.3%, while patients that complied <30% of the time with TTF had median survival of 18.2%[28]. Ultimately, such levels of compliance require lifestyle modifications that could dramatically impact the quality of life for patients with limited survival, and may be an explanation for low patient use.

The majority of these treatment options are differentially administered based on the tumor subtype, now determined by the genetic alterations of the tumor. But how do clinicians know the genetic alterations of a patient before selecting a chemotherapy regimen? The tissue acquired during resection can be evaluated for the aforementioned genetic alterations through histopathology, FISH and/or genetic sequencing. From this information, clinicians are armed with the necessary information to plan treatment.

However, tumors in the basal ganglia, thalamus, and brain stem cannot be resected without vastly damaging quality of life. Compounding these challenges, genetic sequencing and immunohistochemistry are costly and time-consuming expert-mediated processes. In cases where a patient cannot undergo surgical resection or immediate therapeutic decisions could improve patient outcome, clinicians could ideally obtain this essential genetic information through other

means to ensure that the appropriate therapeutic strategy is used for the patient. In Chapter 5, we explore the potential to presurgically identify genetic subtypes of newly-diagnosed glioma.

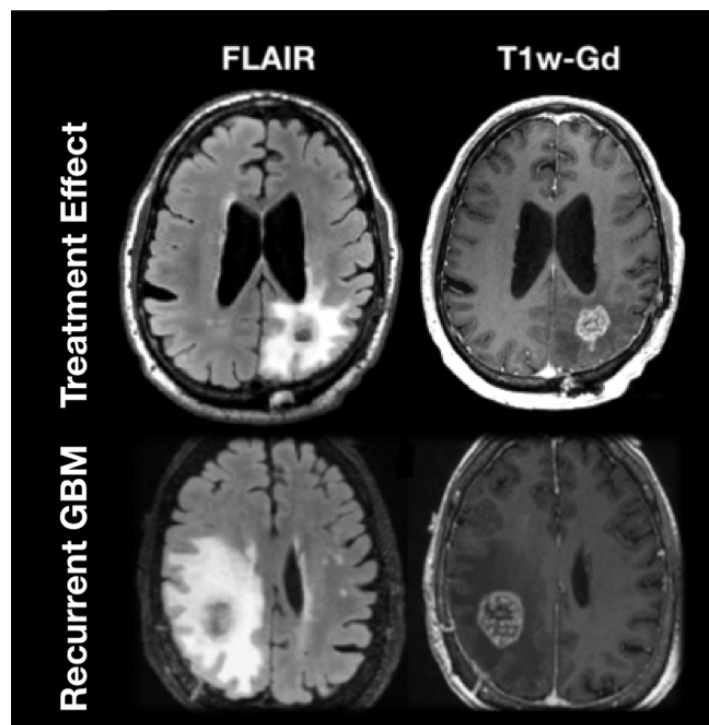
### 2.2.3 Treatment of Recurrent Glioma

There is no known curative therapy for glioma, and therefore these tumors will all eventually recur. Unfortunately, treating recurrent tumors is even more complex than treating newly diagnosed tumors. One consideration is whether a repeat resection is worth the impact on a patient's quality of life. Another complicating factor is that residual tumor cells remaining after first-line therapy are the cells that propagate recurrence, meaning they can be resistant to the original therapies, including radiation. Therefore, it is necessary that clinicians are armed with additional strategies when treating recurrent tumors.

One chemotherapy typically introduced upon recurrence is bevacizumab. Under hypoxic conditions such as that in a growing body of tumor cells, Vascular Endothelial Growth Factor (VEGF) transcription is upregulated[29]. True to name, this growth factor will stimulate recruitment and proliferation of endothelial cells toward the hypoxic mass, enabling the delivery of oxygen. The current standard of care for recurrent tumors includes Avastin, or bevacizumab, a humanized anti-VEGF monoclonal antibody. It selectively binds circulating VEGF so that it cannot conduct downstream signaling, thereby limiting the blood and oxygen supply to tumor tissues. Bevacizumab can be especially effective in combination with RT, as it can sensitize tumor endothelium to RT[30], and results from a 2013 trial support that this combination can meaningfully extend progression free survival (PFS) and overall survival (OS)[31].

During and after the completion of original treatment for Grade 4 Glioblastoma, patients are frequently scanned with magnetic resonance imaging (MRI) to search for evidence of tumor recurrence. Unfortunately, RT often complicates this evaluation period by causing damage to the

damage to the surrounding healthy brain tissue, inducing a new lesion that is not caused by tumor growth. These treatment-induced effects can appear identical to a true recurrence of the tumor both on conventional anatomic imaging (Figure 2.4) and during clinical symptom evaluation. Even further, an estimated 36% of patients experience these treatment-induced effects[32], and its appearance is even more common with the recent advent of immuno- and other targeted therapies in clinical trials. If recurrence is incorrectly diagnosed, a patient may be removed from an effective therapy, which could invalidate the results of a clinical trial or expose a patient to unnecessary surgical intervention. Taken together, this phenomenon presents an enormous challenge for planning treatment when a suspected recurrence appears.



**Figure 2.4. The similarity of treatment-induced lesions and recurrent, high-grade glioma on convention MR imaging.**

Treatment-induced lesions and true recurrence can both exhibit extensive T2-FLAIR hyperintensity, with contrast enhancement toward the central region of the lesion.

## 3. Magnetic resonance imaging of glioma

### 3.1 Conventional anatomic imaging sequences

#### 3.1.1 Introduction to MRI

Magnetic resonance imaging, or MRI, is a powerful imaging modality that produces images of internal physical and chemical characteristics of an object from externally measured nuclear magnetic resonance signals[33]. MRI is incredibly flexible, offering a wide range of physical parameters to image and instrumental parameters to set for control of image content[34]. This section will focus on gaining intuition about the basics of MRI: how magnetization is used to create and record signals emitted from the body, and how this signal acquired over time can be converted into a spatial (2D or 3D) image.

Magnetic resonance can be achieved by any atom with an odd number of protons and/or odd number of neutrons. By far the most studied MR imaging technique is hydrogen ( $^1\text{H}$ ) imaging because it is the most abundant atom in the body and therefore gives rise to the largest signals[34]; for the rest of this introduction, we will discuss these phenomena with respect to the nuclei of  $^1\text{H}$  atoms, or a single proton, with the knowledge that the principles hold for any odd numbered nuclei in the body.

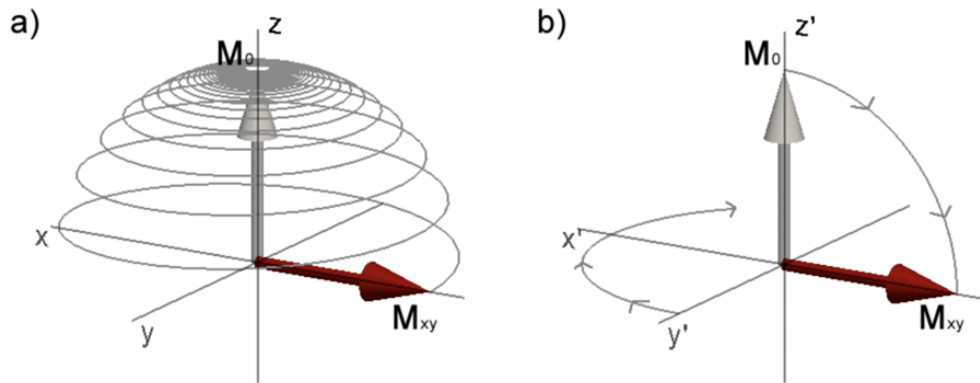
Protons are subatomic particles with a positive charge that precess (or “spin”) with angular momentum. This lends them an electrical current which is accompanied by a magnetic field. When these protons are placed in a strong external magnetic field ( $B_0$ ) like that generated by a modern MR scanner (0.5 - 7 Tesla), they will align parallel or antiparallel with  $B_0$ . Because of the energy differential between parallel and antiparallel states, there are slightly more protons

aligned parallel to  $B_0$  generating a net longitudinal magnetization ( $M_z$ ) along the direction of the patient! However, since it is so minimal compared to the external magnetic field, it is not possible to measure this net magnetization. So how do we measure any signal after all?

Recall that the atomic nuclei are continuously precessing, even once aligned parallel or antiparallel within the magnet. The frequency  $f$  at which the nuclei are precessing can be calculated with respect to the external magnetic field with the following equation:

$$f = \frac{\gamma}{2\pi} B_0 \quad (3.1)$$

where  $\gamma$  is the gyromagnetic ratio and is specific to each element ( $\frac{\gamma}{2\pi} = 42.58\text{MHz/T}$  for  $^1\text{H}$  atoms). Notably, the frequency of the spinning varies directly with an increase in the strength of  $B_0$ . When a radiofrequency (RF) wave oscillating at the Larmour frequency is applied ( $B_1$ ) transverse to the direction of  $B_0$ , the exchange of energy from the RF pulse to the precessing protons can occur due to *resonance*. This excitation will a) knock some protons that were parallel to  $B_0$  into antiparallel configurations; b) knock the protons' precession into the transverse plane, described in Figure 3.1; and c) synchronize the precession of protons such that they are "in phase" with one another, establishing transverse magnetization ( $M_{xy}$ ). This rotating transverse magnetization induces an electromotive force in a receiver coil oriented to detect the changes of magnetization in this transverse plane, which is what will ultimately generate the image.



**Figure 3.1. Excitation and magnetization of a proton depicted in (a) the lab frame and (b) the rotating frame.**

Image courtesy of Kleinnijenhuis, 2014.

Following this excitation, both the longitudinal and transverse components of the magnetization decay and return to equilibrium state  $M_0$ . The longitudinal relaxation is characterized by the energy difference between  $M_z$  to  $M_0$  released into the lattice of the spin system, an exponential process mediated by the time constant  $T_1$  (“spin-lattice” relaxation), or the time it takes for 63% of longitudinal magnetization to recover. Transverse relaxation is governed by interactions between protons and their surrounding spins (“spin-spin” relaxation) such that they are no longer precessing in sync. It is governed by the time constant  $T_2$ , which is the time it takes for 37% of the original signal  $M_0$  to remain in the transverse plane. Both  $T_1$  and  $T_2$  vary based on tissue composition. This evolving magnetization can be described by the Bloch equations[35]:

$$\frac{dM_x}{dt} = \gamma M_y \times B_z - \frac{M_x}{T_2} \quad (3.2)$$

$$\frac{dM_y}{dt} = \gamma M_x \times B_z - \frac{M_y}{T_2} \quad (3.3)$$

$$\frac{dM_z}{dt} = \gamma M_z \times B_z - \frac{M_z - M_0}{T_1} \quad (3.4)$$

or in combined vector form:

$$\frac{dM}{dt} = \gamma M \times B - \frac{M_x i - M_y j}{T_2} - \frac{(M_z - M_0)k}{T_1} \quad (3.5)$$

where where  $i, j$  and  $k$  are unit vectors in the  $x, y$  and  $z$  directions. An excitation with 90 degree flip angle in the rotating frame as described in Figure 3.1 simplifies the solutions to the Bloch equations such that the equation describing longitudinal relaxation becomes:

$$\frac{dM_z}{dt} = \frac{M_0 - M_z(t)}{T_1} \quad (3.6)$$

which yields the expression:

$$M_z(t) = M_0(1 - e^{-\frac{t}{T_1}}) + M_z(0) \cdot e^{-\frac{t}{T_1}} \quad (3.7)$$

which can more intuitively describe the evolution of longitudinal magnetization. In the same vein, the solution for transverse magnetization can be simplified to:

$$\frac{dM_{xy}}{dt} = -\frac{M_{xy}}{T_2} \quad (3.8)$$

and has the solution:

$$M_{xy}(t) = M_{xy}(0) \cdot e^{-\frac{t}{T_2}} \quad (3.9)$$

[36]. The rotating transverse magnetism will release a signal over time that is known as free induction decay (FID), inducing an electromotive force in a coil oriented to detect changes in the  $xy$  plane.

Because the  $T_1$  and  $T_2$  relaxation times are intrinsic to tissue composition, measuring the free induction decay signal generated from exciting these tissues can allow for contrast between

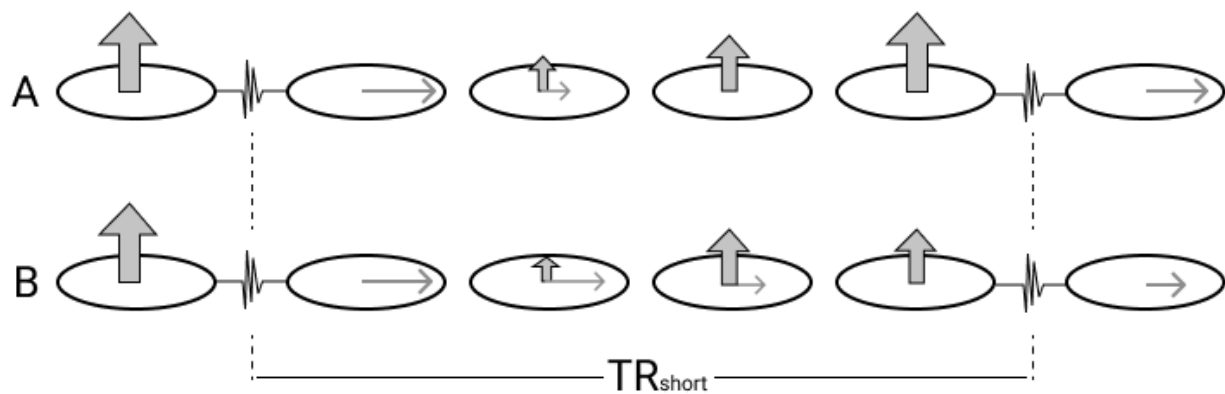


different tissue types. The strength of the MR signal generated from a particular tissue is given by the following equation:

$$S = K \cdot [H] \cdot \left(1 - e^{-\frac{TR}{T_1}}\right) \cdot e^{-\frac{TE}{T_2}} \quad (3.10)$$

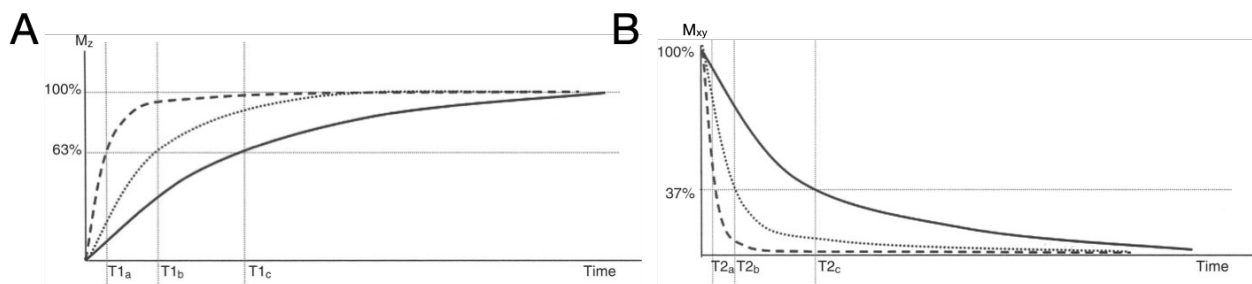
where K is a scaling constant, [H] is the density of protons in a particular area, TR is the time to repetition of the original RF pulse, and TE is the time to echo - or the time it takes to have a repetition of the focused magnetization vector before repeating the original RF pulse. In the simplest spin-echo sequence, TE is mediated by a second 180 degree RF pulse that is given at TE/2 that effectively acts as a wall that the signal bounces off of and refocuses at TE. From Equation 3.10, we notice that TR mediates the weighting of T1 differences between tissues, and TE mediates the weighting of T2 differences between tissues.

To understand the effects of TR on T1 weighting, consider what happens to two tissues A and B when two 90 degree RF pulses are given in succession quick enough (i.e. short enough TR) such that all the protons knocked antiparallel in Tissue B do not have time to recover to their original thermal equilibrium (Figure 3.2b), while Tissue A does fully recover (Figure 3.2a). Upon the second 90 degree RF pulse, there will be a lower magnitude of transverse magnetism, because there was not enough time for  $M_z$  to return to  $M_0$ , and the magnitude of the FID will be diminished. In this scenario, Tissue B has a longer T1 recovery time, which can be represented by the solid line in Figure 3.3a, while Tissue A has a shorter recovery time represented by the dotted black line Figure 3.3a. As a concrete example, in a T1-weighted image, tissues with shorter recovery times (e.g. fat, myelin in white matter) will have greater magnitude of the transverse magnetization at TR compared with tissues with longer T1 recovery times (e.g. water, cerebrospinal fluid), creating an image that appears brighter in fat-dense areas compared with water-dense areas.



**Figure 3.2. Example of magnetization differences between tissues A and B when repetition time (TR) is short.**

(A) Tissue A has shorter recovery time, and when excited again has greater transverse magnetization compared with (B) Tissue B, which takes longer to recover and has lower transverse magnetization upon the time to repetition (TR). Image adapted from Schild 1990 [37].



**Figure 3.3. Example tissue recovery times of (A) longitudinal and (B) transverse magnetization after an RF pulse.**

Image adapted from Lipton et al. [38].

For a discussion regarding the accentuation of T2 recovery time differences, it is important to note that within biological systems, the decay of transverse magnetism with T2 is impacted by the inhomogeneities of the magnetic field in addition to the local impact from spin to spin. These static field variations contribute to additional, nontrivial dephasing with time constant T2', for a final time constant T2\* described by the following equation:

$$\frac{1}{T2^*} = \frac{1}{T2} + \frac{1}{T2'} \quad (3.11)$$

resulting in even faster signal decay. The decay between different tissues and their resulting T2 constants is depicted in Figure 3.3b. To understand how to accentuate differences in T2 recovery times between two tissues, consider what happens when TE is short (e.g. before T2a on Figure 3.3b). The refocused signal echo would occur before sufficient T2 decay has taken place, and the tissues could not be differentiated. Therefore, long TE is necessary to allow sufficient time for differentiating between the T2 recovery times intrinsic to two distinct tissue types. The longer the T2 recovery time of a substance, the greater magnitude of signal upon refocusing; therefore, substances with longer T2 recovery times (e.g. water) will appear bright compared to substances with shorter T2 recovery times on T2-weighted images.

In practice, T1 and T2 recovery times influence the magnitude of the signal generated by a specific tissue. In order to generate a truly T1-weighted image, it is therefore not enough to have long TR - it is also necessary to minimize the effects of T2 recovery time by also including a long TE. In the same way, T2-weighted images are generated by having short TE accompanied by a short TR to minimize the effect of T1 recovery time. Referring to Equation 3.10, choosing a long TR and a short TE would result in the signal being largely dependent on  $[H]$ , and is appropriately named proton density weighted imaging, while short TR and long TE generates poor contrast and is not used at all. The choices of TR, TE and the impact on final image contrast are summarized in Table 3.1 and illustrated in Figure 3.4.

Table 3.1. Effects of acquisition parameters on image contrast.

<b>Image Weighting</b>	<b>Accentuated relaxation</b>	<b>Diminished relaxation</b>	<b>TE values</b>	<b>TR values</b>
<i>T1-weighted</i>	T1	T2	Short	Short
<i>T2-weighted</i>	T2	T1	Long	Long
<i>PD-weighted</i>	None	T1 and T2	Short	Long

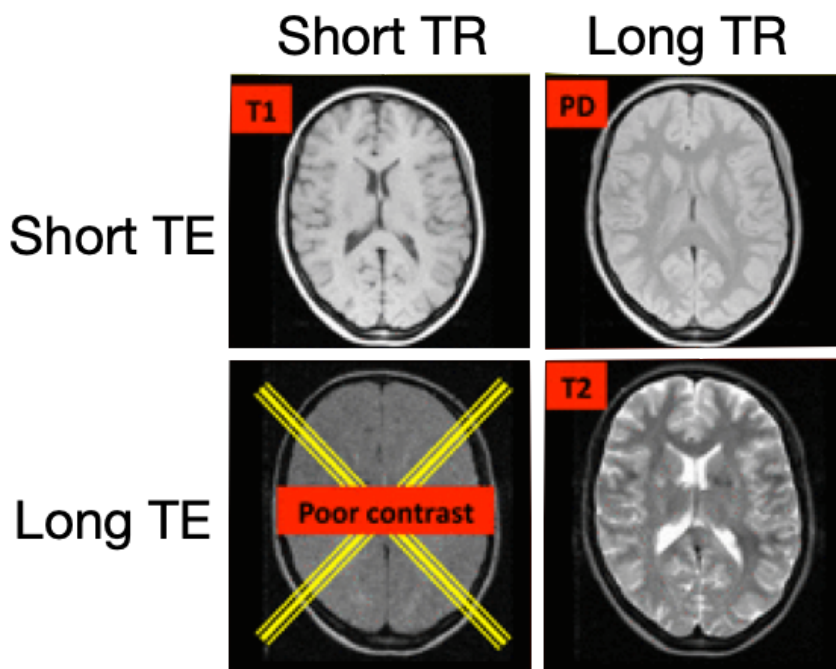


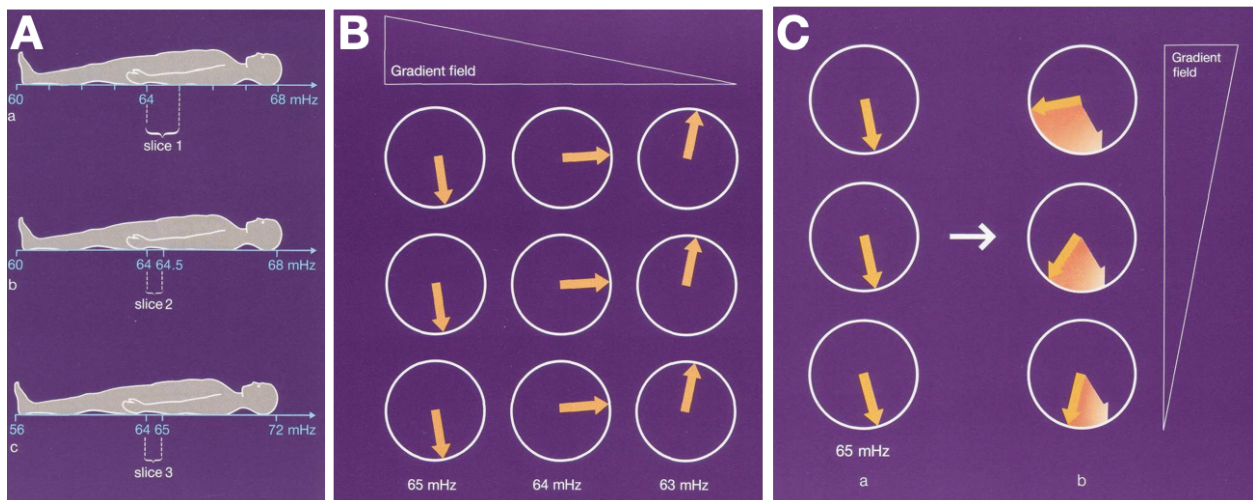
Figure 3.4. Effects of acquisition parameters on image contrast.

Image adapted from Elster, 2021 [39].

### 3.1.2 From signal to image

The previous section described how to leverage intrinsic differences in tissue properties in order to generate characteristic signals, but the question remains: how does one create an image from these signals? Localizing signal received by the coil requires the application of magnetic gradients, or spatially varied magnetism in all three directions. The basic principles of gradient application as it applies to image acquisition will be introduced in this section.

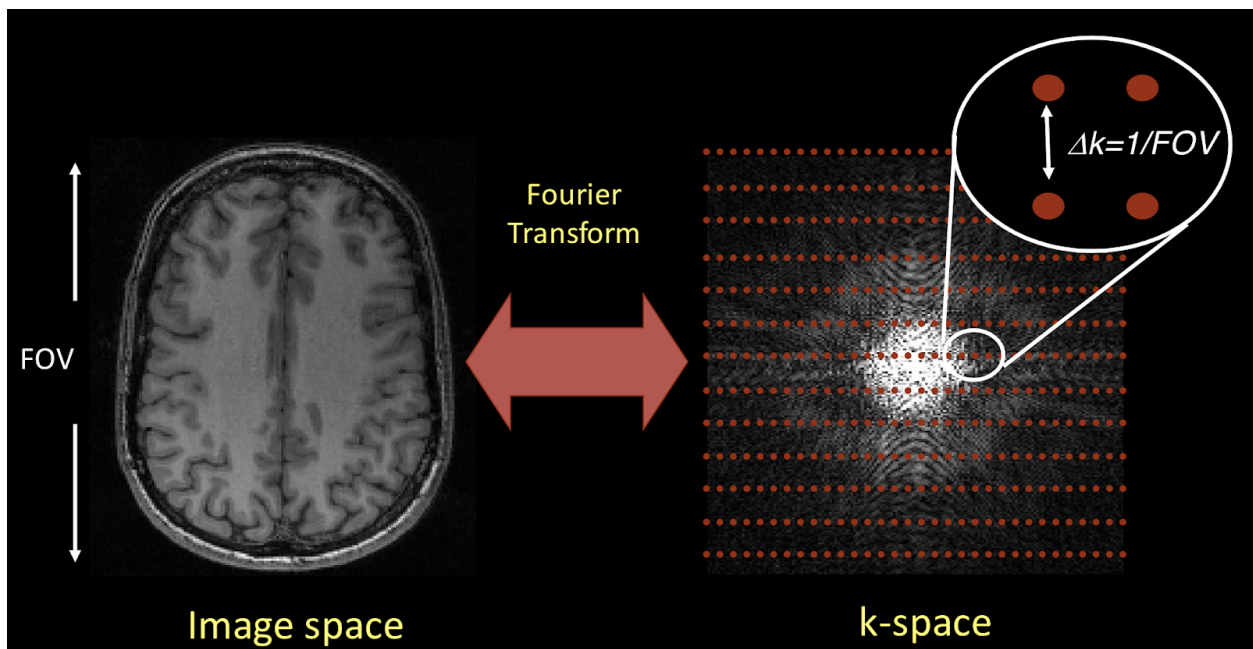
In MRI, the main static magnetic field  $B_0$  is complemented by three orthogonal gradients  $G_x$ ,  $G_y$ , and  $G_z$ , each of which is a linear function of spatial position. Conventionally, the  $G_z$  gradient determines the slice of the body that will be imaged in the axial plane. The thickness of the slice is determined by the range of frequencies that is delivered by the RF pulse (Figure 3.3a); the wider the range of frequencies, the thicker the slice; however, this is also determined by the slope of the linear  $G_z$  gradient. Once the slice is chosen, the remaining two directions are encoded by 1) a frequency encoding gradient (Figure 3.3b) and 2) a phase encoding gradient, demonstrated by Figure 3.5c.



**Figure 3.5. Localizing signal in MRI using (A) a slice selection gradient; (B) frequency encoding gradient; and (C) a phase encoding gradient.**

Image adapted from Schild, 1990 [37].

Consequently, every point in the targeted imaging volume assigned a unique frequency at which the protons in that space precess. Phase encoding gradients are repeated at multiple amplitudes for each frequency encoded line that is read out until an entire plane of k-space, or frequency space, is acquired. Each point in k-space is thus assigned a unique frequency that can be mapped to its corresponding position in the image via an inverse Fourier transform (Figure 3.6). The field of view (FOV) of an image is inversely related to the distance between sampled points in k-space.



**Figure 3.6. Relationship of signal acquired in k-space (time domain) to an MR image (spatial domain) via the inverse Fourier transformation.**

Image courtesy of Dr. Janine Lupo.

### 3.1.3 The role of conventional anatomic imaging in brain tumor diagnosis

A typical brain MRI consists of T1-weighted and T2-weighted MR imaging sequences. Compared with other tomographic imaging techniques such as CT, these MRI techniques offer enhanced soft tissue contrast. The primary role of these images is to determine the lesion location, extent of tissue involvement, and resultant mass effect upon the brain, ventricular system, and vasculature [40].

T1-weighted imaging is typically acquired before and after the administration of an intravenous gadolinium (Gd) based contrast agent. Gadolinium ions possess 7 unpaired electrons which can interact with water nuclei via dipole-dipole interactions and distort the local magnetic fields, affecting both T1 and T2 recovery times. When a Gd-based contrast agent is delivered systemically, it can extravasate in places where the blood-brain barrier is compromised by tumor growth, in turn shortening the T1 and making voxels appear bright on T1-weighted imaging. Breakdown of the blood-brain barrier as visualized by the contrast enhancement on T1-weighted imaging is positively associated with tumor aggressiveness, and its presence and extent are key components of glioma diagnosis. However, enhancement does not always represent active tumor regions and can often be present due to BBB disruption as a result of radiation treatment.

Regions of increased peritumoral edema cause visibly hyperintense regions on T2-weighted imaging. In gliomas, this is typically a mixture of vasogenic and infiltrating tumor cells along white matter tracts [40]. In many lower-grade gliomas that do not enhance, this is the primary signal by which clinicians can diagnose a tumor. The size and location of the T2 hyperintense region are useful for properly diagnosing glioma.

Despite providing the foundation of patient diagnosis, there exists abundant heterogeneity in tissue composition, metabolism, and mutational burden that cannot be fully characterized with

structural imaging methods alone. Therefore, additional MR techniques that capture specific physiologic and metabolic tumor characteristics are complementary for the accurate diagnosis of brain tumors from presurgical imaging.



## 3.2 Diffusion MRI

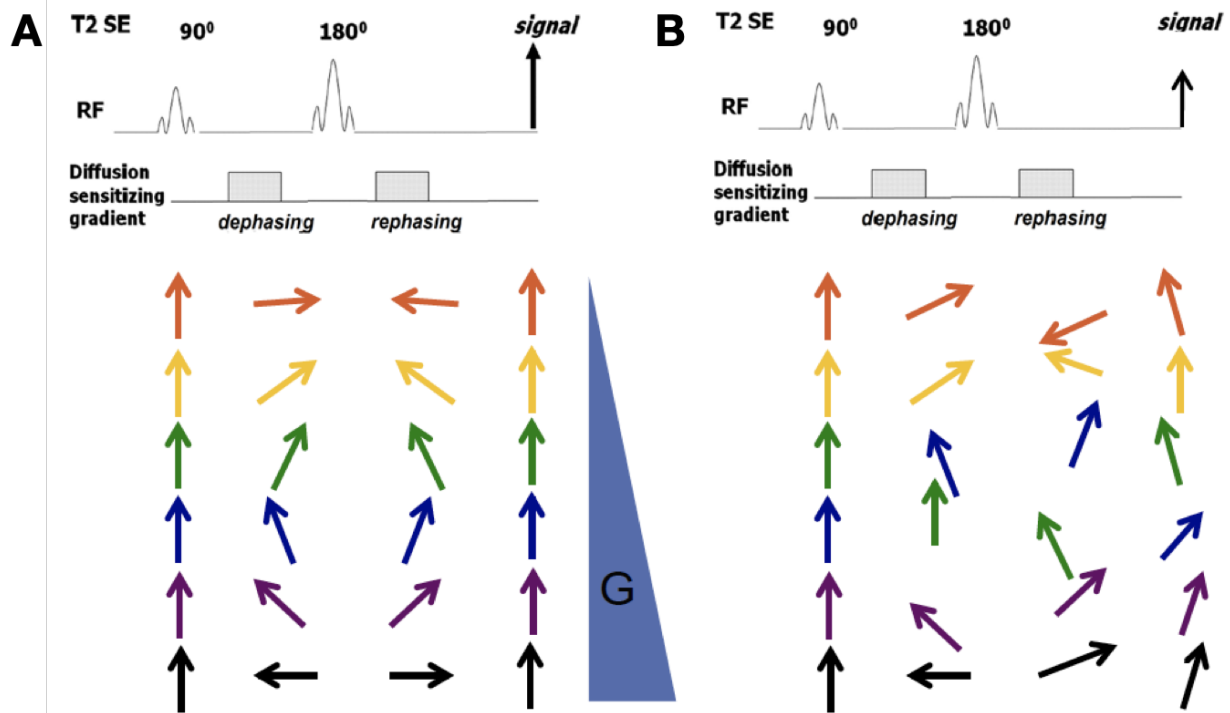
### 3.2.1 Introduction to diffusion-weighted and diffusion-tensor imaging

Diffusion-weighted imaging (DWI) is an MR technique that probes the random (Brownian) intra-voxel motion of water molecules. A basic schematic of the MR pulse sequence that can capture Brownian motion is depicted in Figure 3.7: T2-weighted spin-echo sequence of a 90 degree/180 degree RF pulse sequence together with diffusion sensitizing gradients can rephase protons perfectly in the scenario of zero proton movement (Figure 3.7a). In contrast, when there is increased water motion (Figure 3.7b), the protons cannot be perfectly rephased after a 180 degree pulse, resulting in diminished signal. These signal differences between voxels can differentiate areas where water is restricted and water can move freely.

The diffusion sensitizing gradients are typically described with a single value  $b$  value measured in  $\text{s}/\text{mm}^2$  that can be characterized by the magnitude of the gradient ( $G$ ), the time between the dephasing and rephasing gradient ( $\Delta$ ), and the duration of the gradient application itself ( $\delta$ ), shown in the following equation:

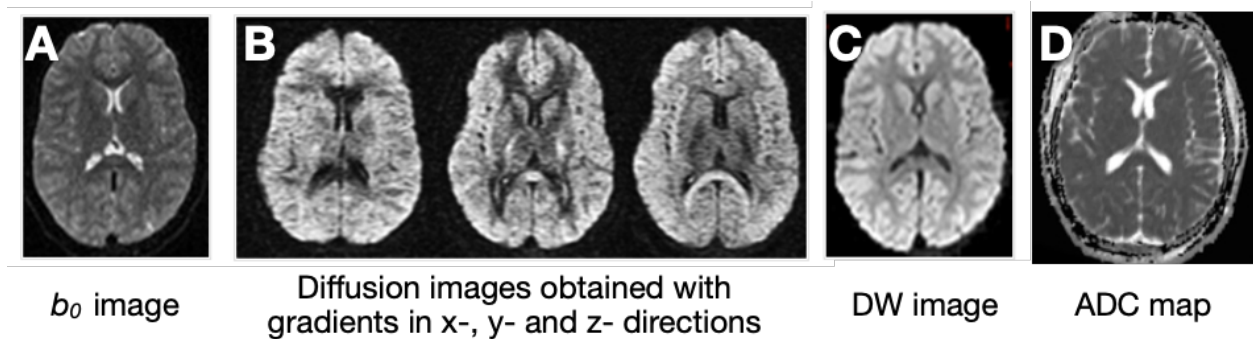
$$b = (\gamma G \Delta)^2 \left( \Delta - \frac{\delta}{3} \right) \quad (3.12)$$

When  $b=0 \text{ s}/\text{mm}^2$ , the resulting image is simply the T2-weighted image described in 2.1.1. These are acquired during DWI sequences to serve as a baseline. Typical  $b$  values for diffusion-weighted images range between  $500 \text{ s}/\text{mm}^2$  and  $2000 \text{ s}/\text{mm}^2$ , with larger  $b$ -values increasing the signal from diffusion in the image. The tradeoff, however, is that there is also greater noise, which can compromise the signal-to-noise ratio.



**Figure 3.7. Schematic of basic pulse sequence sensitive to the diffusion of water molecules within tissues.**

(A) When there is little brownian motion, protons can dephase and rephase with this gradient to recover greater signal; (B) when there is motion of water molecules, the dephasing and rephasing gradients can't realign protons to recover as much signal. On a purely diffusion-weighted image, this means that areas and tissues with greater diffusion experience MR signal loss, appearing darker. Image courtesy of Qiuting Wen.



**Figure 3.8. The creation of diffusion-weighted images and ADC maps from three directions.**

(A) The first image obtained without diffusion gradients applied; (B) x-, y-, and z-direction diffusion-weighted images ( $S_x$ ,  $S_y$ ,  $S_z$ ); (C) the combination of three images in (B) via Equation 3.16; (D) obtaining the ADC map through Equation 3.17.

To create a diffusion-weighted image, a minimum of three diffusion-sensitizing gradients must be turned on in three orthogonal directions; typically, they are in the x, y, and z directions. These images will create diffusion-weighted source images sensitized to diffusion in different directions (Figure 3.8). The relationship between the signal acquired from applying b in the x, y, and z direction ( $S_x$ ,  $S_y$ , and  $S_z$  respectively) to the signal acquired at  $b=0$  ( $S_0$ ) can be described by the following equations:

$$S_x = S_0 e^{-bD_{xx}} \quad (3.13)$$

$$S_y = S_0 e^{-bD_{yy}} \quad (3.14)$$

$$S_z = S_0 e^{-bD_{zz}} \quad (3.15)$$

where D is the x-, y-, or z-directionally-specific diffusion coefficient. They often combined using the geometric mean to produce a diffusion-weighted image (sometimes called “isotropic”) through the following formula:

$$S_{DWI} = \sqrt[3]{S_x S_y S_z} \quad (3.16)$$

which is exemplified in Figure 3.8c.

In these images, regions with increased diffusion appear dark and restricted diffusion appear bright regions. However, it can be more intuitive to display these properties inversely. In addition, lesions with either very long or very short T2 values might suffer from phenomena known as “T2-shine through.” The calculation of the apparent diffusion coefficient (ADC) map is often helpful for both of these, and can be calculated using the following equation:

$$ADC = -\frac{1}{b} \ln\left(\frac{S_{DWI}}{S_0}\right) \quad (3.17)$$

Through the ADC calculation, a pure parametric image is created which removes the confusion generated from T2-shine through and inverts the contrast such that increased diffusion appears bright, while restricted diffusion appears dark. An example of an ADC map is given in Figure 3.8d.

Diffusion tensor imaging (DTI) is an extension of diffusion weighted imaging, where the diffusion of water within a voxel is represented with a tensor with three orthogonal directions called eigenvalues( $\lambda_1, \lambda_2, \lambda_3$ ), that are not specifically in the x, y and z direction of the three gradients applied in DWI. Briefly, the full diffusion tensor is represented by a 3x3 symmetric matrix that must be fully sampled ( $D_{xx}, D_{xy}, D_{xz}, D_{yx}, D_{yy}, D_{yz}, D_{zx}, D_{zy}, D_{zz}$ ); but because  $D_{xy}=D_{yx}$ ,  $D_{xz}=D_{zx}$  and  $D_{yz}=D_{zy}$ , gradients are applied at a minimum of six directions. The shape of the tensor representing diffusivity within the voxel therefore also describes the extent of anisotropy (which can be thought of as “directionality”) within a voxel. This metric, dubbed fractional anisotropy (FA), can be calculated using following equation:

$$FA = \sqrt{\frac{(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_1 - \lambda_3)^2}{2(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)}} \quad (3.18)$$

where the values of FA vary between 0 and 1. In perfectly isotropic voxels,  $\lambda_1 = \lambda_2 = \lambda_3$  and FA = 0. As diffusivity becomes increasingly specific to one direction (i.e. the ellipsoid becomes more and more elongated and narrow),  $FA \rightarrow 1$ .

### 3.2.2 The role of diffusion-weighted and diffusion tensor imaging in glioma

From this description, it is intuitive to imagine how different tissue types might vary with respect to how free or restricted the movement of water might be. In brain tumors, areas of increased tumor cellularity could restrict diffusion, while peritumoral edema could increase diffusivity. Given that tumor growth patterns and edema are the physiological manifestation of

the underlying genetic and epigenetic biology of the tumor cells, it reasons then that ADC measured within the different structural regions of the tumor could be associated with certain genetic subtypes, be prognostic for overall survival, and/or an early marker of recurrence. On top of these potential diagnostic and prognostic applications, the ability to map white matter tracts and eloquent regions has made FA maps an indispensable tool for neurosurgeons[41]. As mounting evidence suggests that extent of resection is the most important prognostic factor for glioma patients of any grade and subtype [42], it has become imperative to remove as much tumor as possible while ensuring the postsurgical retention of motor and communication skills of the patient.

One 2015 study from Wen et al. provides evidence that even when adjusting for clinical prognostic factors, metrics derived from the ADC in the T2-hyperintense region of newly diagnosed glioblastoma is strongly associated with OS ( $p < 0.001$ )[43]. Another simultaneous report uses a slightly different ADC metric but comes to similar conclusions supporting ADC's prognostic value[44]. Rather than OS, Pope et al. report a 2.75-fold reduction in median time to progression when stratifying patients into two folds by one threshold derived from ADC, supporting ADC's prognostic utility[45]. What's more, multiple studies provide evidence that metrics derived from ADC and FA maps are predictive of both IDH mutation and 1p19q codeletion in glioma, the two most important genetic alterations that influence treatment response and OS[46–48].

ADC and FA have also been extensively studied with regard to their ability to differentiate true glioblastoma recurrence from treatment-induced damage (Section 2.2.3). Most studies calculate a threshold from the mean or histogram-related metric from an anatomically delineated region of interest (ROI) and normalize that value against a contralateral normal

appearing white matter region. Though at least 10 studies reporting significant differences in these metrics between patients experiencing recurrent tumor and those experiencing the effects of treatment, they must be qualified: 1) some include as few as 15 patients; 2) each study uses different histogram-related metrics to create a threshold; 3) studies define patients with treatment effect and true recurrence inconsistently; 4) thresholds are often created from and then applied to the same study samples; and 5) thresholds between studies are highly variable due to varying acquisition parameters and lack of normalization[49–58]. Ultimately, using values derived from ADC and FA maps in this way will be difficult to standardize across institutions, as they are sensitive to the diffusion b-value and  $B_0$  magnetic field unless normalization is performed. More recently, a 2019 study reports that incorporating features derived from the radiomics analysis of diffusion-weighted imaging improves diagnostic performance compared with anatomic imaging alone when incorporated into a machine-learning model [59]. The results from this study provide a promising alternative to harness the rich information in ADC and FA maps to solve this challenging problem.

## 3.3 Perfusion MRI

### 3.3.1 Introduction to perfusion MRI

Perfusion is defined as the amount of blood delivered to the capillary beds of a tissue volume over a certain period of time. Measuring perfusion can be used to approximate microvascular density or as an indication of how efficiently oxygen and other nutrients are being delivered in tissue. Perfusion MRI denotes the set of MRI techniques developed to measure perfusion, including arterial spin labeling (ASL), dynamic susceptibility contrast (DSC), and dynamic contrast enhancement (DCE). In both DCE and DSC methods, an intravascular contrast agent is injected into the blood in order to measure blood volume and permeability, while ASL requires no exogenous agents to measure blood flow. In both exogenous cases, a paramagnetic tracer will travel through a capillary network, inducing transient changes in the local magnetic field of surrounding tissue. These changes can be captured over time using fast MRI sequences. They are subsequently mapped to create concentration-time curves within voxels, from which the following useful metrics can be calculated:

- Cerebral blood flow (CBF): the rate at which blood flows through the microvasculature in a tissue region, calculated in  $\text{mL}(\text{tissue grams}^{-1})(\text{sec}^{-1})$
- Cerebral blood volume (CBV): the fraction of the volume of tissue occupied by blood
- Mean transit time (MTT): average time that blood spends passing through a region of tissue before exiting through the venous system.

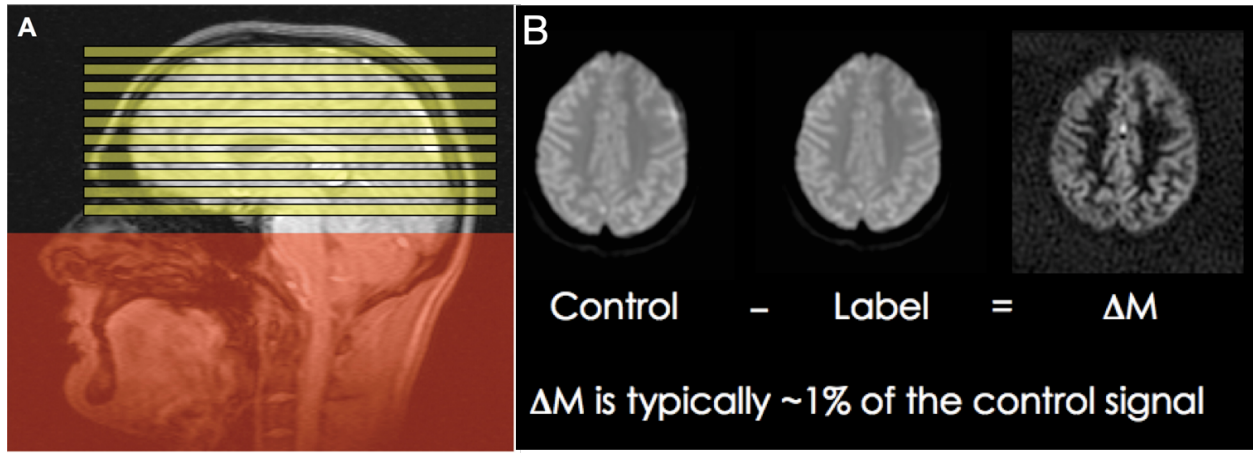
In this section, we will review the basics of each technique while discussing their advantages and disadvantages in the context of brain tumor imaging.

ASL uses a noninvasive endogenous tracer which is an enormous advantage for patients that need serial perfusion scans, as gadolinium (Gd) is a toxic substance that can cause damage if given too often[60]. ASL selectively inverts the longitudinal magnetization vector of a slice or slab containing inflowing blood to the slice of interest, thereby harnessing the water within the blood itself as a tracer (Figure 3.9a). This labeled water flows into capillaries and exchanges with tissue water, which can reduce the total tissue magnetization in a slice of interest by approximately 1%. The subtraction of the labeled image from a control image gives an image proportional to the cerebral blood flow (CBF) within the slice of interest, which can be used in conjunction with several quantitative steps to approximate absolute measures of CBF. It tends to have less variability across subjects and fewer susceptibility artifacts compared with other techniques. Though these advantages make ASL an attractive alternative to Gd-based agents, this technique suffers from low signal-to-noise ratio (SNR) which can only be amended by multiple acquisitions. Unfortunately, patient movement can be detrimental to this perfusion technique, as the subtraction of voxel intensities requires spatial alignment. It also relies on several assumptions that are not always met, especially in pathological conditions, such as an intact blood-brain barrier and the assumption of upward arterial flow.

DCE imaging uses a T1-weighted inversion-recovery gradient-echo sequence that acquires serial images before, during and after administration of an intravenously injected, gadolinium-based contrast agent. These signal changes are used in conjunction with a two-compartment pharmacokinetic model to approximate the permeability between the blood vessels (compartment one) and extracellular space (compartment two) (Figure 3.10). In turn, one can derive the rate of transfer between the two compartments ( $K_{trans}$  and  $K_{ep}$ ), which will describe the permeability of the blood vessels. Resulting maps reflect a composite of tissue perfusion, vessel

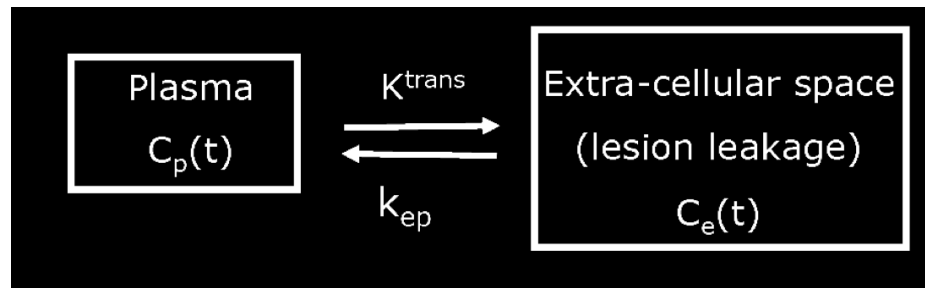


permeability, and extravascular-extracellular space that can be used to estimate the fraction of blood within a given volume.



**Figure 3.9. Arterial Spin Labeling (ASL).**

(A) A slab of blood entering upward through the arteries (red area) is labeled. Signal derived from slices (yellow) are acquired before (B, control) and as the labeled blood flows through (B, label), resulting in ~1% decrease in slice magnetization. The change in magnetization is found using a subtraction of the control and labeled slices and is proportional to the relative cerebral blood flow (rCBF).



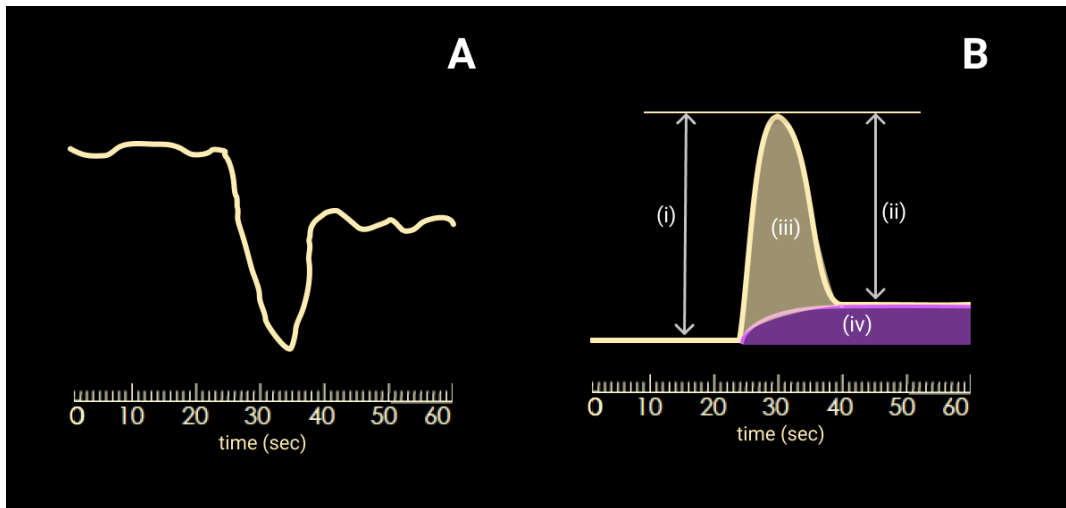
**Figure 3.10. Diagram of the two-compartment pharmacokinetic model used to calculate  $K^{trans}$  in DCE imaging.**

DSC is similar to DCE imaging in that it too requires the injection of an exogenous Gd-based contrast agent. In contrast to DCE, DSC captures its effects on  $T2^*$  recovery through the use of serial  $T2^*$ -weighted MR images. During the first pass of the contrast agent through the region of interest, the strong paramagnetic properties of Gd induce local magnetic field distortions around the vessels, further accelerating  $T2^*$  dephasing and loss of signal. When serial

images are captured in quick succession, the loss of signal intensity in each voxel can be plotted as a function of time (Figure 3.11). Using the following transformation, a curve denoting the change in relaxation rates called  $\Delta R2^*$  can be approximated that is proportional to the concentration of contrast agent:

$$\Delta R2^* \sim \frac{-\ln(S(t)/S_0)}{TE} \quad (3.19)$$

This transformation is also demonstrated in Figure 3.11, where 3.11a represents the  $T2^*$  signal change in a single voxel over time, and Figure 3.11b denotes the approximated  $R2^*$  change of that same voxel.



**Figure 3.11. Changes in  $T2^*$  and  $\Delta R2^*$  signal for DSC metric calculations.**

(A) Example of the decrease in  $T2^*$  signal over time in a voxel containing leaky blood vessels. (B) The  $\Delta R2^*$  curve as calculated by Equation 3.19: (i) peak height; (ii) percent recovery; (iii) relative cerebral blood volume; (iv) leakage factor.

Figure 3.11b also visually demonstrates the values of metrics derived from the  $R2^*$  curve, including (i) peak height, (ii) percent recovery, (iii) relative cerebral blood volume (rCBV), and (iv) leakage factor. rCBV can be approximated using the integral of the  $\Delta R2^*$  curve:

$$rCBV \sim \int \Delta R2^*(t)dt \quad (3.20)$$

However, there are necessary corrections when measuring rCBV in the perforated vessels present in glioma. To quantify rapid signal changes in tumors following contrast injection, high temporal resolution is needed; however, the  $\Delta R2^*$  curve obtained at a high temporal resolution can be contaminated by shortening tissue water T1 when there is BBB breakdown and extravasation of the contrast agent into the extracellular space [61]. These T1-shortening effects oppose the T2\* signal decrease, impacting the accuracy of rCBV calculations. In Figure 3.11b, (iii) rCBV is shown corrected for the (iv) leakage factor. Many approaches have been proposed to correct for T1-shortening effects of leakage: a) using mathematical post-processing correction approaches [62–64]; b) lowering the flip angle during excitation [65]; and c) pre-loading with a dose of contrast agent prior to image acquisition [62]. As early as 2006, Boxerman et al. discovered that rCBV maps corrected for contrast extravasation (leakage) correlate significantly with tumor grade, while uncorrected maps do not [64].

### 3.3.2 The role of perfusion-weighted imaging in brain tumor diagnosis

In Section 2.2.3, the biological mechanisms mediating new blood vessel growth in glioma were introduced. Briefly, as tumor cells multiply in glioma, they stimulate the recruitment and proliferation of endothelial cells in a process called angiogenesis. The result is a reorganized novel microvasculature system made to deliver the nutrients necessary for cell viability inside the growing cell mass. It follows that imaging and quantifying these properties could have important implications for noninvasively probing underlying tumor cell biology.

Finding the best MR imaging metric that indicates progressive disease prior to the manifestation of contrast enhancement would allow for early detection of treatment failure and

remains an important clinical objective. One 2011 study compared metrics derived from DSC, DCE and MR spectroscopy (see Section 3.4.1) to evaluate if any could identify progression earlier than conventional imaging, and if so, which had the most accurate diagnosis. The best performing metrics were tumor blood flow and blood volume derived from DSC imaging, supporting the idea that microvascular changes might manifest in perfusion imaging before structural changes in conventional T1 and T2-weighted MRI[66]. In addition, there is increasing adoption of antiangiogenic therapy into care for both newly diagnosed and recurrent glioma patients, for which DSC-derived parameters have been shown to serve as early markers of treatment failure or response[67]. In short, MR perfusion metrics could serve as biomarkers that manifest earlier than lesions imaged with structural MR alone.

## 3.4 MR Spectroscopy

### 3.4.1 Introduction to MR Spectroscopy

Magnetic Resonance Spectroscopic Imaging (MRSI) differs from classic MRI in that it can differentiate signals from hydrogen protons in metabolites other than water. The theoretical applications of this technology are profound: we can use mechanism-driven imaging strategies for improved diagnosis and prognosis; we can identify metabolic imaging markers of therapeutic response; and we can discover new etiologic and therapeutic biomarkers using MR. So why hasn't it been adopted as a panacea?

The fundamental challenge with all MRSI is low concentration: an average metabolite's concentration is approximately 10% that of water, which results in a 10000x reduction in the signal-to-noise ratio (SNR). Beyond this biological challenge, it is also technically challenging to acquire, requiring additional training for MRI technicians and long scan times. Compounding these, the resulting data requires skilled and extensive post processing because, among other complications, the data suffers from interfering signals from subcutaneous lipids. Not so much of a panacea after all.

Despite all of these challenges, MRSI is a promising technique as improvements in hardware and software inevitably advance and have the potential to overcome some of these challenges. Our lab has developed robust techniques automating the prescription and acquisition of MRSI [68]. In addition, increasing availability of high field scanners and multichannel radiofrequency coils has improved signal-to-noise ratios. From these advances, acquisition time has been shortened and coverage has been increased enough to be clinically useful. Here, we will

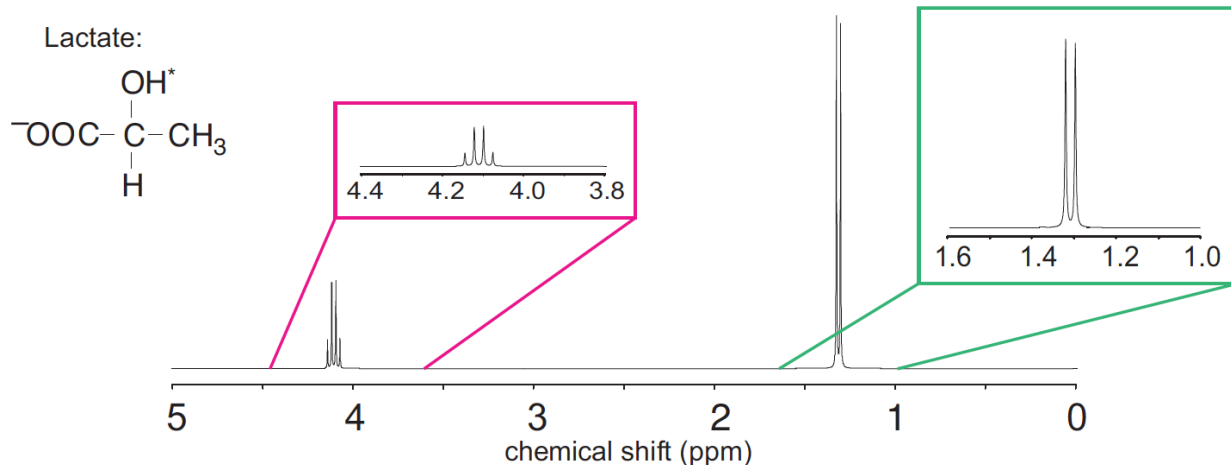
overview the basic principles necessary to understand how metabolite spectra can be acquired using NMR principles.

An MR spectroscopic signal is made up of multiple frequency components (either from different metabolites or different protons from the same metabolite) that can be observed and quantified in the same spectroscopic voxel due to a phenomenon known as chemical shift. Briefly, chemical shift is the shift in frequency due to the chemical environment surrounding the proton. Recalling basic chemistry, one might remember the influence of asymmetric electron clouds on the polarity of molecules. An analogous effect is in play when we consider the magnetic field experienced by the proton nucleus: a dense electron cloud surrounding a proton can “shield” the nucleus from the magnetic field, causing each proton to experience slightly different magnetic field strengths. The variations in different resonant frequencies can be expressed as

$$\delta = \frac{\omega - \omega_{ref}}{\omega_{ref}} \quad (3.21)$$

where  $\omega$  is the absolute frequency of the sample and  $\omega_{ref}$  is the frequency of a reference compound under the same  $B_0$ . Because the numerator is measured in Hz and the denominator in MHz, this measurement is expressed in parts per million (ppm).

The other critical component to analyzing spectra requires an understanding of the phenomenon known as J-coupling, which is the manifestation of neighboring protons as split peaks on an NMR spectrum. In the simplest sense, protons that have  $n$  equivalent neighboring protons resolve on the NMR spectrum with  $n+1$  peaks. This relationship is always mutual; i.e. if nucleus A affects the precession frequency of nucleus B through J-coupling, nucleus B also affects the frequency of nucleus A. An example of an NMR spectrum derived from lactate shown below in Figure 3.12.



**Figure 3.12. The NMR spectrum of lactate.**

This spectrum was chosen as an example to demonstrate the chemical shifts of protons experiencing disparate electronic environments as well as the J-coupling splits from adjacent nuclei.

The first peak at ~4.1 ppm is the signal from the singular hydrogen attached to the central carbon atom, which is split into 4 based on its three equivalent neighbors on the methyl group. The second peak at ~1.3 ppm is split into two based on its neighboring hydrogen just discussed. We observe greater magnitude of the peaks because of the greater number of equivalent hydrogen atoms resonating at that same frequency, experiencing that same shift.

These fundamentals in combination with the introduction to MRI fundamentals in 3.1.1 should provide a sufficient foundation for understanding how we can spatially resolve voxels, each with their own spectra describing differences in metabolic concentrations. A rigorous explanation of MRSI acquisition steps, processing steps, and the challenges at each are beyond the scope of this dissertation; however, I will briefly mention some of these steps so that one can appreciate how much effort is needed to accurately resolve metabolites of such low concentration:

- Water suppression: Suppression of the signal derived from protons in water is necessary to resolve any metabolites at 10% the concentration. The most common technique to suppress the signal from water is to use a chemical shift selective pulse sequence (CHESS), which will leave the spin system in a state where no net magnetization of the component resonating at the Larmor frequency of water is retained.
- Lipid suppression: In tissues that have large fractions of adipocytes (e.g. subcutaneous fat), large signals from lipids can overwhelm the signal derived from other regions and therefore need to be suppressed. On the other hand, lipid within the lesions to be visualized can be important biomarkers for tumor malignancy, as increased lipid is usually associated with increased necrosis and cellular division. One can appreciate the challenge that such a scenario presents. Clever protocols such as outer volume suppression (OVS) and very selective saturation pulses (VSS) are used to selectively suppress lipid signals from subcutaneous fat.
- Combining signal from multi-channel RF receivers: In short, the more coils, the better the spatial encoding and the faster the acquisition time. However, geometric variations, wiring, receive delays and electrical properties make the combination of signal acquired from each additional coil that much more of a post-processing puzzle.

In essence, MRSI still requires skilled technicians and researchers to correctly acquire and analyze its findings; however, it is a powerful tool that has the potential to noninvasively characterize disease metabolism, an extraordinarily important clinical objective.

\* On a personal note, I find MRSI remarkable. Not only can we image different tissues in the body with regular MRI, but we have actually figured out how to probe the *metabolism* of the



human body with MRSI. We can take a human, put them in a strong magnet, and we can actually visualize where certain metabolic functions are upregulated or downregulated according to the concentration of their byproducts. Despite the numerous limitations of MRSI, I will never cease to find this accomplishment astonishing.

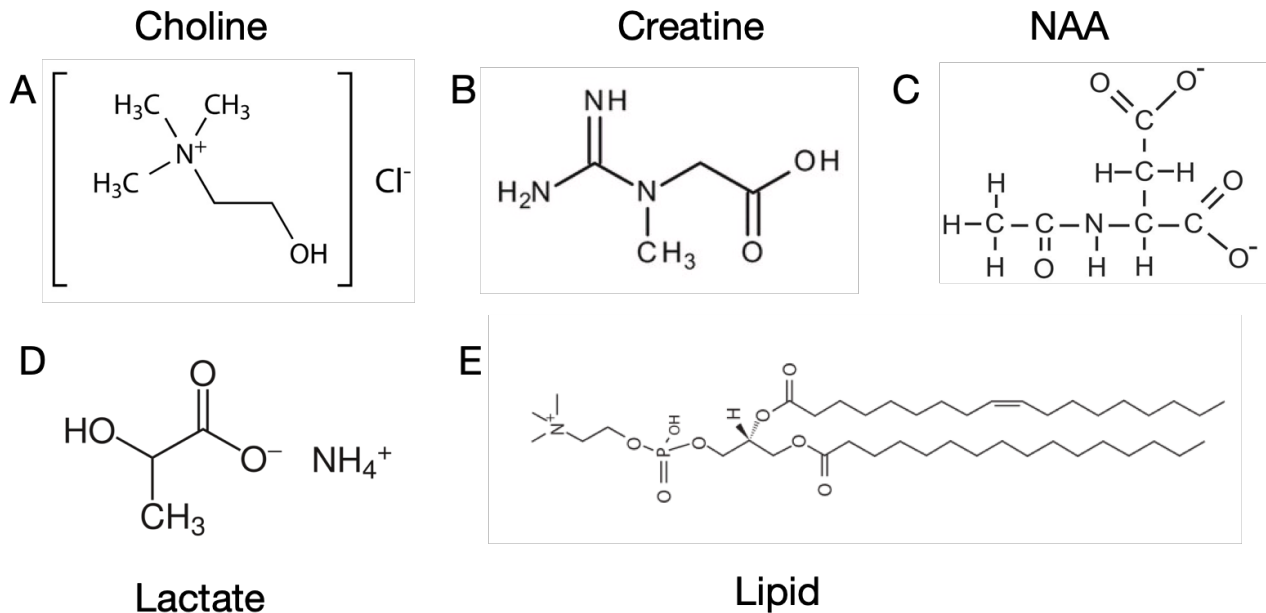
### 3.4.2 The role of MR spectroscopy in brain tumor diagnosis

The major metabolites that are resolved in brain tumor MRSI and their biological relevance are described below:

- Choline containing compounds (nCho): Choline and choline-containing compounds (glycerophosphocholine, phosphocholine) are essential for the synthesis of the cellular membranes[69]. Therefore, it reasons that increased choline is related to increased mitosis and cellular production, as in a rapidly multiplying glioma. Other pathological processes such as active demyelination due to inflammation. Its dominant signal from its 3 equivalent methyl groups resonate as a composite peak from all choline-containing compounds at 3.20 ppm from (Figure 3.13a).
- Creatine and phosphocreatine (Cre): Creatine is involved in ATP metabolism and exhibits low variability across brain regions and across subjects; though there remains debate about its fluctuations within glioma[70–72]. Total creatine (creatine plus phosphocreatine) is most commonly used as the reference metabolite for ratio normalization. Its methyl group resonates as a composite peak from all creatine containing compounds at 3.03 ppm (Figure 3.13b).
- N-acetyl aspartate (NAA): One of the most concentrated molecules in the CNS, NAA can reach concentrations up to 10 mM. It is specific to the nervous system because it is

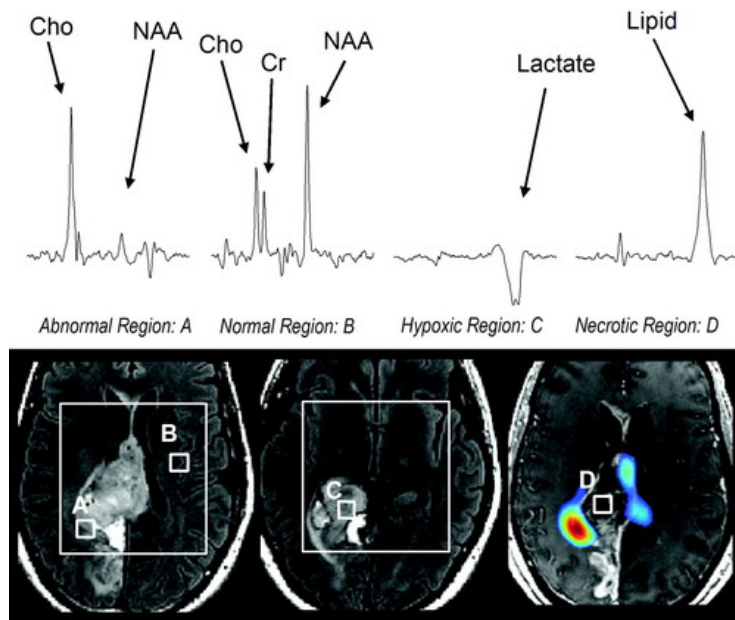
synthesized from aspartate and acetyl-coenzyme A in neurons [73]. Its methyl group resonates as a singlet at 2.02 ppm (Figure 3.13c).

- Lactate (Lac): Lactate is thought to be produced under hypoxic conditions. It resonates as a doublet centered at 1.31 ppm, as well as a smaller quadruplet around 4.1 ppm that is not observed in typical long TE sequences (Figure 3.12, 2.13d).
- Mobile lipid (Lip): An increase in lipid signals observed in tumors is thought to be due to membrane breakdown and subsequent release of mobile fatty acids. Lipid signal can be observed prior to the apparent presence of necrosis, suggesting that membrane breakdown may precede histologically evident necrosis[74]. It resonates as a singlet around 1.3 ppm, which can confound the signal derived from lactate unless advanced protocols are in place to account for separate lactate from lipid using specialized spectral editing sequences. An example of one lipid structure is given in Figure 3.13e.



**Figure 3.13. Chemical structures of metabolites relevant for quantifying in glioma at long TE.**

The vast majority of studies use long TE because it allows for lactate-editing that can resolve lipid concentrations from lactate concentrations. With long TE, these are the main compounds observed in glioma (Figure 3.13); MRS spectra acquired using long TE from voxels with distinct pathologies are depicted in Figure 3.14. However, it is also possible to use a short TE with MRSI and resolve other compounds resonating with larger ppm values, e.g. myo-inositol and 2HG, which are additionally implicated in glioma.



**Figure 3.14. Examples of differential metabolite concentrations in regions with disparate pathological burden.**

Image reproduced from Osorio et al., 2007 [75].

One of the first multi-institutional MRS studies (1996) established the feasibility of acquiring single-voxel spectroscopy data in brain tumor patients [74]. Glial tumors were found to exhibit elevated mean signal intensity of choline, decreased signal of creatine and NAA compared with contralateral normal brain, a promising first step toward metabolically characterizing glioma. Since 1996, MRS has advanced to include multiple voxels in the

prescribed area, allowing for spatial resolution of the concentration of metabolites. Figure 3.14 depicts the results of a 2007 study where region A exhibits elevated Cho and depressed NAA compared with region B, suggesting an increase in cellular proliferation that cannot be neuronal (and are therefore likely glial[75].) In the necrotic region D, elevated lipid signal implies deterioration of cellular membranes as described above. This study concluded a statistically significant increase in the choline-to-NAA ratio in the T2 lesion of glioma. These early findings have since been applied to most pressing questions in brain tumors.

We return to the problem of distinguishing recurrent glioma from treatment-induced lesions. Given that each phenomenon is characterized by distinct cellular populations, it reasons that their metabolic signatures could help distinguish them. It follows that if MRSI is able to resolve these metabolic differences, it could have the unique capability of being able to distinguish true recurrent glioma from the inflammatory and necrotic response from radiation therapy. Several studies have reported significant differences in Cho/Cre and Cho/NAA ratios within the contrast enhancing lesion [53,54,56,76–81]. In two recent meta analyses, these metabolite ratio thresholds demonstrate the most sensitive and specific ability to distinguish these two entities compared with any metrics derived from anatomical, diffusion or perfusion spectroscopy [82,83].

Beyond the metabolites listed above, of particular interest is imaging the concentration of the metabolite 2-hydroxyglutrate (2HG). Because IDH1 mutations occur on only a single copy of the gene, some transcribed enzymes retain the ability to catalyze the conversion of isocitrate to  $\alpha$ -ketoglutarate. However, the mutated IDH1 gene produces an enzyme with a gain of function that converts  $\alpha$ -ketoglutarate to 2HG. In 2012, MR spectroscopy correctly identified the presence of 2HG in all IDH mutant tumors (n = 15) while no detection of 2HG correctly

identified every IDH wildtype patient (n=14) [84]. This study shows remarkable sensitivity and specificity (exceeding that of any machine learning algorithm) for identifying IDH mutation, highlighting the utility of MRS in the context of IDH.

## 4. Introduction to machine learning and deep learning for neuroimaging

### 4.1 Introduction to machine learning

With the advent of the information age, vast amounts of data are being generated and stored at an ever-increasing pace. For a given problem, it is the job of the researcher to make sense of the data, unveiling patterns to understand “what the data says.” In other words, we want to *learn from the data*. Broadly, we can categorize techniques used to learn from data into two categories: supervised and unsupervised learning. When we use supervised learning, the goal is to learn a function  $f: X \rightarrow Y$  that maps the input domain  $X$ , which can be thought of as a set of features, to the output domain  $Y$ , the outcome of interest. In contrast, there is no outcome measure in unsupervised learning, but the goal is to instead describe patterns and associations among a set of input features[85].

Rigorous descriptions of all machine learning and deep learning techniques are the subjects of many books, including a favorite of mine titled the “Elements of Statistical Learning” by Rob Tibshirani, Trevor Hastie, and Jerome Friedman and “Deep Learning” by Ian Goodfellow [85,86]. For the purpose of this dissertation, I will break my discussion of machine learning and deep learning into:

- 4.2, an overview of some “classic” machine learning techniques that I have used in my research;
- 4.3, an overview of methods used when there is spatial correlation between observations; and
- 4.4, an introduction to artificial neural networks, deep learning and convolutional neural network classifiers.

## 4.2 Classic machine learning techniques

In this overview, we will briefly discuss the design of supervised machine learning algorithms that are relevant for this dissertation, as well as their advantages and disadvantages. Inputs are interchangeably denoted as *predictors*, *features* or *independent variables*, and they are typically represented as by the symbol  $X$ ; if it is a vector, its components can be accessed by a subscript  $X_j$ . They can be categorical (e.g. “red”, “blue”), ordinal (e.g. “small”, “medium”, “large”), or numeric (e.g. 0, 1000). In most algorithms, categorical and ordinal variables are numerically encoded. The outcome that we aim to predict is often called the *response* or *dependent variable*, and is represented by a variable  $Y$ . If  $Y$  is continuous, we refer to the task as regression, while if  $Y$  is discrete, we refer to the task as classification. The basics of learning model parameters from these paired data are explored in Section 4.2.1 and the fundamental concepts illustrated here can be applied to deep learning, as seen in Section 4.4.

### 4.2.1 Linear models

Given a vector of inputs  $X^T = (X_1, X_2, \dots, X_p)$ , we predict the output  $Y$  via the model:

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^p X_j \hat{\beta}_j \quad (4.1)$$

where  $\hat{\beta}_0$  is the intercept (or “bias”), and  $\hat{\beta}_j$  is the coefficient of feature  $X_j$ . Often, the magnitude of  $\hat{\beta}_j$  is thought of as the average effect of a one-unit increase of  $X_j$  on the outcome. When equation 4.1 is represented in vector form, it takes the form of equation 4.2:

$$\hat{Y} = X^T \hat{\beta} \quad (4.2)$$

where  $X^T$  denotes the vector or matrix transpose. In order to fit the model to the training set, it is necessary to choose the metric by which to minimize the error. There are many different

methods, but the most popular is minimizing the residual sum of squared error between predictions  $\hat{Y}$  and ground truth outcomes  $Y$ :

$$RSS = L = \sum_{i=1}^N (y_i - \hat{y}_i)^2 = \sum_{i=1}^N (y_i - x_i^T \hat{\beta})^2 \quad (4.3)$$

which after differentiation has the following optimal solution:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (4.4)$$

if  $X^T X$  is nonsingular (i.e. its determinant is nonzero).

Though this is the quick solution used in all practical scenarios utilizing linear regression, it is worth taking an alternative approach that more closely resembles model parameter optimization in more complex algorithms. Let's imagine that we begin with a random initialization of the model, in a sense "guessing"  $\hat{\beta}$ . We can use Equation 4.2 to calculate an initial "guess" of  $\hat{Y}$ , and calculate the residual sum of squares through Equation 4.3. When differentiating Equation 4.3, we obtain the following equation:

$$\frac{\partial L}{\partial \hat{\beta}} = \sum_{i=1}^N 2(y_i - x_i^T \hat{\beta}) \quad (4.5)$$

The value of the derivative (Eq 4.5) will determine the direction and magnitude with which to update our model parameter weights according to the following function:

$$\hat{\beta}_{updated} = \hat{\beta} - \alpha \frac{\partial L}{\partial \hat{\beta}} \quad (4.6)$$

where  $\alpha$  is a hyperparameter that allows the experimenter to control how large or small (fast or slow) the updates will converge. This process can be repeated until convergence or until the loss is changing at a rate less than a certain threshold. When a loss function exists on a high-dimensional hyperplane, choosing the right  $\alpha$ , or learning rate, becomes critical to converging upon a solution to optimal  $\hat{\beta}$ . Though this process of iteratively updating model weights is



decidedly less efficient than Equation 4.4, it is necessary when there are no closed-form solutions in more complex algorithms, and a useful exercise in the simple context of linear regression parameter optimization.

Logistic regression is an extension of linear regression when we are interested in the probability that  $X$  belongs to one of two classes. In the simplest sense, it can be represented with the following equation:

$$p = P(Y = 1|X) = \frac{e^{\hat{\beta}_0 + \sum_{j=1}^p X_j \hat{\beta}_j}}{1 + e^{\hat{\beta}_0 + \sum_{j=1}^p X_j \hat{\beta}_j}} \quad (4.7)$$

Through algebraic manipulation, we can recover the log odds, which allows for an intuitive comparison to Equation 4.1:

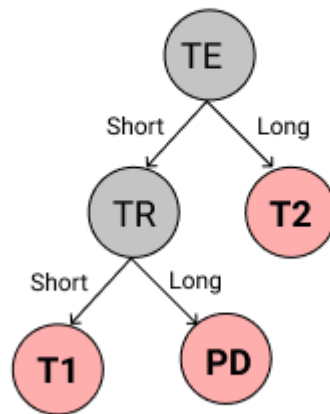
$$\ln\left(\frac{p}{1-p}\right) = \hat{\beta}_0 + \sum_{j=1}^p X_j \hat{\beta}_j \quad (4.8)$$

Both logistic and linear regression are flexible and can be combined with different optimization techniques and parameter weight penalties to achieve simple, interpretable, and useful models. Notably, for both logistic and linear regression, correlations among predictors can cause problems: variance of coefficient estimates tend to increase dramatically, while classic interpretations (a one-unit increase of  $X_j$  impacts  $Y$  in this way) become hazardous.

#### 4.2.2 Decision trees and random forests

Decision trees are a simple, intuitive algorithm that uses a series of decisions to arrive at a conclusion. A feature can be represented as a node, while the decision represents the threshold of that feature that creates a split between data (Figure 4.1).

Is the image T1, T2, or PD weighted?



**Figure 4.1. Example of a simple decision tree to classify the contrast of an MRI.**

As pictured, a decision tree can mimic the way some people make decisions and can use qualitative (categorical) features without encoding. How splits are decided at each level will depend on whether the tree is being used for classification or regression, but the simplest idea is to iterate over all possible features and thresholds of those features and find the threshold that best splits the data with minimal error (regression: RSS (Eq. 4.3); classification: 1-accuracy). In practice, greedy search algorithms are used to reduce the computational cost, but the idea is essentially the same: build a tree with the best split at each point to best predict our outcome. Though single decision trees rarely predict well on their own, algorithms of decision tree ensembles have been quite successful.

One method of ensembling trees is through the use of bagging. In general, we do not have more than one training dataset from which to build a model. Bagging describes the generation of  $B$  different bootstrapped training sets - meaning training sets sampled with replacement. A method is then trained on the  $b^{\text{th}}$  bootstrapped training set, from which a prediction can be

calculated  $\hat{f}^{*b}(x)$ . An average of these predictions is known as bagging and is represented with the following equation:

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{*b}(x) \quad (4.9)$$

Random forest (RF) is an algorithm that combines bagging with a random subset of feature choices per tree. This effectively decorrelates all of the trees, in turn reducing the variance typically observed from a single tree. It is attractive because it can handle categorical or continuous features, the data do not need to be rescaled or transformed, and it is robust to outliers, class imbalance, and nonlinear data. This comes at a tradeoff with model interpretability, and despite reduction of variance through bagging, it still can overfit to training data. The general schematic of a random forest algorithm is depicted in Figure 4.2.

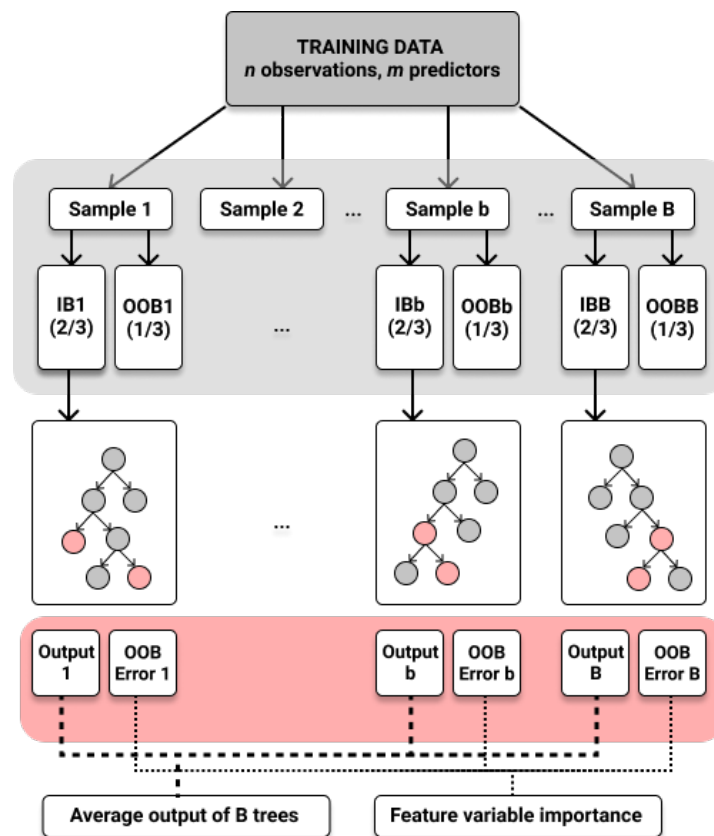


Figure 4.2. Diagram of a Random Forest algorithm.

In contrast to RF where all trees are averaged (or treated equally), boosting is a technique that builds models sequentially, building off of the errors of the prior tree. In brief, each successive tree increases the weight of misclassification error for samples that were predicted incorrectly in the previous tree. Its many hyperparameters provide immense flexibility and often unparalleled predictive accuracy compared with other model paradigms. However, their many hyperparameters might necessitate a large grid search during hyperparameter tuning which contributes to being a computationally expensive task.

### 4.3 Spatially repeated measures

All of the models that are discussed in Section 4.2 rely on the assumption that each observation is independent from one another, including the simplest cases of linear and logistic regression. But what happens when observations are correlated with one another in some way? Data can be “clustered”, which is the case when subjects themselves are grouped together (e.g. patient cohorts from different institutions), but the dependent variable is measured just once per subject. In “repeated measures” data, the dependent variable is measured more than once per subject (e.g. a control knee and injured knee). Both clustered and repeated measures data result in correlations among data points that must be accounted for during statistical analysis, as the observations are no longer independent from one another.

As we will discuss in Chapter 7, we repeat measurements within a patient when resecting multiple tissue samples from brain tumor patients during surgery. Our goal is to predict a binary outcome -- whether our tissue sample was diagnosed as treatment-related injury or real tumor recurrence -- from preoperative MRI parameters on the location of the tissue sample. Without getting into too much detail about that specific problem, using methods designed specifically for

repeated measured data can generate more descriptive models and therefore more accurate predictions to generalized data.

The way to account for spatially repeated measures data varies on the target of inference: 1) subject-specific outcomes; or 2) population-averages. If interested in (1), e.g. uncovering the probability of a particular patient experiencing recurrence, the most appropriate modeling strategy might be a generalized linear model with mixed effects (both fixed and random):

$$\log\left(\frac{P(Y_{ij}=1)}{P(Y_{ij}=0)} \mid x_{ij}, b_i\right) = \beta_0 + \beta_1 x_{ij} + b_i \quad (4.10)$$

where we observe the binary response  $Y_{ij}$  from  $n_i$  samples in  $N$  individuals, where  $Y_{ij} = 1$  if sample  $j$  from individual  $i$  is pathologically tumor and 0 otherwise;  $x_{ij}$  is a continuous MR feature of interest from sample  $i$  of patient  $j$ . The odds ratio of recurrence risk differs based on the value of  $b_i$  which takes the baseline differences in recurrence risk into account on a per-individual basis, and the estimates are therefore subject specific.

However, we instead care to model whether an MR-derived parameter is associated with recurrence in the population (2). In this case, we are interested in fitting a marginal model so that we can uncover the population wise association of an MR parameter with tumor recurrence. In this case, we fit a generalized estimating equation:

$$\log\left(\frac{P(Y_{ij}=1)}{P(Y_{ij}=0)} \mid x_{ij}\right) = \beta_0 + \beta_1 x_{ij} \quad (4.11)$$

where we no longer have a  $b_0$  intercept that is subject-specific. More details can be found in the Analysis of Longitudinal Data, which, despite its name, is also appropriate for spatially repeated measures [87].

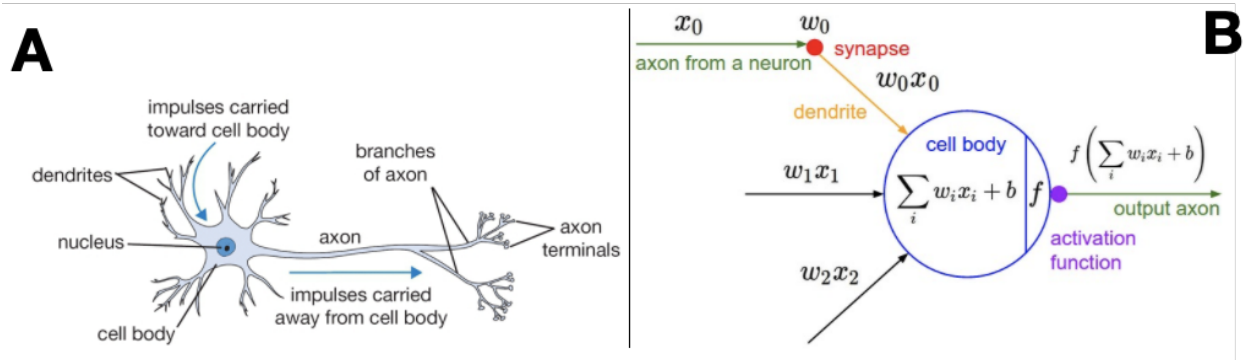
## 4.4 Introduction to deep learning and convolutional neural network classifiers

The Universal Approximation Theorem in its most basic sense states that neural networks can approximate any continuous function to a reasonable accuracy. Neural networks are not just interesting because they can powerfully represent nearly any function, but also because they self-learn features from unstructured data like text, audio, images, and video. Compared with the necessity of predefined features for models discussed in Section 4.2, this feature self-learning capability is a paradigm shift that changed the way we use machine learning together with imaging. In this section, I will briefly discuss the basics of artificial neural networks, as well as the utility of convolutional neural networks for imaging data.

### 4.4.1 Introduction to artificial neural networks

The basic computational unit of the human brain is a neuron, which receives a combination of input signals from its dendrites to produce a single output signal along an axon (Figure 4.3a)[88]. In essence, if the sum of signal derived from all the dendrites is great enough, the neuron will “fire”, sending signal to along an axon and its terminal nodes to eventually connect with other neuronal dendrites and propagate the signal.

An analogous model is generated computationally to create an artificial neuron (Figure 3.3b). The inputs of previous neurons to the current neuron are represented by  $x_i$ , whose weight  $w_i$  can be thought of as the contribution of neuron  $i$  to the probability of the current neuron firing. These contributions  $w_i x_i$  are then sent through an activation function modeled by  $f$ , which will determine whether the sum of all  $w_i x_i$  is great enough to propagate the signal from that neuron forward.



**Figure 4.3. Structural comparison between a biologic neuron and an artificial neuron.** (A) Biologic neurons receive chemical signals from other neurons via dendrites, which, upon hitting a certain threshold, can fire an electric signal down an axon to propagate the signal to downstream neurons. (B) The artificial neuron is based off of this structure, receiving input from prior neurons which are sent through an activation function, controlling whether the artificial neuron will fire. Figure adapted from Stanford CS231n, publicly available neural network course [88].

Activation functions can take many forms, but the original function used was the sigmoid:

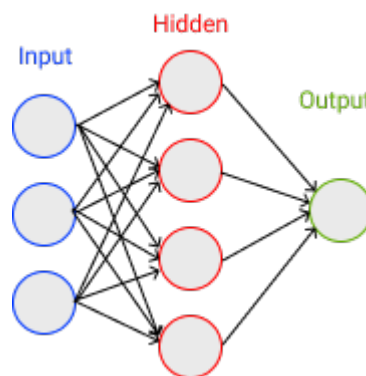
$$s(x) = \frac{1}{1+e^{-x}} \quad (4.10)$$

that maps any real valued number to a value between 0 and 1, which can be interpreted as the probability of a neuron firing. Another commonly used function is the rectified linear unit (ReLU):

$$s(x) = \max(0, x) \quad (4.11)$$

which allows for more favorable behavior when training deep neural networks. Specifically, it was beneficial in two ways: 1) compared with the sigmoid function that involves exponentiating terms, it is implemented with a simple threshold which is less computationally expensive; 2) it was found to accelerate the convergence of stochastic gradient descent (i.e. the process by which a loss minimum is found in a neural network) compared with the sigmoid function (Eq. 4.10)[89].

An example of a simple, two layer feedforward neural network is given in Figure 4.4. In this figure, the leftmost blue layer represents the input features themselves; in this problem, there are only three input features but in practice, there are typically many more. The inner red layer is “hidden” and conducts intermediate processing, where each of four red nodes is simply a linear combination of the input features as described in Figure 4.3. The final green layer represents the output; in this case, we have a single output which is typically sent through a sigmoid function to decide whether the input belongs to one of two classes, numerically represented by 0 and 1. Feedforward networks with more than one hidden layer are typically referred to as deep neural networks.



**Figure 4.4. Diagram of a simple two-layer feed-forward artificial neural network.**

During training, this final output will be compared to the ground truth class and the difference will be calculated according to a set loss function. Here, we return to the concept of iteratively updating model weights introduced in Section 4.2.1. In this example of binary classification, one typical loss function is called “log loss” or “binary cross entropy.” The same concepts are then applied; we differentiate the loss with respect to the inputs, and use a “learning rate” to iteratively change the model weights as we work our way backward through the layers of the model. Understanding the basics of iteratively updating model weights is essential for



understanding how a model “learns” from paired input and output data during training.

Optimizing these processes, called stochastic gradient descent and backpropagation, are active areas of research for all kinds of deep neural networks.

#### 4.4.2 Introduction to convolutional neural networks

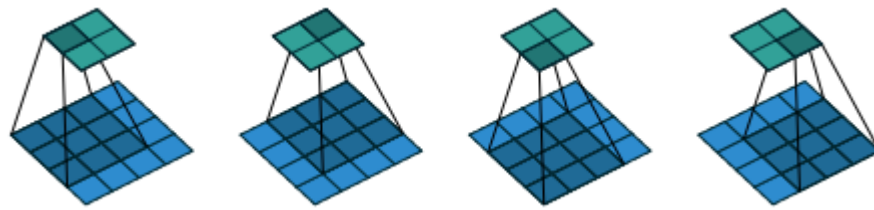
Computer vision tasks related to imaging are faced with specific limitations: 1) input images usually contain thousands if not millions of input pixels, each of which could be considered its own input into the model; and 2) a feedforward network with an architecture similar to that in Figure 4.4 cannot capture spatial relationships between features. Challenge #1 dramatically increases the number of weights in the network that need to be optimized, which increases the risk of overfitting and has much slower training efficiency. Challenge #2 means that the rich information of how proximate pixels relate to one another is lost.

Convolutional neural networks are a class of deep neural networks that aim to combat these limitations using spatial convolutions, which are specialized linear operations. Consider a two-dimensional image  $I$  as input with a two-dimensional kernel  $K$  that will perform the convolution. The output of the convolution of image  $I$  with kernel  $K$  is given by the following equation:

$$S(i, j) = (K * I)(i, j) = \sum_m \sum_n I(i - m, j - n)K(m, n) \quad (4.12)$$

[89]. Complementary to this equation, a more intuitive understanding of convolution can be acquired by visualizing the operation. In Figure 4.5, we demonstrate the convolution of a 4x4 blue input image with a 3x3 shaded kernel to create a 2x2 green output where each of four convolutions is the element-wise multiplication of the 3x3 input area with the 3x3 kernel, and subsequently summed to create the single output value in the green that is shaded. The kernel can

be thought of as “sliding” over the input area, and the 4 distinct spots where this convolution operation can be performed become the spatial 2x2 output. The 3x3 kernel values are the learned model weights during training of the CNN. The activation function is then applied to these values after the convolution function; if ReLU is chosen, we preserve all the positive values and set the negative values to zero.



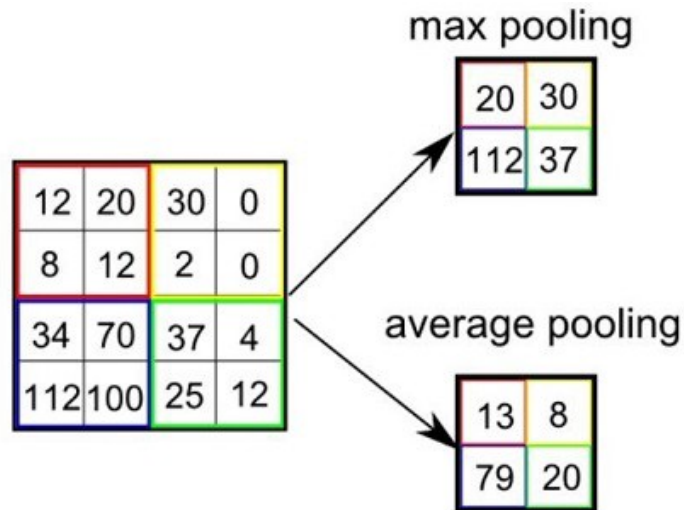
**Figure 4.5. Convolution operation of a 4x4 input image with a 3x3 shaded kernel to create a 2x2 green output.**

Image reproduced from Dumoulin et al., 2016. [90]

Pooling is another essential operation found in most convolutional neural network architectures. In many ways, pooling works just like a convolution; in contrast to convolutions, pooling layers use a set function such as the maximum or average, which in turn does not require learning the weights of a kernel as it is a set operation. The pooling operation can be described by Figure 4.6, where a 2x2 pooling layer is applied to a 4x4 input image (with stride 2) to create a 2x2 output.

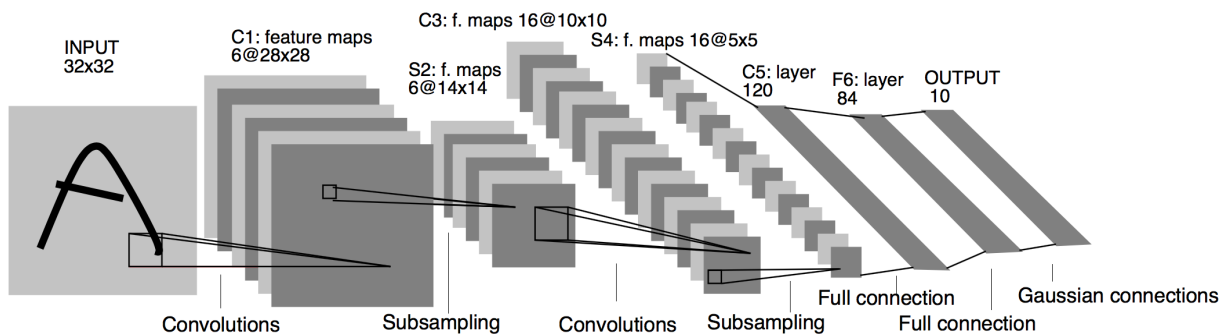
The first great achievement using convolutional neural networks was in the context of recognizing handwritten digits. In 1998, Lecun et al. proposed the LeNet-5 (Figure 4.7) where “subsampling” layers are another way of saying “average pooling”[91]. Therefore, the architecture consisted of first convolving the image, then activating the resulting feature maps, using an average pooling layer to reduce dimensionality, one more convolutional > activation >

average pool set, and finally three fully connected layers that behave like a regular feedforward network, each node receiving input from all nodes in the prior layer.



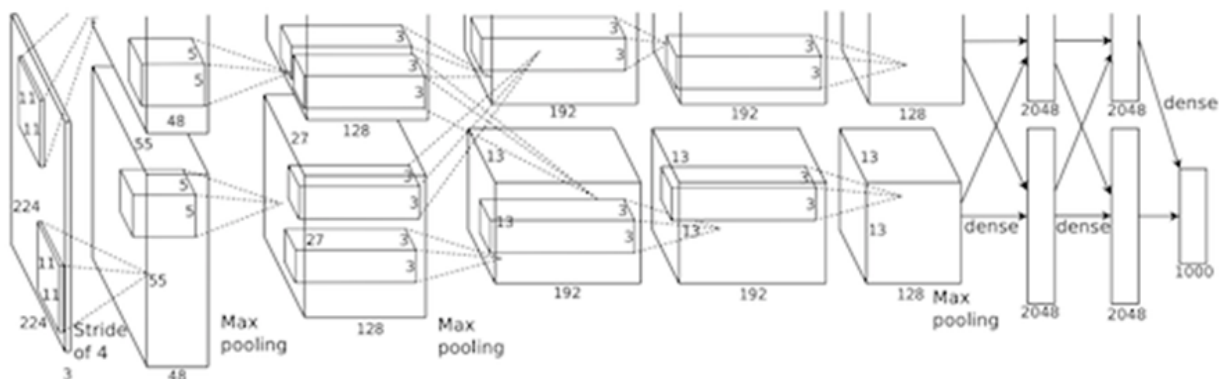
**Figure 4.6. Example of a max pooling and average pooling operations in a convolutional neural network.**

In max pooling, the maximum value of the region is used; in average pooling, a simple average across the values in the region is the output.



**Figure 4.7. Architecture of LeNet-5.**

Figure reproduced from Lecun et al. 1998 [91].



**Figure 4.8. Architecture of AlexNet.**

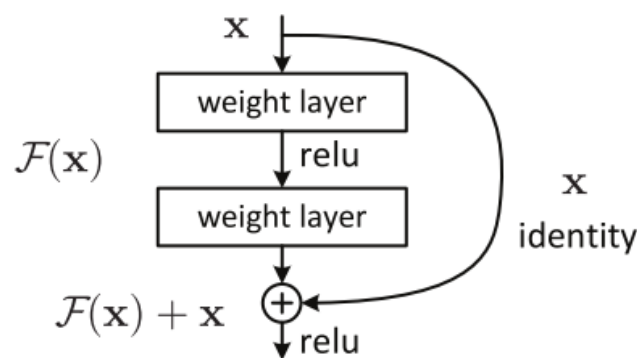
Figure reproduced from Krizhevsky et al. 2012. [89]

In 2012, AlexNet (Figure 4.8) made several important advances. It was the first GPU-accelerated CNN that achieved near-human performance on the ImageNet image recognition challenge, winning the challenge by a large margin. It was a much deeper network with far more parameters (~60M) compared with LeNet-5 (~60k). The important advances can be summarized as 1) It used ReLU instead of tanh as its activation function; 2) it used dropout instead of weight decay to reduce overfitting; and 3) it used overlapping pooling to reduce the size of the network[89].

In 2014, VGGNet’s success in the ImageNet competition showed that an increase in depth of the model resulted in better performance, which was an important advancement regarding our knowledge of CNN model behavior. At this time, the 16-layer VGG known as VGG16 was the deepest model known (~140M parameters). Appealingly, all of the convolutional layers are homogenous, in that they consist of 3x3 kernels with 2x2 pooling, so that it is easier to understand[92].

One challenge of networks growing deeper was that they became more difficult to train. As an answer to this problem, “residual” or “skip” connections were introduced into deep neural network architectures (“ResNets”)[93] (Figure 4.9), which won the ImageNet competition the

following year. In short, residual connections take the output of a prior layer and add these outputs to a downstream layer. Simple in theory, this idea was powerful: it meant that if the layers in between these summation “residual” connections were not useful, they could simply be skipped over rather than trained. It meant that adding additional layers to networks should - in theory - not decrease performance at all. Skip connections have since been incorporated in many more CNN architectures, achieving state-of-the-art performance in numerous domains.



**Figure 4.9. Residual “skip” connections that improve training accuracy of very deep convolutional neural networks.**

Introduced by He et al., 2016. [93].

Although the development of CNNs for image classification first began in 1998, the application of CNNs to medical imaging have only just burgeoned within the past five years. With respect to MR and glioma specifically, the first use of CNNs was first demonstrated by Pan et al. in 2015 [94] to classify glioma by grade using MR imaging. Another major leap forward in 2015 was the publication of the BRATS dataset, which sparked the use of convolutional neural networks to automate glioma lesion segmentation [95]. Other recent applications of CNNs to the analysis of anatomic images have been to predict mutations in the IDH gene, 1p19q co-deletions, and MGMT methylation status. The most promising results with respect to IDH mutation prediction were achieved recently in a combined segmentation-classification algorithm by

Bangalore Yogananda et al. achieving 97% accuracy on the TCIA dataset [96]. These remarkable capabilities of convolutional neural networks on small glioma datasets lay the groundwork for answering the remaining diagnostic questions in glioma research with a combination of convolutional neural networks and MRI.

## 5. Presurgical identification of genetic alterations in glioma

In this chapter, we synthesize ideas introduced in the first four chapters: 1) the importance of genetic subtypes and presurgically identifying the genetic alterations; 2) the ability of MR to provide insight into the underlying biology of glioma; and 3) the power of deep learning for medical image classification. We use these ideas to examine the MR modalities and deep learning frameworks that are best for predicting the IDH mutation and 1p19q codeletion.

### 5.1 Introduction

From 2007 to the present, the World Health Organization (WHO) categorization of gliomas has been restructured to include variations in underlying genetic and epigenetic alterations [97]. During 2019-2020, the cIMPACT-NOW consortium that informs the WHO, has placed even greater emphasis on the delineation of glioma categories by a mutation in the isocitrate dehydrogenase 1 and/or 2 (IDH1 and/or 2) and co-deletion of 1p and 19q chromosomal arms, prioritizing these features over grade[5,6,98,99]. In contrast to the WHO 2016 guidelines that use genetic alterations to further stratify patients within a designated grade, cIMPACT-NOW suggests that the first diagnostic delineation should rely on IDH mutation, followed by 1p19q codeletion status, as supported by evidence that these distinct genetic subtypes indicate drastic differences in overall survival and response to therapy[21,100–102]. Due to this increasing emphasis on genetic alterations as a diagnostic tool, it has become a clinical standard to perform genetic testing on tissue acquired during surgery to decide subsequent treatment.

Because genetic testing can be a costly and timely process and there remain cases where resection is not recommended, an alternative approach for obtaining this crucial genetic

information noninvasively is highly attractive. With a growing body of evidence that imaging features from MR are predictive of genetic alterations in IDH and 1p19q codeletion[47,48,96,103–108], image analysis techniques have the potential to provide a fast, noninvasive complementary pathway for identifying genetic alterations. Once these features are identified, the next step is to automate their extraction and determine the optimal strategy for combining them to improve classification of tumors into their genetic subtype. Several prior studies have implemented radiomics, machine learning, and/or deep learning to accomplish this task. In 2017, Li et al. reported that the automatic extraction of radiomic features using deep learning successfully predicts IDH mutation status in grade 2 glioma; however, this study was limited by requiring *a priori* knowledge of the tumor grade obtained through pathological tissue evaluation, limiting its application in the presurgical setting. It was also marked with an uncommon enrichment of IDH WT grade II glioma in their patient cohort[109]. Since then, many studies have leveraged The Cancer Imaging Archive either alone or together with internal datasets to evaluate the ability of deep learning and radiomics to predict a patient’s IDH mutation[96,110–113]. All of these studies use only anatomical MR imaging, which is advantageous in that they are universally acquired with standard imaging protocols and require minimal preprocessing, but lack the associated benefits of physiological imaging that more closely reflect the underlying tumor biology. Even less emphasis has been placed on predicting 1p19q codeletion, with only 2 of these studies reporting attempts to separately classify this mutation. We hypothesize that using a 3-class model that predicts genetic subgroup rather than individual mutation status plus a strategy that incorporates models that have been pretrained on classifying large, publicly available images, will improve the accuracy over prior tiered approaches for predicting IDH and 1p19q mutations because of the shared imaging features of these mutations.

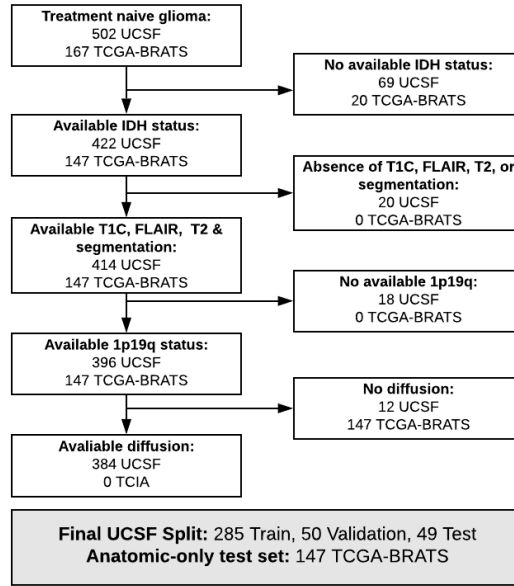


As diffusion-weighted imaging has become a standard in mainstream clinical imaging of gliomas at most institutions, there is a growing body of evidence that features derived from MR diffusion-weighted imaging are predictive of both IDH mutation and 1p19q codeletion [46–48,114]. As part of this work, we also sought to evaluate whether the addition of maps of Apparent Diffusion Coefficient (ADC) derived from diffusion-weighted imaging that are indicative of tumor cellularity would improve both the accuracy and generalization to an unseen test set when included as one of the inputs to a deep convolutional neural network (CNN) trained to predict genetic subtype. Because the “T2-FLAIR mismatch” signal has been shown to identify IDH mutated gliomas with 1p19q intact [103,104], we hypothesize that using T2 imaging together with FLAIR and ADC will improve the accuracy of the IDH-mutant, 1p19q intact subgroup, whereas including post-contrast T1-weighted images (T1c), ADC, and either the T2-weighted or T2-FLAIR images will improve the classification of IDH-wildtype and IDH-mutant, 1p19q codeleted subgroups.

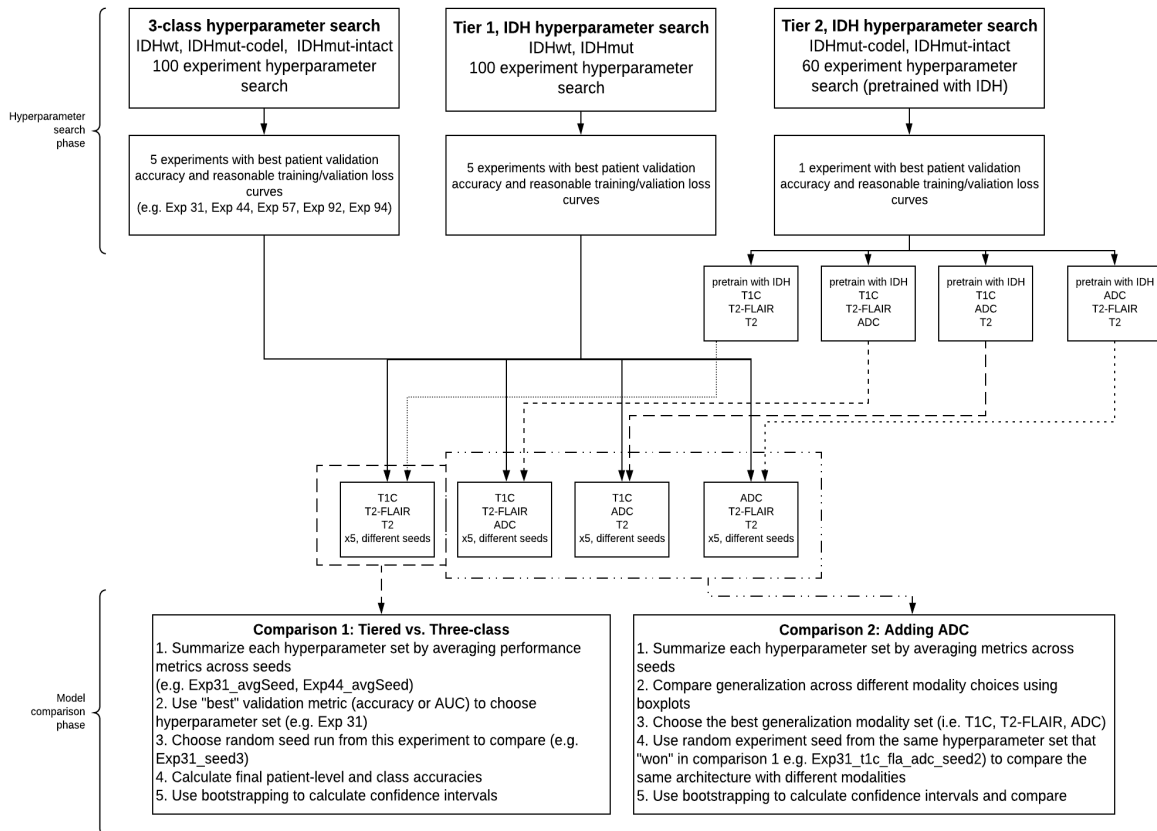
## 5.2 Methods

### 5.2.1 Patient Characteristics and Study Design

Imaging, pathological, and clinical data from a total of 502 adult patients that were newly-diagnosed with a pathologically confirmed glioma at our institution between 2007 and 2019 were assessed in this retrospective, IRB approved study. Patients were excluded if either their IDH status was indeterminable ( $n = 69$ ) or their preoperative MRI acquisitions did not include T1-weighted post-contrast (T1c), T2-weighted (T2), or T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) images ( $n = 20$ ) (Figure 5.1). The study design comprised two parts: hyperparameter search and model comparison phases (Figure 5.2). The hyperparameter search phase was first performed to select the best hyperparameters for each of our three deep learning models before our model comparison phase then investigated the benefits of 1) using a 2-tiered, binary classification approach opposed to a single-tiered, 3-class classifier, and 2) including apparent diffusion coefficient (ADC) images as an input channel. The best performing models were then tested on 147 BRATS images from the independent TCGA dataset to determine generalization to external, multi-institutional cohorts.



**Figure 5.1. Inclusion and exclusion criteria that led to the final numbers used in the study.**



**Figure 5.2. Study design detailing the hyperparameter search and model comparison phase.**

### 5.2.2 Assessment of genetic alterations

IDH mutation status for UCSF cases was evaluated by Sanger sequencing of IDH1 and IDH2 genes or by IHC (IDH1R132H, H09, Dianova GmbH, Hamburg, Germany) using standard techniques. Negative IDH mutation results based on immunohistochemistry (IHC) were either validated by sequencing or excluded. IDH mutation status for all TCGA cases was assessed via Sanger sequencing. Confirmed negative IDH1 and IDH2 mutated samples were classified as IDH-wildtype (“IDHwt”).

Our IDH mutated tumors were further classified into either “oligo-like” or “infiltrating astro-like” molecular subgroups based on either 1p19q codeletion status or ATRX alterations for UCSF cases, or solely 1p19q co-deletion status for TCGA data. Since tumors with 1p/19q codeletion (“IDHmut-codel”) almost invariably have IDH and TERT promoter mutations and are almost mutually exclusive with ATRX mutations, IDHwt gliomas and IDH-mutant (“IDHmut”) gliomas with ATRX alterations were not tested for 1p/19q codeletion unless it was performed clinically. ATRX was assessed by IHC (HPA001906, Sigma Aldrich, St. Louis, MO) performed at the UCSF Brain Tumor Research Center using previously published methods, while the presence of a 1p/19q codeletion was determined with clinical FISH assays [115]. IDHmut tumors that either had an ARTX alteration or were lacking a 1p19q codeletion were classified as 1p/19q-intact (“IDHmut-intact”).

### 5.2.3 Image Acquisition and Processing

All patients underwent MR examinations performed on a 3T MR750 GE scanner (GE Healthcare Technologies) using an eight-channel phased-array head coil within 48 hours prior to surgical resection. Standard anatomical imaging included T2-weighted FLAIR and fast spin echo

(FSE) images, along with 3D T1-weighted IR-SPGR imaging pre- and post- the injection of a gadolinium-based contrast agent. Diffusion-tensor images (DTIs) were obtained in the axial plane with  $b = 1000 \text{ s/mm}^2$  and either 6 gradient directions and 4 excitations or 24 gradient directions and 1 excitation or  $b = 2000 \text{ s/mm}^2$  and 55 gradient directions (repetition time [TR]/echo time [TE] = 1000/108 ms, voxel size =  $1.7\text{--}2.0 \times 1.7\text{--}2.0 \times 2.0\text{--}3.0 \text{ mm}$ ). To calculate the ADC map, a pipeline that utilized components of FMRIB's Diffusion Toolkit was applied to estimate relevant diffusion parameters from the DWI and DTI data as previously described [116].

All images from the UCSF cohort were registered to the T1c image volume using either FMRIB's FSL Linear Image Registration Tool (FLIRT) or Slicer's BRAINSFit tool with B-spline warping, and resampled to an identical 1-mm isovoxel spatial coordinate [116–118]. Brain masks were derived using the Brain Extraction Tool (BET) (FSL, FMRIB) and were visually verified to have worked properly [116]. All images were subjected to signal intensity normalization through a multistep process: (i) the images were multiplied by the brain mask, (ii) pixels above the 99.9 percentile were thresholded to the pixel intensity denoting the 99.9th percent; (iii) the mean was subtracted from each pixel and the result divided by the standard deviation; (iv) the images were scaled to lie between a value of 0 and 1 by subtracting the minimum and dividing by the difference between the maximum and the minimum pixels. The T2-lesion (T2L), defined as hyperintense signal on FLAIR images, and contrast enhancing lesion (CEL), or hyperintense signal on the T1c images that was not enhancing on the original T1-weighted images, had been previously contoured for the UCSF dataset using either 3Dslicer, or in-house software [119]. The 2019 BRATS dataset from the TCGA cohort were already pre-processed, segmented, and curated as part of this publicly available imaging dataset for annual

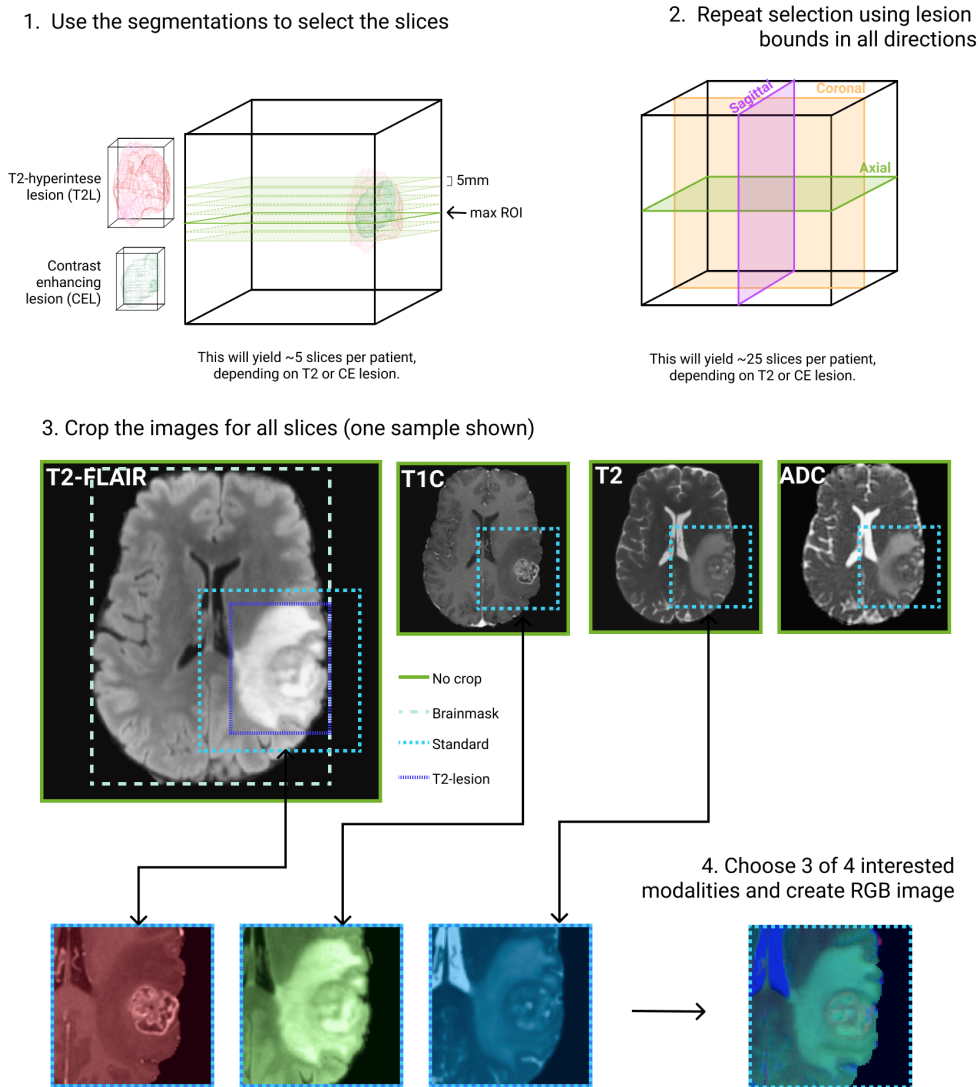
challenges as described previously [95,120,121].

Input images containing tumor were automatically selected and processed to form multi-contrast RGB colormaps according to Figure 5.3. Masks of the 3D segmented lesion volumes were used to first select the slice containing the largest tumor area in each direction. Additional slices spaced 5mm apart were added in each direction until the edge of the tumor mask was reached. Images were also automatically cropped to a rectangular bounding box surrounding each lesion of interest. For each cropping strategy, combinations of three image modalities were then merged to create a multi-contrast RGB image for each lesion slice.

#### 5.2.4 Baseline models from clinical metrics

Since age and the presence or absence of contrast enhancement are known predictors of IDH mutation status and anatomical MRI images of the brain can predict age with high accuracy [122,123], we first used basic logistic regression models in the sklearn package to establish an interpretable baseline prediction accuracy for which to compare our models [124]. Age and the presence of contrast enhancement were included as independent variables and used in: 1) a tiered binomial logistic regression structure to predict IDH mutation status followed by 1p19q codeletion status and 2) a 3-class multinomial logistic regression model to directly predict molecular subgroup. The presence of contrast enhancement was automatically quantified as having a CEL volume greater than 150 mm<sup>2</sup>, the cutoff for the lower 10th percentile, and included in the first tier and 3-class models. The 3-class model minimized multinomial loss fit across the entire probability distribution and balanced class weights. We also tested whether the two datasets (UCSF and TCGA) inherently differed significantly from each other using the

Mann-Whitney U test for age and the  $\chi^2$  test for categorical variables sex, mutation status, and presence of contrast enhancement.



**Figure 5.3. Schematic of image processing strategy.**

(1) The manual segmentation of the contrast enhancing lesion (CEL) or the T2-lesion (T2L) was used to select the slices; the “central slice” was the slice containing the maximum area determined by the segmentation and all following slices were selected by emanating every 5mm until the boundary of the lesion was reached. (2) If selected for during the hyperparameter search, the process was repeated in each direction: axial, coronal and sagittal. Some networks used just one direction, others used all three. (3) Another hyperparameter (“crop”) determined whether the image was cropped to the brainmask, to the T2L, to a standard size, or was not cropped. (4) The modalities of interest (e.g. three of T2-FLAIR, T1C, T2, and ADC) were placed in the R, G, and B channels of an image and used as input to the network.

### 5.2.5 Hyperparameter search

In order to find a reasonable starting point to train our models, we first searched through a set of randomly generated hyperparameters that included various model architectures, learning parameters, and image pre-processing strategies to find a reasonable starting point for each of our 3 classification models: IDH mutation only, 1p19q co-deletion only, and 3-class molecular subgroups. The individual hyperparameters that were tested are listed in Table 5.1, while hyperparameters that remained fixed were the learning rate cycling strategy (“One Cycle”),[125], the optimization algorithm (“Adam”)[126], and the weight decay coefficient (0.01). During this phase, model inputs were restricted to the T2, T2-FLAIR, and T1c images and the 1p19q co-deletion experiments began with the model pre-trained on the IDH mutation status classification. Overall patient accuracy and quality of the validation loss curves were used to evaluate the models’ efficacy. The top hyperparameter sets from each outcome were rerun 5 times with different seeds to account for stochasticity introduced during gradient descent. The set of hyperparameters with the best performance based on the mean overall classification accuracy of the 5 seeds on the validation set that also had acceptable training and validation loss curves was then chosen (see Figure 5.4 for examples of acceptable vs. unacceptable training/validation loss curves). The details describing the hyperparameter search phase and the model comparison phase, are found in Figure 5.1 and Table 5.1.

### 5.2.6 Model comparison

Using the set of hyperparameters for the top performing model for each classification experiment determined during the hyperparameter search phase, we investigated the impact of using a 2-tiered vs single 3-class structure approach and the addition of ADC as one of the input



image channels. As in the hyperparameter search phase, training and validation loss plots were visually compared in order to ensure that there was appropriate reduction in validation loss as the model trained (Figure 5.4). For the 2-tiered approach, IDH mutation status was predicted in the first tier, while 1p19q codeletion was then classified from the IDH mutated tumors in the second tier. The final performance accuracy for the IDHwt subgroup was determined from the output of the first tier while the accuracy of predicting the other 2 subgroups were determined by the prediction accuracies of the second tier. For each classification approach (2-tiered and 3-class), ADC maps were then included as one of the 3 input channels (either in place of the T1c, T2, or T2-FLAIR image volumes) and trained with the same set of hyperparameters run with 5 different seeds. On the final models, confidence intervals were calculated using bootstrapping 1000 times on the predictions.

Model explanation and feature attribution was performed using GradCAM, a heatmap-based feature attribution method. In contrast to methods that use “guided” back-propagation as a part of feature attribution, GradCAM has been validated in deep learning literature to assign feature importances to areas of the image better than random [127]. This was the most appropriate feature attribution technique because it allows for quick visual confirmation that the model is behaving as expected by extracting features in the areas that align with human interpretation. We used these GradCAM maps to help interpret our best and worst predicted patient examples in order to gain insight into the models’ behavior in these cases.

### 5.2.7 Model generalization to an independent test set

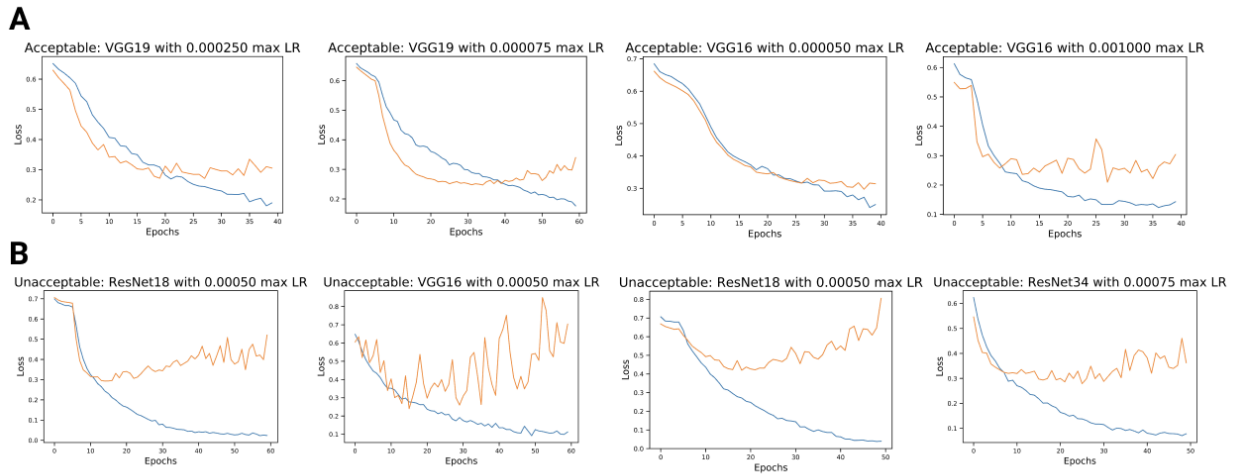
In order to evaluate whether our developed anatomical models were able to generalize to data from multiple institutions acquired using different scanners and acquisition parameters that result in variations in image contrast and resolution, the publicly available TCGA dataset together

with the post-processing and labeling performed for the BRATS challenge were used to establish an independent dataset for testing. The BRATS imaging dataset was preprocessed with the same specifications as our data with expert segmentations, but because this dataset did not include diffusion data, we were only able to validate our best anatomical 2-tiered and 3-class models with this dataset.

**Table 5.1. Hyperparameter search space.**

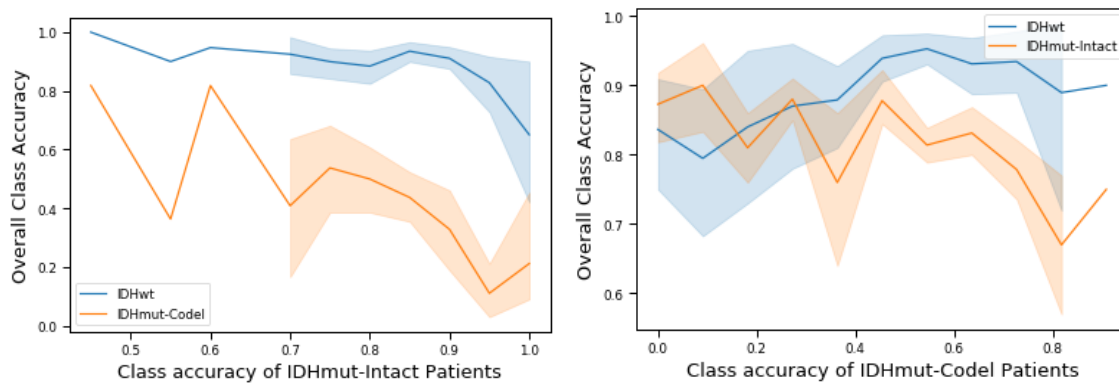
All of the potential choices for hyperparameters during the search phase.

Architecture	VGG16	VGG19	ResNet18	ResNet34									
Pretrained	Yes	No											
Slice selection	5mm	10mm	Only center slice										
Crop selection	T2-hyperintense lesion	Standard crop (varied pixel size)	Brainmask	No crop									
Learning rate	1.00E-05	2.50E-05	5.00E-05	7.50E-05	1.00E-04	2.50E-04	5.00E-04	7.50E-04	1.00E-03	2.50E-03	5.00E-03	7.50E-03	1.00E-02
Data augmentation strategy	"Light" - vertical flip: p=0.5; horizontal flip p=0.5; random rotation: -20 and 20 degrees;		"Medium" - vertical flip: p=0.5; horizontal flip p=0.5; random rotation: -20 and 20 degrees; random affine: -20 and 20 degrees+0.5 translation		"Heavy" - vertical flip: p=0.5; horizontal flip p=0.5; random rotation: -20 and 20 degrees; random affine: -20 and 20 degrees+0.5 translation; color jitter: brightness = 0.1, contrast = 0.1, saturation = 0.2, hue = 0.1								
Final resampling size during data augmentation	100	120	140	160	180	200	220						
Dropout probability between the convolutional and fully connected layers	0	0.1	0.2	0.3									
Parameter freezing strategy	Train all at once	Freeze all but FC layers	Unfreeze all but FC layers at certain number of epochs										
Plane	Coronal	Sagittal	Axial	All 3									
How to combine slice predictions	Max pooling	Average pooling	Majority voting	Averaging	Weighted averaging based on distance from max slice								



**Figure 5.4. Acceptable and unacceptable training/validation loss curves.**

Training/validation loss curves were visually inspected in order to ensure that there was reduction in both training and validation loss throughout training. (A) Examples of training (blue) and validation (orange) loss curves that were acceptable. (B) Examples of training (blue) and validation (orange) loss curves that were rejected.



**Figure 5.5. 3-class model development tradeoff.**

## 5.3 Results

### 5.3.1 Characteristics of the Study Sample

The clinical characteristics of the entire dataset, consisting of 384 patients from UCSF and 147 patients from the TCGA dataset, are summarized in **Table 5.2**. While sex was not statistically significantly different between the two cohorts (59% male in UCSF vs 52% male in BRATS), patients at UCSF were statistically significantly younger than patients in the TCGA data set, with mean age of  $47.4 \pm 15.3$  years compare to  $53 \pm 14.9$  years ( $p < 0.001$ ). UCSF and TCGA datasets also significantly differed in the proportion of IDH mutation status, with 269 mutated UCSF patients (62%) and 56 mutated TCGA patients (22.8%),  $p < 0.00001$ ) and similarly the frequency of enhancement ( $p < 0.00001$ ), with 206 UCSF patients enhancing (47%) and 120 TCGA patients enhancing (82%).

**Table 5.2. Patient demographics and differences between UCSF and TCGA data.**

		UCSF n=384		TCIA n=147		Difference between UCSF and TCGA-BRATS	
		Mean	Std	Mean	Std	Mann-Whitney U p-value	p-value (Chi <sup>2</sup> )
	<b>Age</b>	47.7	15.6	53	14.9	0.0074	n/a
	<b>Sex</b>	Male	Female	Male	Female	0.05	0.088
		231	153	76	70		
	<b>IDH status</b>	IDHwt	IDHmut	IDHwt	IDHmut		
		151	233	87	59	4.71E-06	2.78E-05
	<b>WHO Grade</b>	Non-enhancing	Enhancing	Non-enhancing	Enhancing		1.02E-12
<b>2</b>	<b>IDHwt</b>	0	1	2	0		
	<b>IDHmut-intact</b>	69	4	11	8		
	<b>IDHmut-codel</b>	63	8	3	3		
<b>3</b>	<b>IDHwt</b>	6	1	2	7		
	<b>IDHmut-intact</b>	54	5	7	14		
	<b>IDHmut-codel</b>	6	7	1	6		
<b>4</b>	<b>IDHwt</b>	0	143	0	76		
	<b>IDHmut-intact</b>	1	13	0	6		
	<b>IDHmut-codel</b>	0	2	0	0		

### 5.3.2 Clinical baseline modeling results

The detailed results of baseline clinical logistic regression models built on UCSF training patients and validated on the UCSF validation, UCSF test, and TCGA test patients are presented in Table 5.3. The tiered and 3-class results were very similar, achieving 67% and 71% average accuracy on the UCSF test set, respectively, which served as the basis for comparison for our deep learning models. However, for both the tiered and 3-class baseline models, the bootstrapped confidence intervals were very wide for every metric, with the best prediction accuracy achieved for IDH-wt tumors (95%/91% for the UCSF/TCGA test sets) and worst for the IDHmut-codel group (40%/23% for the UCSF/TCGA test sets).

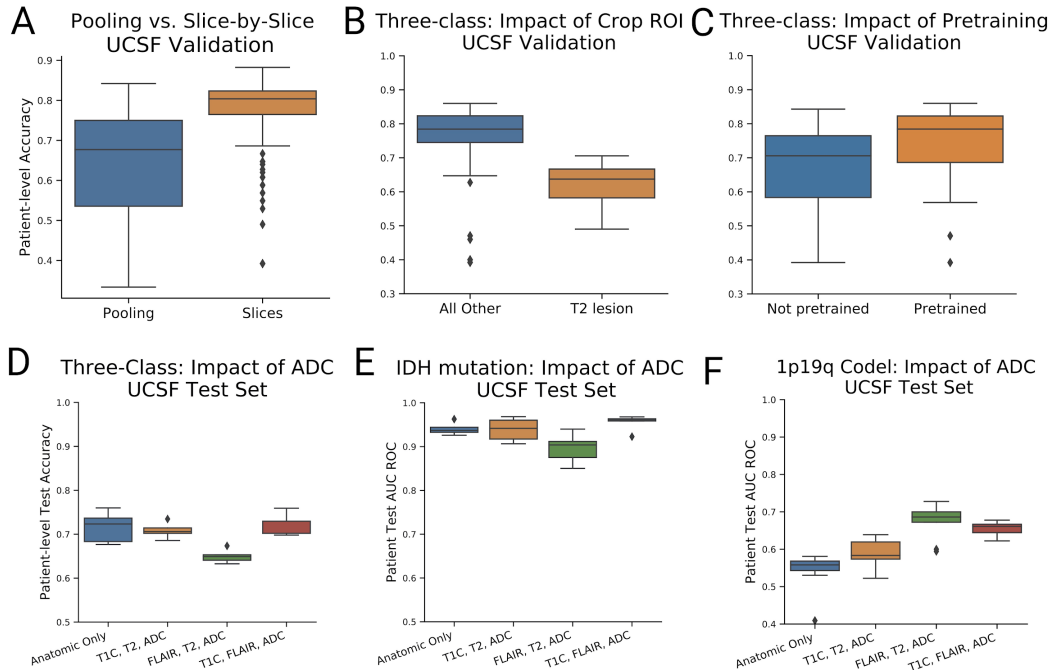
**Table 5.3. Results of all six models compared.**

2-tiered logistic regression	IDHwt	IDHmut-intact	IDHmut-codel	Overall	3-class Logistic regression	IDHwt	IDHmut-intact	IDHmut-codel	Overall
<b>Training</b>					<b>Training</b>				
Accuracy	93.8%	62.6%	51.5%	72.3%	Accuracy	95.5%	54.2%	53.0%	70.2%
Confidence Interval	[0.895, 0.973]	[0.547, 0.707]	[0.412, 0.621]	[0.681, 1.0]	Confidence Interval	[0.921, 0.983]	[0.46, 0.62]	[0.429, 0.634]	[0.656, 1.0]
<b>Validation</b>					<b>Validation</b>				
Accuracy	100.0%	70.0%	45.5%	76.0%	Accuracy	100.0%	55.0%	54.5%	72.0%
Confidence Interval	[1.0, 1.0]	[0.524, 0.857]	[0.222, 0.727]	[0.66, 1.0]	Confidence Interval	[1.0, 1.0]	[0.364, 0.739]	[0.286, 0.818]	[0.620, 1.0]
<b>Test</b>					<b>Test</b>				
Accuracy	95.0%	47.4%	50.0%	67.3%	Accuracy	95.2%	61.1%	40.0%	71.4%
Confidence Interval	[0.864, 1.0]	[0.538, 0.889]	[0.125, 0.667]	[0.653, 1.0]	Confidence Interval	[0.857, 1.0]	[0.364, 0.739]	[0.286, 0.818]	[0.571, 1.0]
<b>TCGA</b>					<b>TCGA</b>				
Accuracy	90.1%	25.6%	23.1%	65.3%	Accuracy	91.2%	23.3%	23.1%	65.3%
Confidence Interval	[0.825, 0.935]	[0.321, 0.575]	[0.071, 0.444]	[0.63, 1.0]	Confidence Interval	[0.916, 1.0]	[0.118, 0.333]	[0.059, 0.444]	[0.596, 1.0]
<b>2-tiered anatomical</b>					<b>3-class anatomical</b>				
<b>Training</b>					<b>Training</b>				
Accuracy	97.3%	95.3%	86.4%	94.0%	Accuracy	92.9%	95.3%	53.0%	84.6%
Confidence Interval	[0.946, 0.973]	[0.916, 0.963]	[0.742, 0.848]	[0.902, 1.0]	Confidence Interval	[0.902, 0.938]	[0.925, 0.962]	[0.492, 0.591]	[0.826, 1.0]
<b>Validation</b>					<b>Validation</b>				
Accuracy	89.5%	90.0%	63.6%	84.0%	Accuracy	100.0%	65.0%	81.8%	82.0%
Confidence Interval	[0.895, 0.947]	[0.850, 0.900]	[0.455, 0.636]	[0.780, 1.0]	Confidence Interval	[0.947, 1.0]	[0.5, 0.65]	[0.727, 0.909]	[0.74, 1.0]
<b>Test</b>					<b>Test</b>				
Accuracy	81.0%	77.8%	30.0%	69.4%	Accuracy	90.5%	77.8%	70.0%	81.6%
Confidence Interval	[0.714, 0.810]	[0.684, 0.842]	[0.100, 0.400]	[0.633, 1.0]	Confidence Interval	[0.9, 0.95]	[0.579, 0.737]	[0.6, 0.9]	[0.735, 1.0]
<b>TCGA</b>					<b>TCGA</b>				
Accuracy	91.2%	86.0%	0.0%	81.6%	Accuracy	95.6%	32.6%	0.0%	68.7%
Confidence Interval	[0.868, 0.912]	[.744, 0.857]	[0.0, 0.077]	[0.769, 1.0]	Confidence Interval	[0.956, 0.967]	[0.279, 0.372]	[0.0, 0.077]	[0.678, 1.0]
<b>2-tiered with ADC</b>					<b>3-class with ADC</b>				
<b>Training</b>					<b>Training</b>				
Accuracy	97.3%	96.3%	80.3%	93.0%	Accuracy	92.0%	94.4%	62.1%	86.0%
Confidence Interval	[0.946, 0.973]	[0.916, 0.963]	[0.712, 0.803]	[0.891, 1.0]	Confidence Interval	[0.911, 0.938]	[0.916, 0.963]	[0.561, 0.652]	[0.839, 1.0]
<b>Validation</b>					<b>Validation</b>				
Accuracy	89.5%	90.0%	45.5%	80.0%	Accuracy	94.7%	80.0%	54.5%	80.0%
Confidence Interval	[0.895, 0.947]	[0.800, 0.900]	[0.364, 0.545]	[0.760, 1.0]	Confidence Interval	[0.895, 0.947]	[0.7, 0.8]	[0.364, 0.545]	[0.72, 1.0]
<b>Test</b>					<b>Test</b>				
Accuracy	81.0%	83.3%	50.0%	75.5%	Accuracy	95.2%	88.9%	60.0%	85.7%
Confidence Interval	[0.700, 0.800]	[0.684, 0.842]	[0.221, 0.500]	[0.633, 1.0]	Confidence Interval	[0.857, 0.952]	[0.778, 0.944]	[0.4, 0.6]	[0.771, 1.0]

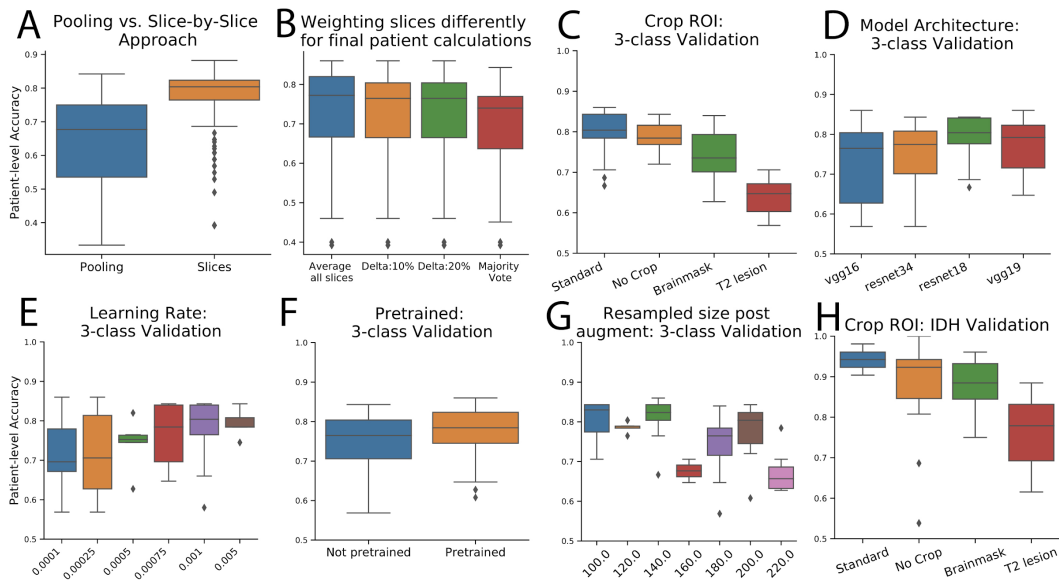
### 5.3.3 Model development results

The first comparison made was whether pooling slice predictions was advantageous compared with slice-by-slice predictions. Briefly, slice-by-slice predictions on average achieved higher patient-level validation accuracies compared with pooling, and the rest of the experimentation used slice-by-slice predictions only (**Figure 5.6a**). Once the slice-by-slice paradigm was chosen, it was possible to create patient-wise predictions from slices using many strategies. **Figure 5.7b** depicts that using a simple mean average achieved better or similar patient validation accuracy compared with other strategies. All subsequent metrics were then evaluated with slices combined by averaging the probability predictions per slice.

We observe the impact of top features on the patient validation accuracy in **Figure 5.7c-g**. Our main finding from this analysis is that in both three-class and IDH experiments, cropping to the T2-hyperintense lesion decreases performance. Additionally, using pretrained models on ImageNet improves the performance on models.



**Figure 5.6. Main results from the hyperparameter search and the model comparison phase.** (A-C) Hyperparameter search: (A) A slice-by-slice approach improved the ability to achieve high accuracy on the validation data compared with the average pooling approach. (B) Cropping to the T2-lesion hindered the ability of a 3-class model to achieve good validation accuracy. (C) Pretraining increased the ability of a model to achieve high accuracy on the validation set. (D-F) Model comparison: (D) For the 3-class models, lower generalization accuracy was observed when TIC was removed; the best performance was obtained with TIC, T2-FLAIR, and ADC imaging. (E) For the IDH-only tier, using TIC, T2-FLAIR, and ADC was slightly favored compared with all other modality combinations. (F) For the 1p19q-codeletion prediction, T2-FLAIR, T2, and ADC performed best.



**Figure 5.7. Additional insights from the hyperparameter search phase.**

### 5.3.4 Model comparison

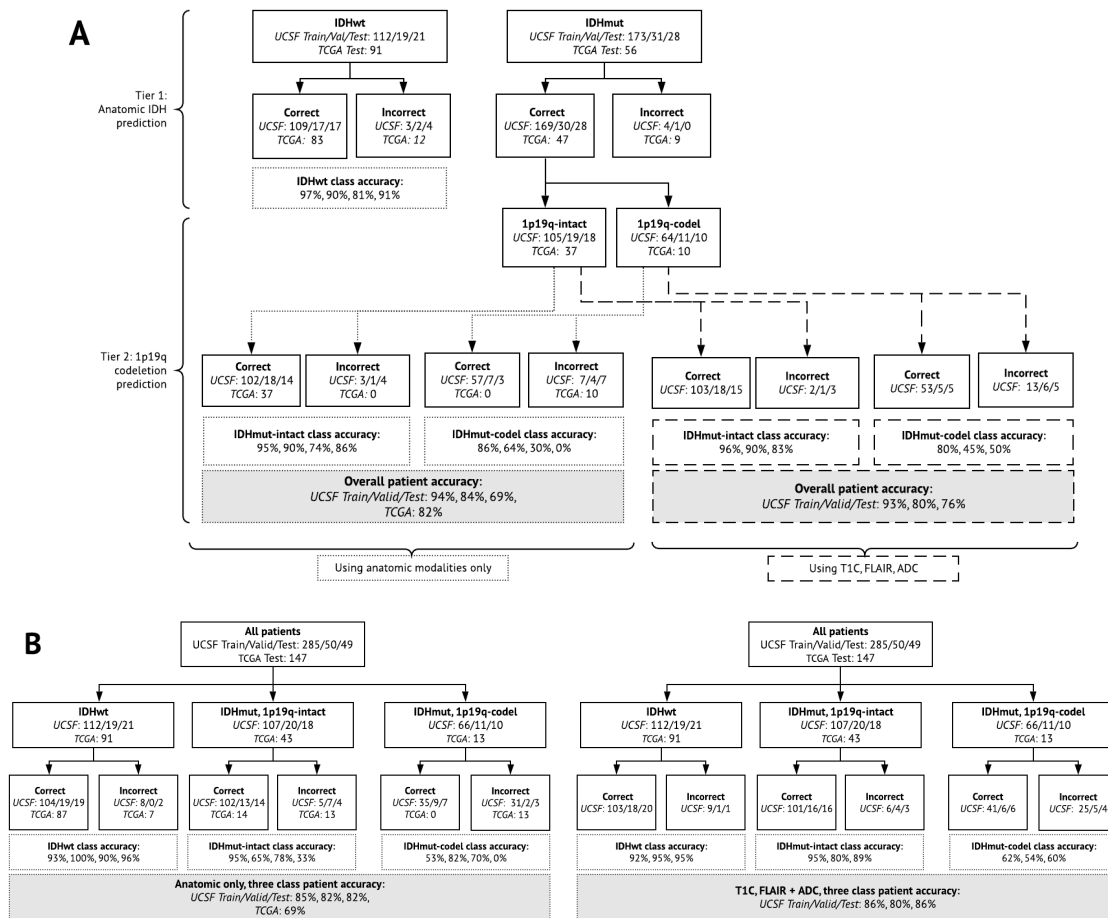
#### 5.3.4.1 Two-tiered vs. Single three-class classifier

When using anatomical images only, the best 3-class model resulted in an overall patient accuracy of 84.6%, 82.0% and 81.6% for the training, validation and testing sets of UCSF data, with individual test class accuracies of 90%, 78%, and 70% for the IDHwt, IDHmut-intact, and IDHmut-codel subgroups, respectively. The final model parameters are shown in **Table 5.4**. Although the best performing 2-tiered structure resulted in a higher overall patient accuracy in training (94%) and relatively similar accuracy as the 3-class model in validation (84.0%), this model did not generalize as well to the UCSF test set (69.4% accuracy). Detailed class accuracies for each training, validation, and testing cohort along with confusion matrices are shown in **Table 5.3**, while the final predictions of the 2-tiered vs. 3-class models are shown in **Figure 5.8**. However, when evaluating the ability of each approach to generalize to the multi-institutional TCGA data, the 2-tiered structure outperformed the 3-class model (82% compared to 69% overall accuracy). Although both of these approaches were able to predict the IDHwt group with high accuracy (96% for 2-tiered and 91% for 3-class), the 2-tiered model was also able to predict the IDHmut-intact subgroup with 86% accuracy while the 3-class model accuracy was only 33% for this subtype. Both approaches failed at correctly predicting any of the tumors in the IDHmut-codel subgroup.



**Table 5.4. Final hyperparameters of each of the deep learning models in the model comparison phase.**

	<b>Final 3-class</b>	<b>Final IDH</b>	<b>Final 1p19q-Codel</b>
<b>Network Architecture</b>	VGG16	VGG16	ResNet18
<b>Pretraining on natural images</b>	TRUE	TRUE	True, Then IDH
<b>Planes</b>	All planes	All planes	Axial
<b>Slice cropping</b>	Standard size	Standard size	T2-Hyperintense ROI
<b>Max learning rate (with OneCycle policy training)</b>	0.001	0.001	0.0005
<b>Data Augmentation</b>	Medium	Heavy	Heavy
<b>Resampling size during augmentation</b>	100 px	100 px	140 px
<b>Dropout probability between convolutional and fully connected layers</b>	0.1	0.2	0.3
<b>Number of epochs trained</b>	60	60	40
<b>Parameter freezing/unfreezing strategy</b>	Unfreeze all parameters at 12th epoch	Unfreeze all parameters at 6th epoch	Unfreeze at 8th epoch



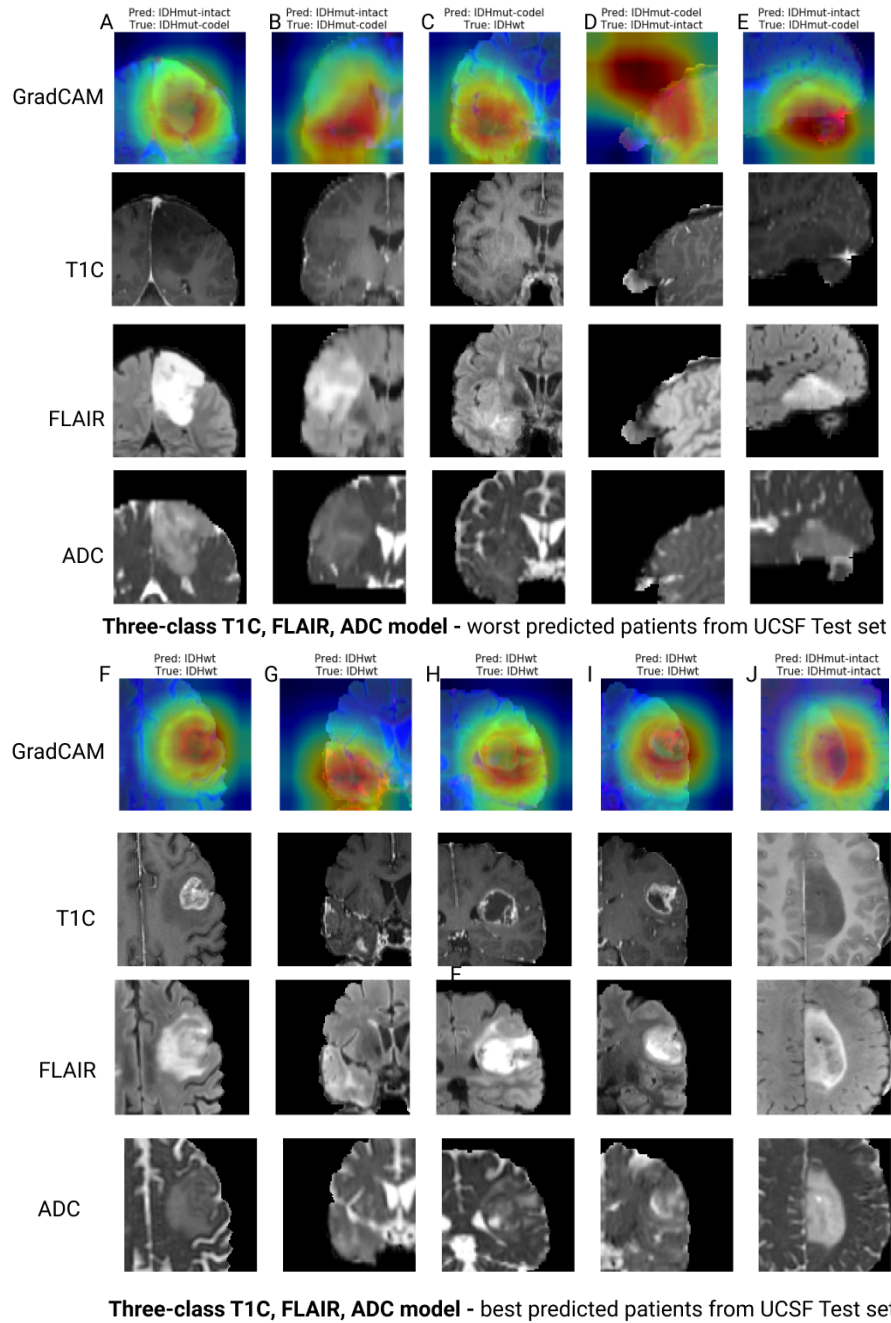
**Figure 5.8. Detailed patient classification of each of the final models.**

(A) Using a tiered structure, we assess the impact on patient classification using ADC in place of T2 imaging. IDHmut-codel accuracy increases in the test set, and the overall UCSF test set accuracy increases. (B) Using a 3-class structure, we compare the performance of anatomic-only and T1c, T2-FLAIR, and ADC modalities. We notice an increase in the IDHmut-codel accuracy as well as overall increase in UCSF Test set accuracy. The best model is the 3-class model that includes ADC.

### 5.3.4.2 Benefit of adding ADC

We next investigated whether the addition of ADC was advantageous compared with using anatomical images only as inputs to the model for both the 3-class and 2-tiered approaches. In both the 3-class and IDH models (first tier), the best performance was achieved when ADC maps were used along with T1c and T2-FLAIR images as inputs, while replacing the T1c image with ADC decreased performance from using anatomical imaging alone (**Figure 5.7D-E**). For

the 1p19q codeletion classification (2<sup>nd</sup> tier) of the 2-tiered approach, however, the best performance was achieved when ADC replaced the T1c image (**Figure 5.7F**), resulting in an increase in test accuracy from 60.7% to 70.6% for this tier with the largest improvement observed for the IDHmut-codel subgroup from 30% to 50% (**Table 5.3**). Although, in general, including ADC in both models outperformed their anatomical imaging-only counterparts, the best generalization power to the test set was achieved with a 3-class model that replaced T2-weighted images with ADC maps. The final overall patient accuracies achieved were 86%, 80%, and 86% on training, validation, and test UCSF sets, with final test set class accuracies of 95% for IDHwt, 89% for IDHmut-intact, and 60% for IDHmut-codel subgroups (**Figure 5.8B and Table 5.3**). **Figure 5.9** illustrates the GradCAM results for the best (>90% confidence that the genetic alteration was ground truth) and worst (>50% confidence that the genetic alteration was other than its ground truth) predictions for the best 3-class model with ADC, T1c, and T2-FLAIR images as inputs. Incorrectly predicted IDHwt tumors were non-enhancing, while IDHmut-codel tumors were frequently predicted as IDHmut-intact because the network was not looking at the right part of the image.



**Figure 5.9. GradCAM analysis.**

GradCAM was performed using the final 3-class model on the worst (A-E) and best (F-J) predicted patients from the UCSF test set. The worst predicted patients tend to have GradCAM maps that are not looking precisely at the tumor region compared with the well-predicted patients.

## 5.4 Discussion

In this study, we systematically investigated different sets of hyperparameters to achieve the optimal deep learning framework and MRI modalities for jointly identifying the IDH mutation and 1p19q codeletion status of a glioma patient prior to surgery. To our knowledge, this is the first study to: a) classify molecular subgroup using imaging and deep learning; b) investigate the impact of including ADC; c) incorporate a pre-training strategy that includes the generation of an RGB color image from 3 grayscale MR images; and d) thoroughly evaluate differences between various deep learning strategies through extensive hyperparameter searching complemented by training/validation loss curves and a feature attribution technique. Our best performing model was a 3-class, VGG-16 model pre-trained on ImageNet that included T1c, T2-FLAIR, and ADC images as inputs and predicted patients in our UCSF test set with promising overall (85.7%) and individual class accuracy (IDHwt: 95.2% 95%CI (0.857, 0.952)]; IDHmut-intact: 88.9% [95% CI (0.778, 0.944)]; IDHmut-codel: 60.0% [95% CI (0.4, 0.6)]). A 3-class model approach was advantageous compared to a tiered strategy that first predicted IDH, and then 1p19q codeletion mutations. Adding ADC as one of the input images increased generalization to test sets for both the 3-class models and 2<sup>nd</sup>- tier 1p19q models. All deep learning models outperformed the corresponding logistic regression baseline models developed and tested on the same patient datasets, implying that imaging features can provide additional insight to genetic alterations compared to age and the presence of contrast enhancement alone. Our GradCAM analyses confirmed that the final algorithm was in fact learning features derived from tumor regions and not surrounding areas.

As age and the presence of contrast-enhancing tumor are known predictors of IDH mutation status, we constructed logistic regression models using these variables to serve as a benchmark for our models to outperform. This approach also ensured that the deep learner was

more than a complex detector of age or the presence of contrast enhancement [128]. Using contrast enhancement as an input to our baseline logistic regression models improved their generalization to the UCSF validation, UCSF test, and TCGA test sets for the 3-class and IDH model to 70-72% overall accuracy. These final patient and class accuracies served as a benchmark for which to compare our deep learning models.

Before implementing our hypothesis driven comparisons on the influence of ADC and modeling approach, we next performed an extensive hyperparameter search to determine the optimal set of network training parameters for each set of experiments. First, two different slice-combining paradigms were compared: 1) pooling slices for a single prediction per patient as performed in MRNet from Bien et al. [129]; and 2) treating each slice individually while training and combining slice predictions afterward as described by Chang et al. [112]. In (1), all slices from a single patient are used in the same batch such that the number of slices becomes the effective batch size. Average or max pooling is then employed in the final layer to condense all slices into a single feature vector which generates a single prediction per patient. As a result, a single value is back propagated through the network for each patient after the loss is calculated. In contrast, (2) treats each slice independently such that a batch often contains slices from many patients; in turn, backpropagating gradients on a slice-by-slice basis and calculating a final patient-level prediction only after training is complete. Updating network weights based on individual slices resulted in better training/validation loss curves as well as an increase in the overall patient-level accuracy as shown in Figure 5.2A. Also of note was that a marked decrease in prediction capability for both the 3-class and 2-tiered settings was observed when cropping to the T2 lesion compared with no cropping or cropping to a standard sized rectangle. This is in line with the notion that the location of the lesion within the brain is associated with IDH mutation status [130]. Although the

ResNet-18 architecture most frequently resulted in models with higher average accuracy, the VGG-16 architecture had the achieved the highest accuracy when other hyperparameters were optimized. This is not all that surprising given VGG16's 3 terminal fully-connected layers that could be helpful in capturing the heterogenous characteristics present in these lesions by allowing for different interactions among features, and the fact that the benefits of the residual connections in ResNet architectures typically are not realized until an order of magnitude of more data is used in training.

We hypothesized that a single 3-class model that was trained to predict molecular subgroup by classifying both IDH mutation status and 1p19q codeletion simultaneously would outperform a 2-tiered cascaded approach because: 1) the second tier predicting 1p19q codeletion had a limited number of patients from which to learn imaging features; and 2) learned imaging features could be shared between tasks. Our results supported this hypothesis, regardless of whether or not ADC was included in our models. The reduced overall training accuracy of the 3-class model also suggests that the model was less likely to overfit when capturing features of 3-classes, boosting its performance on the test set compared to the 2-tiered approach. **Figure 5.5** shows class accuracies plotted from the 3-class experiments that were performed during the hyperparameter search phase, depicting the tradeoff between a model's ability to predict IDHmut-codel patients and IDHmut-intact patients correctly. As the ability to predict IDHmut-codel patients increased, the prediction accuracy for IDHmut-intact patients diminished, while the ability to predict IDHwt patients remained stable. This result implies that even in the multiclass setting, the power of deep learning models to discriminate the 1p19q codeletion was still limited. Although the 2-tiered approach more accurately classified the TCGA cohort compared to the 3-class model, the second tier incorrectly predicted all of the TCGA IDH mutated patients as 1p19q intact, further supporting improved

generalizability with the 3-class model. This is likely because the enriched number of IDHwt and IDHmut-intact patients in the TCGA cohort compared to the UCSF data.

Using ADC in place of an anatomical imaging sequence conferred an advantage in test accuracy for both the tiered and 3-class setting (Table 5.3). This advantage was particularly evident when comparing experiments predicting 1p19q codeletion (Figure 5.8), where we observe the greatest generalization power in models using ADC together with T2 and T2-FLAIR. This finding was expected given that the mismatch in the T2 and T2-FLAIR signal contains imaging features specific to IDHmut-intact patients. When comparing the 3-class models with and without ADC, a more balanced result between IDH mutated classes was achieved with ADC: 70% IDHmut-codel and 74% IDHmut-intact accuracy compared with the 60% IDHmut-codel and 84% IDHmut-intact accuracy. In contrast, including ADC as a modality in place of either T1c, T2, or T2-FLAIR images did not confer an advantage in predicting IDH mutation status alone, despite prior evidence that features derived from diffusion-weighted imaging can help differentiated IDH mutation status [46,47,114]. This result, however, does not mean that ADC is not valuable, but rather that the loss of another more informative sequence outweighs the benefit of ADC. For both the 3-class and IDH model tier, replacing the T1c images with ADC substantially decreases generalization power to the UCSF test set as expected given the known association between presence of contrast enhancement and IDH wildtype tumors. Although a limitation of our study is that we were not able to validate these findings on the multi-institutional external cohort, the promise of incorporating ADC into deep learning models that predict molecular subgroup, especially for the 1p19q-mutation, is still clear from the results presented and a valuable contribution to the scientific community.



Despite the therapeutic relevance of 1p19q codeletion status, the vast majority of prior studies have focused exclusively on IDH mutation prediction. In 2018, a radiomics-based machine-learning algorithm utilized the BRATs portion of TCGA data to predict 1p19q codeletion vs. intact patients and achieved 80% accuracy [111]. The validation set used to assess this accuracy, however, consisted of only 5 subjects. Another study since reported a deep learning based 5-fold cross validation with remarkable 94% accuracy for the prediction of 1p19q codeletion in 2019 [110]. However, this study also lacked a separate test set for generating this metric and overall accuracy measures included IDH-wildtype tumors in the 1p19q-intact class, artificially boosting baseline accuracy to 88% even if all of the 1p19q-codeleted tumors were predicted incorrectly. There was also no specification about whether early stopping was employed, which in our experience, when used in conjunction with cross-validation approaches, results in a >30% drop in accuracy between the validation and test sets. Without reporting an independent test set or at least the loss curves observed during training and validation, it is not possible to assess whether a model would work on unseen data. Although van der Voort et al. in 2019 [131] used the BRATS images of the TCGA dataset as an external test cohort to validate their radiomics-based machine learning model of 1p19q-codeletion classification in low-grade glioma and demonstrated clinical relevance by comparing model results to the predictions of expert clinicians, they also did not first stratify by IDH before predicting 1p19q mutation status, elevating their accuracy in 1p19q-intact patients by 25% which translated to a 0.73 AUC ROC on the external test set. In contrast, our study is the first analysis that aims to predict 1p19q codeletion for patients already determined to be IDH mutant, without first segregating based on tumor grade obtained through pathology. Although the sample size of our IDH-mutated subgroups is still limited and our external test set does not reflect the same distribution of 1p19q-codeleted to intact patients as our internal dataset,

it is still the first deep learning study that attempts to validate models incorporating 1p19q codeletion status within IDH-mutant gliomas on an external, multi-institutional cohort.

Our results from the UCSF test cohort and downstream GradCAM analysis indicated that a deep learning model is in fact learning from signals in tumor regions and it is possible to learn generalizable imaging features when the patient samples are of similar enough outcome distribution. GradCAMs provide some amount of interpretability of an otherwise “black-box” CNN by displaying a combination of semantically meaningful features in the form of a heatmap, which can be thought of as a map of where the network is looking in the image to draw its predictions. Our analysis included the generation of GradCAM heatmaps for both well- and poorly-predicted patients for the best 3-class model, including ADC maps in place of T2-weighted images. The GradCAM heatmaps in Figure 5.9 provide confidence in our results because the network focuses on the lesion in all correctly predicted tumors, while in the patients who were misclassified, the GradCAM heatmaps show that the network often gets confused by other parts of the image outside the lesion boundaries that confound the overall prediction. Although GradCAMs provide insight into where the model is looking, they do not attribute importance of different features and have limited spatial resolution based on the size of the final output layer of the chosen model.

In conclusion, we created a model that was able to generalize to unseen data with a promising overall accuracy of 86%, and individual class accuracies of 95% for IDHwt, 90% for IDHmut-intact, and 60% for IDHmut-codel subgroups. From our extensive hyperparameter search during model development, we derived insights that support the use of a network that has been pre-trained on ImageNet for classification tasks combined with a slice-by-slice approach for updating weights in training and a cropping strategy that extends beyond the boundaries of the T2-

lesion. Using this framework, we concluded that classifying both IDH and 1p19q mutations together in a single step was advantageous compared to implementing a tiered structure that first predicted IDH mutation status before 1p19q codeletion using two separate binary models. The addition of ADC increased the generalization capacity of our models regardless of the modeling structure chosen, highlighting the utility of incorporating diffusion-weighted imaging in future multi-site analyses of molecular subtype. Although larger studies that focus on accumulating enough IDH-mutant patients are still desperately needed to improve the accuracy of 1p19q-codeleted gliomas for implementation in clinical practice, the insights gleaned from this study will be highly valuable once such datasets become publicly available.

## 6. Automatic classification of MR image contrast

### 6.1 Introduction

Automated quantification of data acquired as part of an MR exam requires identification of the specific series of relevance to a particular analysis. This motivates the development of methods capable of reliably classifying MR series according to their nominal acquisition contrast, e.g. T1-weighted (T1), T1 post-contrast (T1C), T2-weighted (T2), T2-weighted FLAIR (T2-FLAIR), proton-density weighted (PD). For example, a machine learning model may be trained to segment lesions from T2-weighted FLAIR images. To perform this analysis automatically requires robust programmatic identification of a T2-weighted FLAIR series within an MR exam comprising potentially tens of different acquisitions. Similarly, analysis of disease progression and response to therapy with MRI involves quantification of serial changes that occur on images acquired with similar tissue contrast, such as change in T1-weighted or T2-weighted lesion load [REF of examples]. In addition, development of new AI and machine learning models depends on the ability to train on large quantities of specific classes of data and therefore MR contrast classification models are an important element of content-based retrieval systems (CBRS, [132]) to enable large medical centers to leverage vast amounts of retrospectively acquired data for population-level computational health research.

Features that identify the acquisition contrast of an MR series may comprise both DICOM (<https://www.dicomstandard.org/>) header data and intrinsic properties of the imaging pixel data. However, identification of relevant series is not always straightforward for single MR exams let alone for longitudinal data in PACS at large medical centers which include data acquired on different scanners (manufacturer, field strength, software), at different sites and with

different scanning protocols. There is a vast range of physical scanning parameters associated with each acquisition type and moreover, descriptive attributes such as the DICOM “series description” are free-text attributes subject to local conventions, technologist choice and even human error. As a result, the header features associated with MR scans are extremely heterogeneous with DICOM attributes that often do not explicitly identify the type of acquisition, limiting the ability to automatically retrieve images acquired with the same contrast weighting.

This presents challenges for developing high-throughput methods for automated analysis of MR data. A driving use case for this work is the automated longitudinal analysis of MR exams from multiple sclerosis (MS) patients at UCSF [133]. MS is a chronic disease and patients are regularly monitored with MRI to assess progression and response to therapy, often for decades. The goal of this use case is to automatically provide a longitudinal, aligned view of the patient’s imaging data for each contrast type together with the patient’s medication, clinical and medication history in a unified view to aid patient-clinician consultations, relieving the clinician of the burden of synthesizing the vast amounts of data and providing it in a patient accessible view (Figure 6.1). This work thus requires retrieving the entire MR imaging record for each patient from PACS and identifying images with key types of tissue contrast (T1, T2, T1C, T2-FLAIR, PD) for analysis and comparison. As mentioned, a given patient often has numerous exams acquired over a long period of time, on different scanners, with different protocols (Figure 6.1A). In addition, exams often consist of images from different anatomical regions. Neurological exams may contain both supratentorial brain as well as spinal cord images. Patients may also have images acquired for reasons unrelated to MS (e.g. breast cancer, etc.). DICOM defines a “body part examined” tag, though this may contain the term brain for both brain and spine images that are part of the same exam, and in other cases is not filled in. This is therefore

not a reliable indicator of imaged anatomy and further confounds the preliminary analysis step of identifying the relevant subset of data for analysis (Figure 6.1A).

There exist only a few prior studies that focus on automatically classifying MR series acquisition contrast. Two past studies have used convolutional neural network architectures to classify MR images based on the imaging pixel data alone. In Ranjbar et al.[134], brain tumor MR images are classified into four contrasts (T1, T1C, T2, T2-FLAIR) with high accuracy (99.2%). However, this study was limited to a specific cohort of brain tumor MR scans acquired using research protocols. Another study [135] achieves similar results (~99%) on even more contrast types (adding proton density (PD), and magnetic transfer ON and OFF) and deploys the algorithm in a real clinical environment. Though the description of the dataset includes over 100 institutions with over 45,000 MR series, they mention only that they were acquired through clinical trials and do not describe the pathology or heterogeneity in scanner acquisitions. Clinical trials are often more uniform in their scanning acquisition parameters across institutions, and they do not necessarily reflect performance in heterogeneous data. In addition, this study uses 30 axial images and a second neural network in order to classify a single MR series into its contrast, limiting the applicability of this algorithm to volumes containing over 30 slices and adding a computationally expensive preprocessing step of resampling the volume to the axial direction. Finally, though both studies achieve great results, neither address how to classify brain MR series that are acquired with other contrast mechanisms, which is necessary for real-world algorithm deployment.

Another recent study achieves impressive results from MRI-associated DICOM metadata alone[136], e.g. echo time, repetition time, whether gadolinium was used. This approach is appealing because inference time is less computationally expensive compared with methods based

on pixel-level imaging data. Despite training and validating on large datasets from different institutions with impressive results, the authors of this study define the ground truth contrast label from the DICOM “series description” attribute. However, the authors acknowledge the limitation of this approach, noting that for up to 10% of series the contrast mechanism can not be accurately identified by the series description alone.

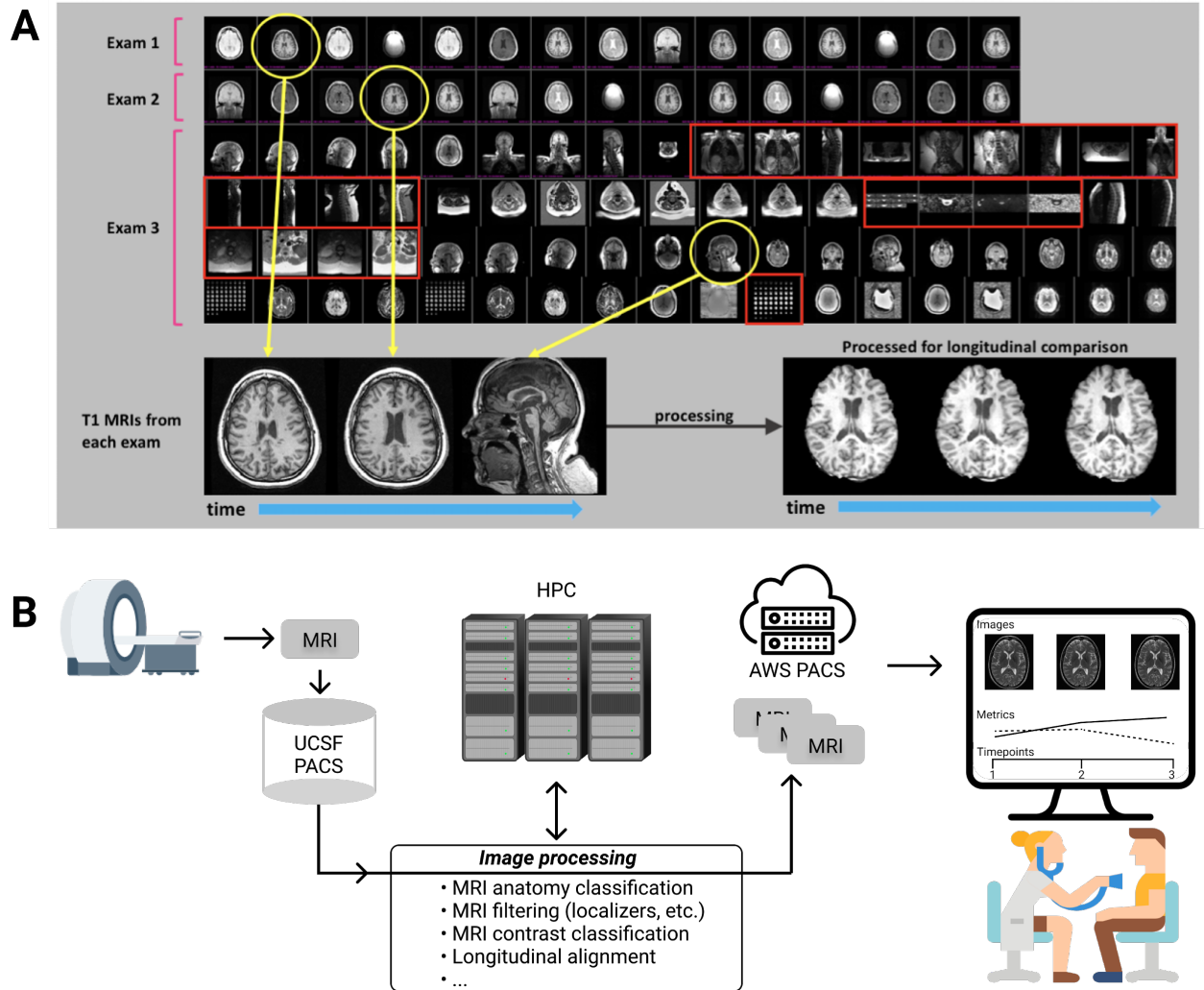
Our efforts to classify the data from our institutional PACS using series descriptions alone were even less successful, and in fact are what originally inspired us to use a pixel-based algorithm to classify contrast. For context, the frequency of unique series descriptions within a clinical trial dataset to a clinical PACS dataset was compared: despite comprising 60% fewer scans, there was an 18-fold increase in the number of unique series descriptions in the PACS data. Therefore, the difference in successful labeling might be due to an increased percentage of well annotated clinical trial data within the dataset used in Gauriau et al.[136].

The present study describes work aimed at developing and validating methods to automatically classify brain MR images acquired with specific contrast types. Specifically, the main objective is to obtain high accuracy for arbitrary real-world MR exams of MS patients sampled from the UCSF clinical PACS. These data are not subject to strict protocols of clinical trials and therefore much more challenging to automatically classify. This distinction is crucial, as this scenario describes the real clinical setting in which models and analysis pipelines are deployed. A secondary objective of this study is to demonstrate the feasibility of deploying the algorithm to other datasets with vastly different pathological profiles by testing the classification algorithm on brain tumor patients. Optimal contrast classification models were identified based on a comparison of the prediction accuracy of 1) a metadata-only, rule-based approach; 2) a

metadata-only machine learning model; 3) an imaging-only convolutional neural network; and 4) a combined ensemble model that uses both metadata and imaging data.

This work overcomes some of the limitations in prior work through 1) using just a single slice from each patient that can be in any direction: coronal, sagittal, or axial; and 2) ground truth classification was determined by a process of automatic rule-based derivation of weak labels, followed by visual review and correction by multiple reviewers; 3) inclusion of an additional “OTHER” category that does not force the algorithm to return the fixed set of contrast types. This study hypothesized that the combined model that uses metadata and imaging data together will obtain the highest classification accuracy. The results and tradeoffs of each model are discussed.





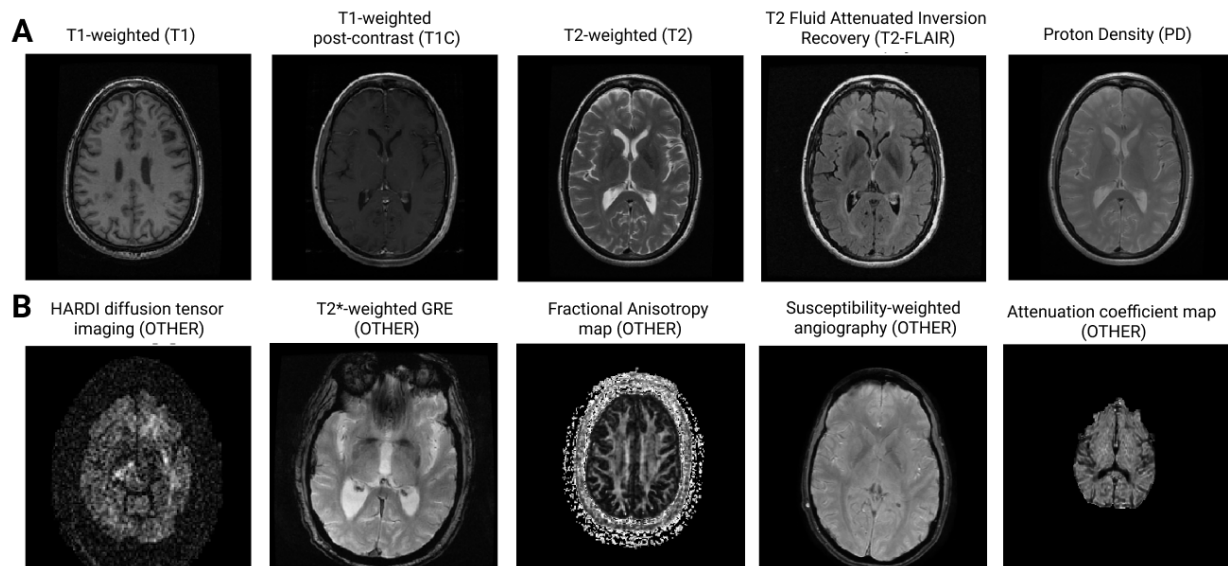
**Figure 6.1 The application of contrast classification of MRI in clinical practice.**

(A-top): Representative longitudinal exams from a single patient retrieved from the clinical PACS. Exams 1-3 consist of different numbers and types of acquisition sequences and even different anatomical regions. Acquisitions indicated in red are spine images despite being labelled as brain in the DICOM headers. Typical downstream applications require identification of input images acquired with specific tissue contrast. Images circled in yellow represent T1 weighted images from each exam used as input to a downstream application. (B-bottom): Representative downstream application to align longitudinal T2-weighted images for visual review.

## 6.2 Materials and methods

### 6.2.1 MR Sequences

Magnetic resonance imaging is incredibly flexible, offering a wide range of physical parameters to image and instrumental parameters to set for control of image content [34]. The specifications of the instrument define an MR sequence, which controls the soft tissue contrast and is especially useful in the brain. The most common and clinically relevant neuroimaging MR contrast mechanisms are T1-weighted (T1), T1-weighted after administration of a gadolinium-based contrast agent (T1C), T2-weighted (T2), T2 Fluid Attenuated Inversion Recovery (T2-FLAIR), and proton density (PD) (Figure 6.2A). In our data, there are many additional kinds of MR image contrasts acquired, including diffusion-weighted and T2\* weighted images. In order to correctly classify these images, we use a catch-all category of other contrasts called “OTHER”. Some of the images classified as OTHER are demonstrated in Figure 6.2B.



**Figure 6.2. Examples of images in each category.**

(A) T1, T1C, T2, T2-FLAIR, and PD are the most common image contrasts acquired during neuroimaging exams. (B) Category “OTHER” is comprised of many kinds of MR sequences, including T2\* weighted images that can look similar to T2 or PD.

## 6.2.2 MR exams and cohorts

Four MRI datasets were used for this analysis: 1) a multiple sclerosis research (MSR) dataset consisting of 1731 MS exams (<https://epicstudy.ucsf.edu/>); 2) a glioma research (GR) dataset consisting of 179 newly diagnosed and recurrent glioma exams; 3) a post-traumatic stress disorder research dataset (ADNIR) consisting of 116 exams from the publicly available ADNIDOD dataset; and 4) a MS clinical (MSC) dataset consisting of 311 exams representative of typical, poorly labeled institutional PACS data acquired at multiple external sites. Datasets MSR, GR, and ADNIR, were obtained with well-defined research acquisition protocol compared with MSC. GR has much more extensive pathology per exam compared with MSR, MSC, or ADNI. For all four datasets, each exam consists of multiple MR series that must be individually labeled based on the contrast mechanism.

**Table 6.1. Distribution of MR exams and series used for this study.**

<b>Dataset</b>	<b>Dataset type</b>	<b>No. of exams</b>	<b>Total series</b>	<b>T1</b>	<b>T1C</b>	<b>T2</b>	<b>T2 FLAIR</b>	<b>PD</b>	<b>OTHER</b>
<b>MS Research (MSR)</b>	Well annotated	1731	11106	3562	1679	1332	887	1392	2254
<b>MS Clinical (MSC)</b>	Poorly annotated	311	3244	655	722	384	593	75	815
<b>Glioma (GR)</b>	Well Annotated	179	607	147	170	124	163	0	3
<b>PTSD (ADNI)</b>	Well annotated	116	477	101	0	117	125	0	134

## 6.2.3. Labels

Each MR series in the MSR and MSC datasets were assigned preliminary weak labels using a rule-based model into the following categories: T1 weighted (T1), T1 post-contrast weighted (T1C), T2 weighted (T2), proton density (PD), T2 FLAIR (T2\_FLAIR) and OTHER. The image volumes from both MSC and MSR were subsequently visually reviewed for accuracy

by two brain imaging scientists (J.C.C. and J.G.C.). Though the GR and ADNIR datasets were obtained from prior studies and already accurately labelled, many MR series from these datasets were visually reviewed to ensure correctness (J.G.C.).

The number of slices in the volume together with the DICOM header SeriesDescription attribute were used to eliminate MR localizer and asset calibration volumes from the OTHER category. This step was applied because a) these volumes were 100% identifiable using these two features alone; and b) localizers are often acquired with an MR contrast of interest (T1, T1C, T2, T2-FLAIR, or PD), but due to their low resolution are almost never the volume of interest to a radiologist or researcher. If included as OTHER, localizers confound the learning process of the imaging-based models; if included as the MR contrast mechanism that they were obtained with, then a post-processing technique of choosing the correct (full-resolution) series from the exam must be applied. For these reasons, we chose a pre-processing step. In turn, the final number of MR series of each contrast from each dataset are detailed in Table 6.1.

#### 6.2.4 Train, validation, and test splits

To answer whether a model could be built to label the heterogeneous dataset (MSC), the data was split such that 100% of MSR and ADNIR data were in the training. MSC was randomly split by exam into 33.3% training, 33.3% validation, and 33.3% test. GR was split 50% validation and 50% test to evaluate whether the developed models were robust enough to predict image contrast even when presented with MR images containing more extensive pathology that it had not previously been exposed to.

#### 6.2.4 Rule-based approach

MR image contrast is determined by multiple scanning parameters, for example the echo time (TE) and repetition time (TR), that are stored as attributes in the DICOM header. In addition, administered contrast agents are often indicated in the DICOM header. Other metadata fields such as the “series description” may explicitly define the acquisition contrast in some data sets. A rule-based model using metadata derived from DICOM attributes was developed to derive weak contrast classification labels from the MSR and MSC cohorts. The model was based on a priori knowledge of attribute values used to scan with specific image contrast weighting together with detection of specific keywords (e.g. “T1”) found in series descriptions[137–139]. The rule-based approach was developed in-house in Python using Pydicom[140] to extract DICOM header attributes. Overall and per-class accuracy were calculated; notably, there is no “training”, but training, validation, and test sets are separated to serve as a comparison point for the following models that require training.

#### 6.2.5 Metadata models

DICOM metadata attributes were extracted using the Pydicom python package (Mason et al. 2020) and are listed in Table 6.2. Missing string type attributes were replaced with “None” and empty numeric attributes were replaced with the mean of the feature in the training data. The majority of these attributes were numeric; those that were string-type features were hashed to numeric values using the sha256 algorithm in the native Python library hashlib. Support Vector Machine (SVM) and Random Forest (RF) models using these features were developed in the python package scikit-learn (LinearSVC, SVC, RandomForestClassifier)[124]. Five-fold cross-

validation on the training data was used to evaluate which algorithm was best suited for predicting MR contrast from DICOM metadata.

Once random forest (RF) was chosen, a randomized cross validation (RandomizedSearchCV in sklearn) was used to search for the optimal set of hyperparameters (n\_estimators, max\_features, max\_depth, min\_samples\_split, min\_samples\_leaf, bootstrap) using the training set only. An RF using all training data (instead of 4/5) was retrained using the optimal hyperparameters returned from the search. Impurity-based feature importance scores derived from the trained RF are biased toward high-cardinality features and because of this the permutation importance function in sklearn was used to calculate the relative feature importance for the training, validation and test sets. Briefly, permutation importance is defined as the decrease in a model score function when a single feature is removed [141]; this calculation was permuted five times. Finally, the algorithm was tested on the validation and test MSC and GR datasets. This algorithm is referred to as the “metadata only” algorithm.

## 6.2.6 Image processing and imaging model development

The LPS coordinate system three-directional ranges were each divided by two in order to calculate the center location of the DICOM volume. The original unprocessed DICOM slice passing through the center location was recorded as the center slice of the volume. During training, volumes were transformed with random horizontal and vertical flips ( $p = 0.5$ ), random affine rotations and translations, and slight alterations to brightness, contrast, saturation and hue and resized to 224x224. A pre-trained ResNet-50 convolutional neural network (CNN) architecture was chosen to initialize the model weights. Data were normalized using the ImageNet normalization means and standard deviations. The final layer of the ResNet-50 was replaced with a fully connected layer with 6 outputs representing the 6 contrast categories. A cosine differential learning rate with a maximum value of 0.0003 was used together with a

weight decay coefficient of 0.0001. The model was trained for 40 epochs total, but early stopping was employed such that the highest accuracy model on the validation model was saved. This model is referred to as the “imaging only” model.

From this trained CNN, each center MR image was sent through the network the final 6 logit outputs were saved before application of the softmax function. First, t-distributed stochastic neighbor embedding (t-SNE)[142] was performed on these final 6 features. Regions of the t-SNE clusters were visually investigated in order to assess 1) whether there were obvious visual differences among regions within the same cluster; and 2) whether the misclassified images had visual similarities to their neighbors. Next, these logit values were combined together with metadata and the best machine learning metadata model was refitted on the training data, resulting in a combined imaging-based and metadata-based machine learning model. This model is referred to as the “combined” model.

## 6.2 Results

### 6.2.1 Rule-based approach

The rule-based approach utilized combination of the acquisition parameters obtained from the following DICOM attributes to predict the acquisition contrast type: 1) EchoTime; 2) RepetitionTime; 3) InversionTime; 4) FlipAngle; 5) ScanningSequene; and 6) the presence of select key words in the series description (e.g. “FRFSE”, “T2”, etc.). On the MSC dataset, the rule-based approach achieved 59.8% on the validation data, and 60.8% on the test data. On the GR dataset, it achieved 53.2% and 51.8% on validation and test data, respectively. Notably, this approach was not programmed to distinguish pre- and post-contrast images. If programmed to perfectly distinguish pre- and post-contrast images from one another, it would have achieved 70.7%, 71.6%, 80.6%, 79.8% accuracy on MSC validation, MSC test, GR validation, and GR test sets, respectively.

### 6.2.2 Modeling results

The main objective was to obtain high accuracy on the MS clinical dataset (MSC). Table 6.3 presents the comparison among the metadata-only, imaging-only, and combined models using validation and test set accuracies as well as per-class accuracies for the test set. In summary, the models that included imaging features outperformed the metadata-only models in validation and test set accuracies (>97% compared with ~95%). The details of the performance of each model are provided below.

The final metadata-only RF was trained with the following parameters: `n_estimators = 450`, `min_samples_split = 2`, `min_samples_leaf = 4`, `max_features = sqrt`, `max_depth = 66`,

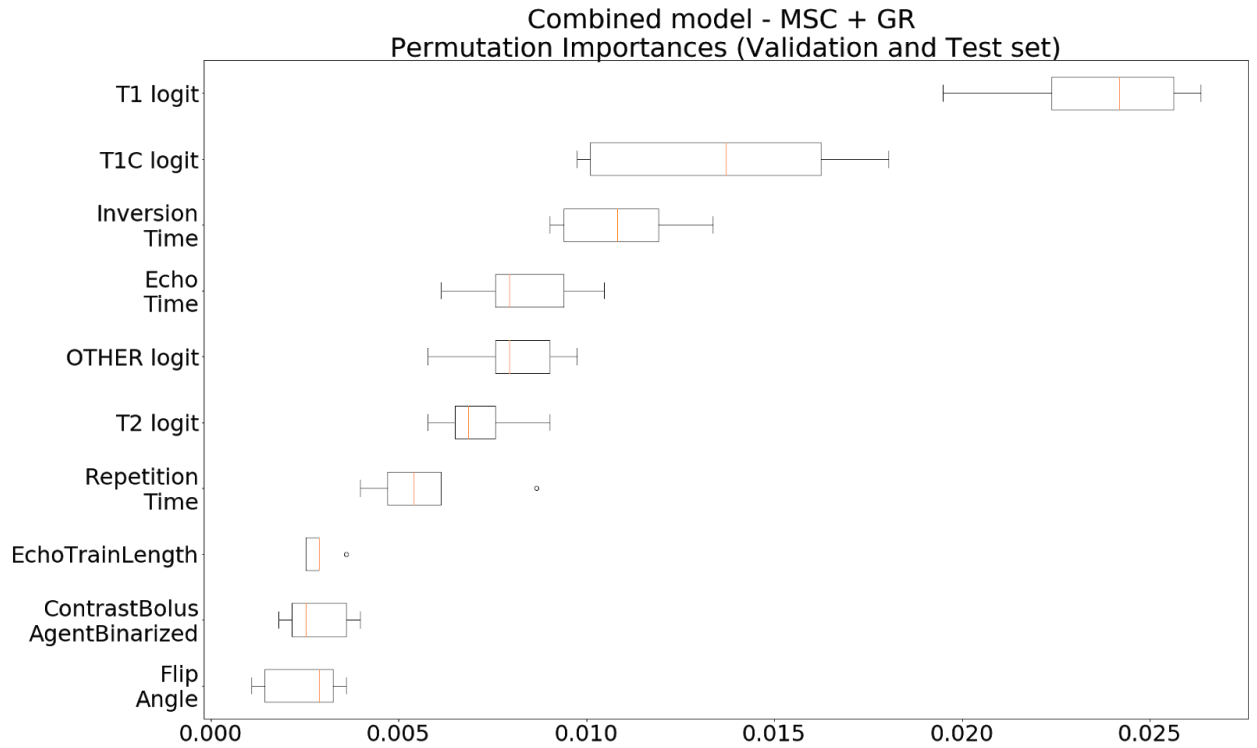
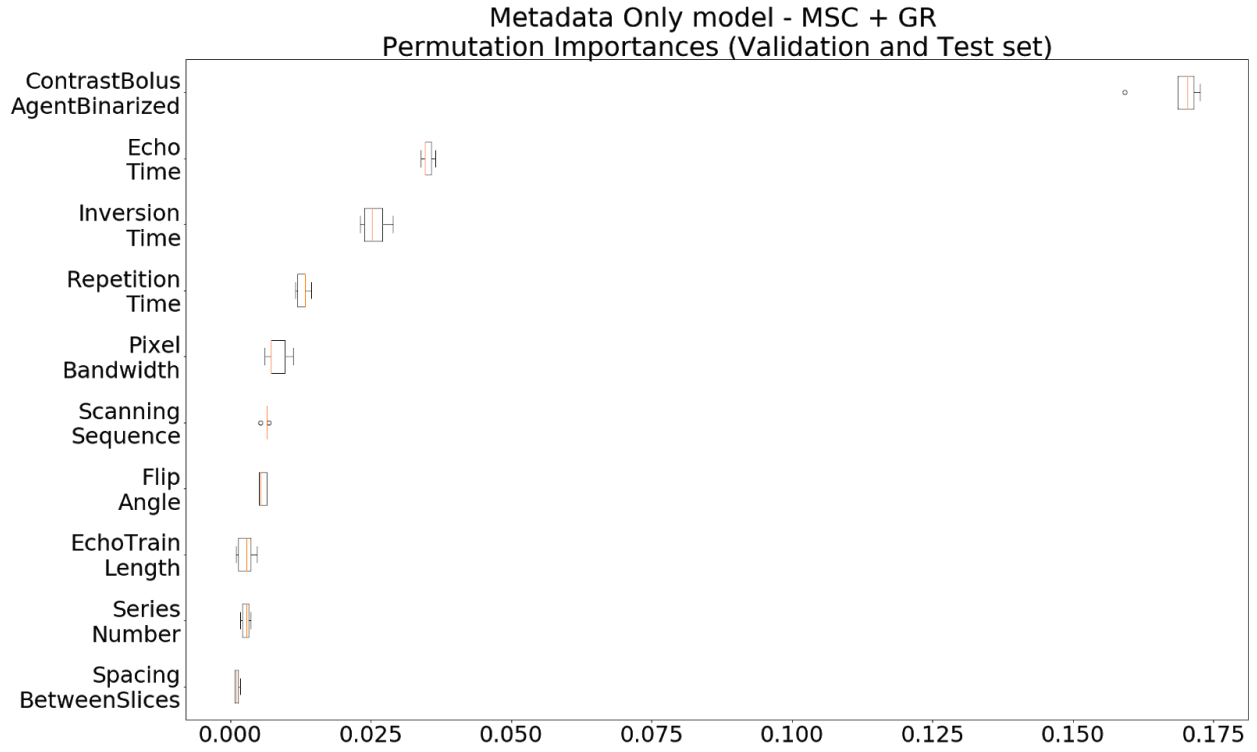


bootstrap = True. Permutation importance calculations are listed in Table 6.2 and depicted in Figure 6.3; the feature importance calculations suggest that removal of EchoTime decreases the accuracy approximately 10.6%, 4.1%, and 3.3%, for training, validation and test sets, respectively. Removing the ContrastBolusAgentBinarized decreased accuracy by 1.2%, 17.6%, and 16.2% for training, validation and test sets in comparison. InversionTime and RepetitionTime were increasingly important in the validation and test sets compared with the training set (Table 6.2, Figure 6.3). The final result on the MSC dataset achieved 99.7%, 94.0% and 95.4% on the training, validation and test set, respectively. The most common mistake made by the metadata-only RF on the MSC dataset was classifying images that were labeled as OTHER due to artifact as their original acquisition contrast (e.g. Gibbs ringing on T1C labeled as OTHER). The next most common mistake was classifying T1 as T1C and vice versa. Upon inspection of all misclassified MR series and their series descriptions, a few series were found to have the incorrect ground truth label (Table 6.5) but classified into their correct contrast by the classifier. Adjusting for the incorrect ground truth labels, the MSC validation and test set accuracies increase to 94.5% and 95.6%, respectively. Finally, the model was tested on the GR dataset. The metadata-only model performed better than those that included imaging on the GR dataset, classifying just 2 images incorrectly; one from each of the GR validation and test set (Table 6.5, Table 6.7).

The imaging-only CNN achieved higher accuracy for the MSC validation and test set compared to the metadata-only model. In total, it classified just 27 and 34 images incorrectly for a final MSC validation and test set accuracy of 97.4% and 96.9%. The most common mistake was misclassifying T1 for T1C and vice versa; this misclassification is explored further in the t-SNE analysis (Figure 6.4E). Misclassification analysis revealed that the CNN properly identified

7, 6, and 1 train, validation and test images that were associated with the incorrect ground truth label, a much higher number than either of the other two modeling types. When adjusting for the proper identification of these images, the imaging-only CNN achieved 98.0% and 97.1% on the validation and test MSC set. In comparison to the performance of the metadata-only GR dataset, the imaging-only CNN performed worse, classifying 10 MR series incorrectly, 5 each from the GR validation and test set. All of these mistakes were on sagittal T2-weighted images obtained with extremely high contrast, of which there were few similar training examples (Figure 6.4I).

The combined model was a trained random forest using the same features as the metadata-only model together with 6 additional features obtained from sending the MR series center slice through the trained CNN and saving the final 6 logit values in the model output. Each logit value is associated with one of the 6 final categories and is labeled accordingly. Compared with metadata-only and imaging-only models, it performed the best on the MSC dataset achieving 97.7% and 97.5% on validation and test data, respectively. When adjusting for the misclassified MSC data, the accuracy increased to 97.9% and 97.7% for the validation and test set, which was extremely comparable to the imaging-only model. Compared with the metadata-only RF, the validation and test feature importances weighted the T1 logit, T1C logit, T2 logit, and the OTHER logit much higher, and the feature importance was diminished for EchoTime and ContrastBolusAgentBinarized (Figure 6.3). Interestingly, the accuracy dropped significantly for the GR validation and test sets, 32 out of 34 misclassifications due to the same specific high-resolution, high-contrast 3D T2-weighted image classified as OTHER instead of T2-weighted.



**Figure 6.3. RF feature importance graphs on validation and test sets (MSC + GR together).**

**Table 6.2 Metadata-only random forest feature importances.**

Each feature importance represents the decrease in accuracy when the feature is permuted.

Metadata Only					
Rank	DICOM Tag	Name	Feature Importance - Train	Feature Importance - Valid	Feature Importance - Test
0	(0018,0010)	ContrastBolusAgentBinarized	0.01199	0.17513	0.16201
1	(0018,0081)	EchoTime	0.10565	0.04165	0.03331
2	(0018,0082)	InversionTime	0.00128	0.02531	0.02608
3	(0018,0080)	RepetitionTime	0.00074	0.01325	0.01304
4	(0018,0095)	PixelBandwidth	0.00060	0.00618	0.01049
5	(0020,0011)	SeriesNumber	0.00022	0.00088	0.00567
6	(0018,1314)	FlipAngle	0.00047	0.00530	0.00567
7	(0018,0020)	ScanningSequence	0.00043	0.00780	0.00454
8	(0018,0091)	EchoTrainLength	0.00014	0.00338	0.00241
9	(0018,0093)	PercentSampling	0.00005	-0.00044	0.00184
10	(0018,0022)	ScanOptions	0.00046	0.00162	0.00113
11	(0018,0088)	SpacingBetweenSlices	-0.00019	0.00074	0.00099
12	(0018,0089)	NumberOfPhaseEncodingSteps	0.00000	0.00000	0.00085
13	(None)	NumberOfFiles	-0.00005	0.00000	0.00071
14	(0018,0094)	PercentPhaseFieldOfView	0.00011	0.00044	0.00057
15	(0018,0025)	AngioFlag	0.00000	0.00000	0.00057
16	(0018,0083)	NumberOfAverages	0.00019	0.00000	0.00028
17	(0018,0050)	SliceThickness	-0.00002	-0.00029	0.00028
18	(0020,1002)	ImagesInAcquisition	0.00006	0.00088	0.00028
19	(0018,0086)	EchoNumbers	0.00002	0.00015	0.00014
20	(0018,0024)	SequenceName	0.00000	0.00074	0.00000
21	(0018,0087)	MagneticFieldStrength	0.00000	0.00044	0.00000
22	(0018,1310)	AcquisitionMatrix	0.00002	0.00000	0.00000
23	(0008,0016)	SOPClassUID	0.00002	0.00000	0.00000
24	(0018,1251)	TransmitCoilName	0.00000	0.00000	0.00000
25	(0018,0085)	ImagedNucleusQuantized	0.00000	0.00000	0.00000
26	(None)	NumberOfVolumes	-0.00002	0.00000	0.00000
27	(0018,0021)	SequenceVariant	0.00006	0.00059	0.00000
28	(0018,0023)	MRAcquisitionType	0.00003	-0.00029	0.00000
29	(0018,0015)	BodyPartExamined	0.00000	0.00000	-0.00014
30	(0028,0030)	PixelSpacing	0.00014	0.00000	-0.00043
31	(0018,1312)	InPlanePhaseEncodingDirection	-0.00003	0.00029	-0.00043
32	(None)	NumberOfImagePositions	0.00014	-0.00059	-0.00043
33	(0018,0084)	ImagingFrequency	0.00017	0.00074	-0.00128
34	(0028,0011)	Columns	0.00002	0.00074	-0.00213
35	(0028,0010)	Rows	0.00014	0.00088	-0.00227

**Table 6.3 Combined random forest feature importances.**

Metadata+Imaging Combined Model					
Rank	DICOM Tag	Name	Feature Importance - Train	Feature Importance - Valid	Feature Importance - Test
0	None	T1_logit	0.0020	0.0311	0.0214
1	None	T1C_logit	0.0003	0.0141	0.0137
2	(0018,0082)	InversionTime	0.0000	0.0088	0.0119
3	(0018,0081)	EchoTime	0.0003	0.0066	0.0089
4	None	OTHER_logit	0.0013	0.0091	0.0044
5	(0018,0080)	RepetitionTime	0.0002	0.0065	0.0044
6	(0018,0010)	ContrastBolusAgentBinarized	0.0002	0.0012	0.0044
7	None	T2_logit	0.0003	0.0121	0.0037
8	(0018,0095)	PixelBandwidth	0.0002	0.0012	0.0026
9	(0018,0091)	EchoTrainLength	0.0000	0.0031	0.0021
10	None	T2FLAIR_logit	0.0000	0.0022	0.0020
11	(0028,0010)	Rows	0.0000	0.0003	0.0020
12	None	PD_logit	0.0005	0.0010	0.0018
13	(0018,0020)	ScanningSequence	0.0000	0.0022	0.0018
14	(0018,1314)	FlipAngle	0.0000	0.0028	0.0016
15	(0018,1312)	InPlanePhaseEncodingDirection	0.0000	0.0007	0.0011
16	(0028,0011)	Columns	0.0001	0.0003	0.0006
17	(None)	NumberOfImagePositions	0.0000	0.0004	0.0004
18	(0018,0023)	MRAcquisitionType	0.0000	0.0001	0.0004
19	(0018,0021)	SequenceVariant	0.0001	0.0006	0.0003
20	(None)	NumberOfVolumes	0.0000	0.0000	0.0003
21	(0018,0094)	PercentPhaseFieldOfView	0.0000	0.0003	0.0001
22	(0018,0086)	EchoNumbers	0.0000	0.0000	0.0001
23	(0018,0050)	SliceThickness	0.0001	0.0003	0.0000
24	(0018,0015)	BodyPartExamined	0.0001	0.0000	0.0000
25	(0018,0089)	NumberOfPhaseEncodingSteps	0.0001	0.0000	0.0000
26	(0018,0024)	SequenceName	0.0001	0.0000	0.0000
27	(0020,0011)	SeriesNumber	0.0001	0.0000	0.0000
28	(0018,0084)	ImagingFrequency	0.0001	0.0000	0.0000
29	(0018,0087)	MagneticFieldStrength	0.0001	0.0000	0.0000
30	(0018,1310)	AcquisitionMatrix	0.0000	0.0000	0.0000
31	(0018,0025)	AngioFlag	0.0000	0.0000	0.0000
32	(0018,0085)	ImagedNucleusQuantized	0.0000	0.0000	0.0000
33	(0020,1002)	ImagesInAcquisition	0.0000	0.0000	0.0000
34	(0008,0016)	SOPClassUID	0.0000	0.0000	0.0000
35	(0018,1251)	TransmitCoilName	0.0000	0.0000	0.0000
36	(0018,0083)	NumberOfAverages	0.0001	-0.0006	0.0000
37	(None)	NumberOfFiles	0.0001	0.0001	0.0000
38	(0018,0022)	ScanOptions	0.0000	0.0003	-0.0001
39	(0018,0093)	PercentSampling	0.0001	-0.0006	-0.0001
40	(0018,0088)	SpacingBetweenSlices	0.0001	0.0003	-0.0004
41	(0028,0030)	PixelSpacing	0.0000	0.0001	-0.0006

**Table 6.4 Final model comparison for the MSC cohort.**

Cohort: MSC Data Only								
	<u>Overall Accuracy</u>		<u>Testing Class Accuracy</u>					
<u>Classifier</u>	<u>Validation</u>	<u>Test</u>	<u>T1</u>	<u>T1C</u>	<u>T2</u>	<u>T2 FLAIR</u>	<u>PD</u>	<u>OTHER</u>
<b>Heuristic</b>	59.81%	60.83%	57.80%	0.00%	34.71%	100.00%	72.00%	96.58%
<b>Metadata RF</b>	94.03%	95.40%	97.25%	95.53%	98.35%	100.00%	72.00%	91.44%
<b>Imaging CNN</b>	97.44%	96.93%	99.08%	95.53%	100.00%	95.63%	88.00%	96.92%
<b>Combined RF</b>	97.73%	97.47%	99.08%	96.75%	99.17%	99.51%	88.00%	95.55%

**Table 6.5 Final model comparison for the GR cohort.**

Cohort: Glioma (GR)								
	<u>Overall Accuracy</u>		<u>Testing Class Accuracy</u>					
<u>Classifier</u>	<u>Validation</u>	<u>Test</u>	<u>T1</u>	<u>T1C</u>	<u>T2</u>	<u>T2 FLAIR</u>	<u>PD</u>	<u>OTHER</u>
<b>Heuristic</b>	53.29%	51.82%	93.15%	0.00%	11.29%	100.00%	0.00%	100.00%
<b>Metadata RF</b>	99.67%	99.67%	100.00%	100.00%	100.00%	100.00%	0.00%	50.00%
<b>Imaging CNN</b>	98.36%	98.36%	100.00%	100.00%	92.19%	100.00%	0.00%	100.00%
<b>Combined RF</b>	94.74%	94.06%	100.00%	100.00%	70.97%	100.00%	0.00%	100.00%

**Table 6.6 Misclassification analysis of MSC cohort.**

Misclassification analysis: MSC							
		<i>Correct</i>	<i>Artifact</i>	<i>OTHER as original contrast</i>	<i>Bad Slice</i>	<i>Unknown reason</i>	<i>Total</i>
<b>Metadata RF</b>	<i>Train</i>	2	0	16	0	18	36
	<i>Valid</i>	5	0	12	0	46	63
	<i>Test</i>	2	0	17	0	32	51
<b>Imaging CNN</b>	<i>Train</i>	7	3	0	3	14	27
	<i>Valid</i>	6	1	0	10	10	27
	<i>Test</i>	1	5	0	5	23	34
<b>Combined Model</b>	<i>Train</i>	6	0	0	0	4	10
	<i>Valid</i>	2	3	0	7	12	24
	<i>Test</i>	2	7	0	4	15	28

**Table 6.7 Misclassification analysis of GR cohort.**

Misclassification analysis: GR							
		<i>Correct</i>	<i>Artifact</i>	<i>OTHER as original contrast</i>	<i>High-res, High-contrast, 3D T2 as OTHER</i>	<i>Unknown reason</i>	<i>Total</i>
<b>Metadata RF</b>	<i>Valid</i>	0	0	0	0	1	1
	<i>Test</i>	0	0	0	0	1	1
<b>Imaging CNN</b>	<i>Valid</i>	0	0	0	5	0	5
	<i>Test</i>	0	0	0	5	0	5
<b>Combined Model</b>	<i>Valid</i>	0	0	0	16	0	16
	<i>Test</i>	0	0	0	18	0	18

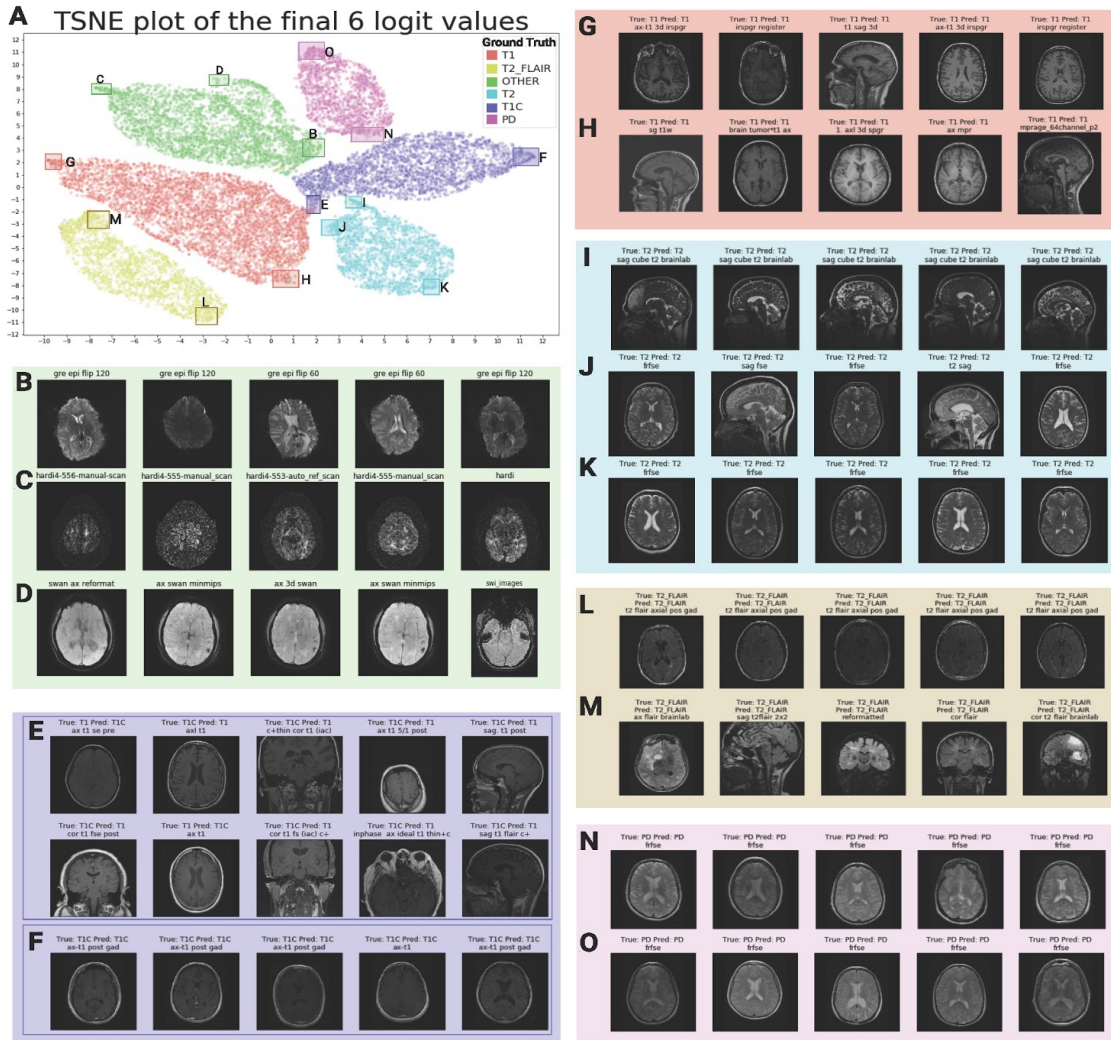
### 6.2.3 t-SNE results

Figure 6.4A depicts the results of the t-SNE analysis. In order to evaluate whether there were visual differences among regions within the same cluster, the coordinates of the t-SNE were used to delineate regions of interest and the images corresponding to those regions were visualized. Given that a catch-all category of OTHER was used to represent less-frequently acquired MR, it was of great interest to explore whether the different regions in the OTHER group comprised different MR series acquisition types. Figure 6.4B-D depicts samples of images corresponding to those regions. Region B is largely composed of axial gradient-echo echo-planar images that share severe distortion. HARDI diffusion tensor images make up Region C, while susceptibility-weighted images (SWI) make up Region D. An additional region proximate to Region D represented a cluster of T2\*-weighted images (not pictured).

In addition, maximally separated regions in each cluster were visualized. Regions G and H correspond to the cluster of T1-weighted images. Both regions contained a mixture of sagittal and axial images that had no obvious difference in gray-white matter contrast. Regions I, J, and K corresponded to the T2-weighted cluster. Region I was chosen due to its protrusion from the main cluster shape, and the high-contrast, 3D T2-weighted images that were classified incorrectly in the prior analyses were found in this region. Regions J and K were similar, but region K was more uniformly axial compared with region J that contained sagittal images as well. Due to the presence of non-yellow points in Region M of the T2-FLAIR cluster, Region M was visually compared with Region L, a set of points maximally separated from Region M within the same cluster. Region L was composed of highly-uniform axial T2-FLAIR images compared with Region M, which contained sagittal, axial, and coronal images - many with extensive pathology.

To complement the analysis of within-cluster differences, the overlapping region of the T1 and T1C clusters was examined to answer whether the misclassified images had visual similarities to their neighbors. Examples from Region E demonstrate that many of the T1 images predicted as T1C had low gray-white matter contrast, typical of T1C images. In addition, T1C images misclassified as T1 had very little contrast enhancement, and retained their gray-white matter contrast even in the post-contrast setting. Compared with Region E, Region F was much more uniform in appearance with similar contrast and axial orientation.





**Figure 6.4. t-SNE of the 6 logits derived from the final layer of the convolutional neural network.**

(A) t-Stochastic Neighbor Embedding of the logits output from the CNN. (B) Example of images and series descriptions that correspond with the Region B on the t-SNE plot, represented largely by gradient-echo echo-planar images. (C) HARDI diffusion volumes largely representing Region C on the t-SNE plot. (D) Susceptibility-weighted images corresponding to Region D on the t-SNE plot. (E) Misclassified T1 and TIC images in Region E that appear similar to the contrast of the other. (F) In contrast to Region E, Region F contains highly uniform axial TIC images. (G-H) images corresponding to Regions G and H respectively. Though maximally separated, these regions both contain axial and sagittal images that appear similar in contrast. (I) Region I contains high-contrast, high-resolution 3D T2-weighted images that have their own distinct area within the T2-weighted cluster. (J-K) Region J contains both axial and sagittal images with varied contrast compared with Region K which appears more uniform. (L) Examples of images located in Region L depicting uniform, axial T2-FLAIR images with little evidence of pathology. (M) Examples of images located in Region M of the yellow cluster, representing coronal, sagittal and axial T2-FLAIR images with extensive pathology. (N-O) This cluster of PD images does not appear different when comparing maximally separated Regions N and O.

## 6.3 Discussion

In this study, we develop multiple algorithms capable of predicting the contrast mechanism of diverse MR series with >97% accuracy. Importantly, we developed these tools to satisfy a real clinical objective within our institution. Our internal tool -- the *UCSF Multiple Sclerosis Bioscreen* -- is a web-based application that displays aligned longitudinal MR images acquired with similar contrast to clinicians alongside other pertinent biometrics (Figure 6.1B). The MR series contrast classifiers presented in this study represent an important step in the pipeline, after classifying the MR series by the anatomical region and before longitudinal image registration. By streamlining the agglomeration and presentation of relevant data, our tool increases the amount of time clinicians have to assess disease status. If the contrast of a particular MR series within an exam is misclassified or if a low-quality image is chosen for all other timepoints to be aligned to, valuable clinician and patient time might be wasted. Therefore, retrieving high-quality series of the correct contrast within an exam is the main priority of this step within our tool. It is with this lens that we can compare the performance of each kind of model developed in this study.

We compare the performance of 1) a rule-based classifier; 2) a metadata-only random forest classifier; 3) an imaging-only convolutional neural network; and 4) a combined model that uses both imaging data outputs from (3) and metadata in a random forest. The primary goal was to obtain high accuracy on the MSC dataset, a heterogeneous clinical cohort, while the secondary goal was to obtain good classification accuracy on the GR dataset, containing images with more extensive pathological burden. All three trained algorithms vastly improved the performance of the rule-based approach, with models including imaging (imaging-only and combined models) performing the best on the MSC validation and test sets.

The imaging-only and combined model were extremely comparable in their performance on the MSC dataset, with the imaging-only model achieving 98.0% validation and 97.0% test set accuracy, and the combined model achieving 97.9% validation and 97.7% test set accuracy, respectively (Table 6.3). The combined model improved the classification of T2-FLAIR images in the test set and decreased the number of T1 pre- and post-contrast mistakes made by the algorithm. The imaging-only CNN improved upon the classification of the OTHER category compared with the combined model for the MSC dataset. With the lens of delivering the image contrast of interest of this MSC dataset, the combined model outperforms the imaging-only model.

We dive deeper into the details of the imaging-only CNN by conducting a tSNE analysis (Figure 6.4). Our between-cluster t-SNE analysis of the T1 and T1C images in Region E in Figure 6.3 confirmed that the CNN misclassified T1 and T1C images that were visually similar to one another. The canonical features that visually differentiate T1C from T1 images are regions of bright contrast enhancement and lower gray-white matter contrast. We observe that the T1C images that are classified as T1 by the CNN are those that retain their gray-white matter contrast and show little to no enhancement.

Our within-cluster t-SNE analysis of the OTHER images (Figure 6.4B-D) provided a valuable qualitative addendum to our main findings. From this analysis, we infer that the logit values output from the imaging-only CNN model contained a rich diversity of imaging features that were able to separate different kinds of images without being explicitly trained to. When we formulated the experimental design for this project, we did not originally label images explicitly as HARDI, SWI, or GRE EPI due to lower sample size and lower priority in the context of the MS Bioscreen application. However, the ability of the imaging-only CNN to separate these

image types suggests that future efforts to refine our MR series acquisition labeling to more specific contrast mechanisms will be successful.

In addition, we demonstrate the feasibility of generalizing our algorithms to a research glioma dataset. This goal was specifically relevant to the future direction of the Bioscreen tool, as the extension of the tool to the Neurosurgery department has been initiated. The best performing algorithm was the metadata model achieving 99.7% on both validation and test GR datasets, classifying just one image wrong in the GR validation and test sets each. We surmise that this is likely due to the homogeneity of the metadata in the strict research protocol that guided the acquisition of GR data within our institution. Contrary to expectations, the combined model had lower accuracy on the GR dataset due to the classifying a set of high-resolution 3D T2-weighted images acquired with BrainLab protocol as OTHER (Figure 6.4I). The training data included no T2-weighted images that resembled these and the RepetitionTime was much longer on average for higher T2-weighting compared with those in the training dataset (2912.5 ms vs. 2370.9 ms). In addition, the imaging-only CNN segregates these images into their own small, separated section within the T2-weighted cluster. Taken together, the information stored in the imaging-based logits and TR difference contained in the metadata and the lack of similar acquisitions in the training data were the likely causes that created the conditions in the combined RF to segregate these specific images from the rest of the T2-weighted images. In order to advance the algorithm to deployment in the context of real-world glioma MR series, all three algorithms should be tested on clinical glioma data stored in institutional PACS systems, where we expect imaging-based algorithms outperform metadata-only models, similar to the results on the MSC validation and test sets. In addition, to achieve even greater results, we will

gather additional glioma datasets and include some of these high-resolution, T2-weighted 3D images in MR series during training.

Though the combined model slightly outperformed the imaging-only CNN on the MSC dataset which was the primary purpose of this investigation, there remain advantages of using the imaging-only model. First, compared with the combined model, it performs nearly just well on the MSC data while increasing the accuracy on the GR dataset, due to generalizing better to the high-contrast, high-resolution 3D T2-weighted images that it hadn't seen before. Second, the number of misclassified images in the training data that were in fact labeled incorrectly was increased for the imaging-only model compared with the others, indicating it was more robust to overfitting to the training set. Thirdly, compared with the metadata-only model, the imaging-only CNN can identify series impacted by severe artifacts as OTHER instead of the contrast that it was acquired with, which is beneficial in our pipeline. Finally, the only preprocessing step is calculating the center of the image volume to identify the slice most representative of the image contrast, which is beneficial during model deployment. Though using imaging data is more computationally expensive compared with metadata alone, the inference time difference is nominal in the context of the *UCSF Bioscreen* application. Though metadata-only models shorten inference time at scale, they have lower overall accuracy and will never have the ability to filter out MR sequences with artifacts, ensuring increased downstream issues with alignment and display. Therefore, in the context of delivering high-quality images to the *UCSF Bioscreen* application, the imaging-only CNN model would be the most appropriate as the contrast classification step.

In this study, we investigate different modeling paradigms for classifying MR series based on their acquisition contrast. We achieve classification accuracy acceptable to deploy

within the pipeline detailed in Figure 6.1B for the MS clinical data stored in PACS, as well as on unseen glioma images with more extensive pathology. In addition, our t-SNE analysis indicates that the imaging-based CNN has more refined discriminatory power for ADC maps, HARDI, and T2\* weighted images, imaging series commonly acquired during MS and glioma exams. Taken together, our immediate future directions include extending our model to more refined categories and including more kinds of MR acquisitions (e.g. high-resolution, high-contrast 3D T2-weighted images) in our training data to improve our accuracy on clinical data even further. Overall, we recommend an imaging-based model within a pipeline such as ours.

## 7. Distinguishing recurrent tumor from treatment-induced effects

As introduced in Section 2.2.3, recurrent glioma and the effects of treatment can appear identical on MRI, resulting in unnecessary surgical intervention and confounding the results of clinical trials. When I heard about this problem, it became the central question I was interested in trying to solve throughout my PhD. This problem is top of mind for many practicing neuro-radiologists and neuro-oncologists. During my time in Dr. Lupo's lab, I have taken two approaches. The first approach was useful to overcome the heterogeneity within individual patients experiencing a mixture of treatment-induced damage and real tumor growth (residual or recurrent). I was lucky to have access to a unique dataset of image-guided tissue samples that allowed pairing of structural, physiologic, and metabolic MR imaging to the pathological outcome of individual tissue samples. Using this approach, we discovered that spectroscopic MR in the non-enhancing region of the lesion was the most sensitive and specific marker for distinguishing the two phenomena.

When my interest in deep learning in imaging was burgeoning, my colleague Dr. Paula Alcaide-Leon discovered that centrally restricted diffusion (inside of the necrotic region of a ring-enhancing lesion) was predictive of treatment-related changes. I hypothesized that a convolutional neural network might be able to harness this signal, together with other texture-level features to predict treatment-related changes. The work using this approach is adapted from a submission to the International Society for Magnetic Resonance in Medicine conference that goal is presented in Section 7.2.

## 7.1 Recurrent tumor and treatment-induced effects have different MR signatures in contrast-enhancing and non-enhancing lesions of high-grade gliomas

### 7.1.1 Introduction

Tumor recurrence in patients with high-grade glioma (HGG) is difficult to diagnose because treatment-induced injury often appears identical on conventional anatomic magnetic resonance (MR) imaging. It is estimated that 25 to 35 percent of patients who undergo standard-of-care radiation and chemotherapy in the form temozolomide for HGG experience treatment-related injury, and its appearance is even more common with the recent advent of immuno- and other targeted therapies in clinical trials [32,143–145]. If recurrence is incorrectly diagnosed, a patient may be removed from an effective therapy, which could invalidate the results of a clinical trial or expose a patient to unnecessary surgical intervention. To mitigate these risks, identifying the exact location and extent of treatment-related changes within newly enlarging lesions is critical.

Despite the known pitfalls of using anatomical imaging, the current Response Assessment in Neuro-Oncology (RANO) criteria for HGG relies solely on standard T1- and T2-weighted MR imaging [146]. These techniques allow for visualization of anatomical abnormalities, but are limited in their ability to capture the underlying biology that differentiates true recurrent glioma from treatment effects. Emerging data suggest that incorporating more advanced MRI techniques may be useful for probing underlying biological differences: diffusion-weighted imaging for capturing the restricted water movement from the density of proliferating tumor cells [51,56,57,147–149]; perfusion-weighted imaging for evaluating the increased vasculature recruited to support a growing mass [55,63,76,150–152]; and



spectroscopic imaging for elucidating the metabolic differences between inflammation and proliferating tumor cells [56,76,77]. Although several groups have investigated the potential for distinguishing between treatment effects and recurrent tumor using these techniques, the majority of prior studies typically involve calculating the mean diffusion-, perfusion-, or spectroscopic-derived parameter value from an anatomical region of interest (ROI) and normalizing that value against that obtained from contralateral normal-appearing white matter in order to obtain a threshold that can distinguish treatment effects from true tumor recurrence. These ROI-based methods suffer from widely varying cutoff values for parameters due to inter-observer dependence, intratumoral heterogeneity, and the coexistence of treatment effect and tumor within the same lesion. Such studies also typically use radiographic observations or a single tissue sample for outcome determination, ascribing a single diagnosis to a mixture of tissue types that could mask the heterogeneity of the lesion and limit the accuracy of the cutoff value concluded from the study and overall clinical diagnosis.

To overcome the complications introduced by tissue heterogeneity in ROI-based studies, one strategy is to use image-guided tissue samples of known coordinates to directly map MRI characteristics to histopathology. In 2002, Rock et al. pioneered this technique in distinguishing radiation necrosis from recurrent disease using metabolite ratios derived from <sup>1</sup>H MR Spectroscopic Imaging (MRSI) at the location of sampled tissue [153]. Their findings suggested that the ratios of choline and lactate/lipid to creatine could differentiate samples with pure necrosis and tumor, but not those with mixed pathology. In 2009, Hu et al. utilized this technique in conjunction with Dynamic Susceptibility Contrast (DSC) perfusion-weighted imaging to distinguish post-treatment radiation effect from recurrent tumor with high sensitivity and specificity using relative cerebral blood volume (rCBV) values from 13 patients.

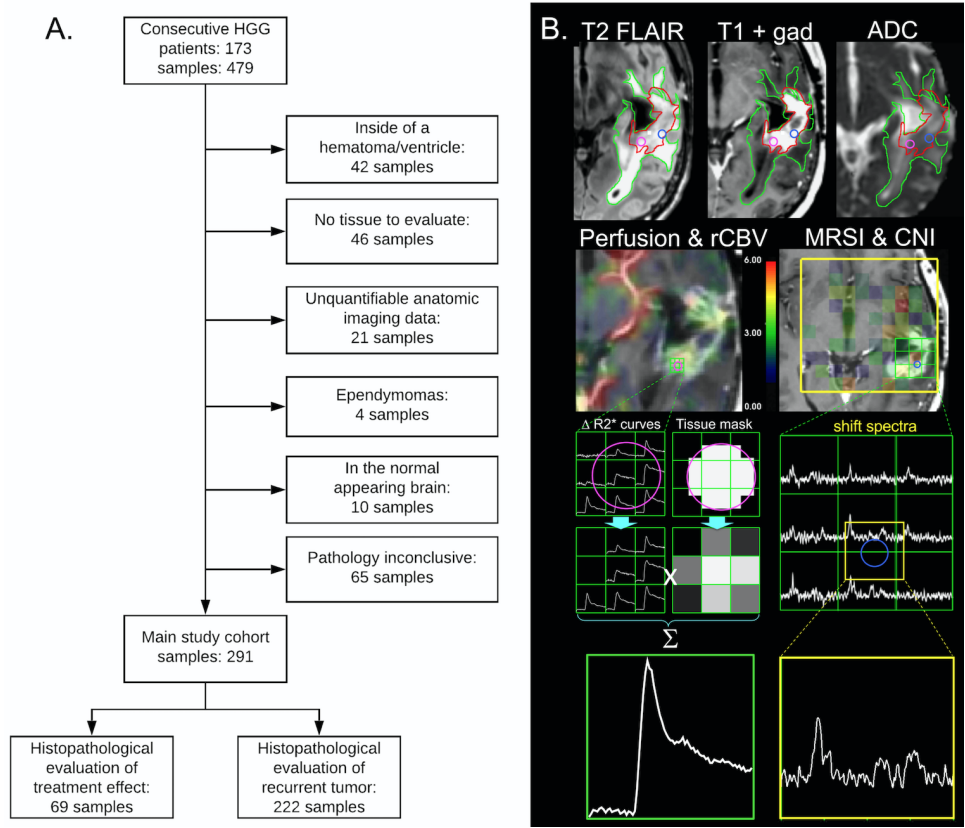
The goal of this study was to determine whether different MR characteristics were relevant for distinguishing pathological features of recurrent tumor from the effects of treatment in the contrast enhancing and non-enhancing lesions of recurrent high-grade gliomas by leveraging a unique dataset of image-guided tissue samples of known coordinates to avoid complications of tissue heterogeneity that confounds most lesion-level analyses. Based on prior literature, we expect that samples composed of recurrent tumor will have increased blood volume and abnormal metabolism, with decreased diffusion compared to samples containing treatment-effect. We also hypothesize that: 1) this difference would be more pronounced in diffusion and perfusion metrics for samples within the contrast enhancing lesion (CEL), while metabolic measures would be equally effective at differentiating recurrent tumor from treatment effect in both the contrast enhancing and non-enhancing lesions (NEL); and 2) the addition of multiparametric physiologic and metabolic MRI in conjunction with tissue sample level analyses will provide increased sensitivity and specificity in distinguishing recurrent tumor from the effects of treatment in both types of lesions compared to anatomical imaging.

## 7.1.2 Methods

### 7.1.2.1 Patient recruitment

Institutional Review Board approval was obtained at our institution to prospectively enroll consecutive patients with an initial pathological diagnosis of a WHO grade III or IV glioma that were suspected of recurrence between 2007 and 2017. A total of 173 patients (median age, 52; range, 21-84) that provided written informed consent to participate underwent MR imaging between 1 and 3 days prior to surgical resection and tissue sample collection were enrolled. Demographics were typical for this population as shown in Table 7.1. The majority of patients received an initial surgical resection followed by the standard of care RT (119 patients) plus temozolomide (114 patients). 41 patients also received 1 or more of 19 different additional therapies (18 bevacizumab; 8 immunotherapy). Prior treatment history was unknown for 12 patients.

From each patient, 1-4 tissue samples were obtained (mean samples/patient was 2.3, with 1.18 samples/patient within the CEL and 0.91 samples/patient within the NEL). Of the initial 479 samples from 173 patients, a sample was only included in the analysis if it: 1) had a conclusive histopathological outcome; 2) came from a patient whose initial diagnosis was as high-grade glioma; 3) did not come from a region of hematoma or extensive necrosis; 4) had quantifiable anatomic imaging data; and 5) was located within either the contrast-enhancing or non-enhancing lesion (Figure 7.1A). This resulted in a total of 291 samples from 139 patients, 26 of which had a diagnosis at surgery of grade III, 90 as grade IV, and 23 were exclusively treatment effect.



**Figure 7.1 Overview of tissue samples.**

(A) Inclusion and exclusion criteria for tissue samples. (B) Examples of MR modalities and parameters used per tissue sample.

**Table 7.1. Clinical demographics of the patient population.**

Clinical/Demographic Characteristics		Patients	Percentage of patients
<b>Totals</b>		139	100
<b>Gender</b>	Female	56	40.3
<b>Race</b>	White	112	80.6
	American Indian	1	0.7
	Asian	6	4.3
	Pacific Islander	2	1.4
	Other	18	12.9
<b>Clinical Diagnosis</b>	Grade III Astrocytoma	11	7.2
	Grade III Oligodendroglioma	15	10.8
	Grade IV Glioblastoma	87	62.6
	Grade IV Gliosarcoma	3	2.2
	Treatment Effect	23	16.5
<b>Age at recurrent surgery</b>	median, range	53	21-84

### 7.1.2.2 MR acquisition

MR examinations were performed on a 3T scanner (GE Healthcare Technologies) using an eight-channel phased-array head coil. Standard anatomical imaging included T2-weighted FLAIR and fast spin echo images, along with 3D T1-weighted IR-SPGR imaging pre- and post- the injection of a gadolinium-based contrast agent. Diffusion-tensor images (DTI) were obtained in the axial plane with  $b=1000$  s/mm<sup>2</sup> and either 6 gradient directions and 4 excitations or 24 gradient directions and 1 excitation or  $b=2000$  s/mm<sup>2</sup> and 55 gradient directions [repetition time (TR)/echo time (TE) = 1000/108 milliseconds, voxel size = 1.7-2.0 × 1.7-2.0 × 2.0-3.0 mm]. DSC perfusion-weighted images were obtained following a 3-ml/s bolus injection of 0.1 mmol/kg body weight gadolinium diethyltri-amine pentaacetic acid using a series of T2\*-weighted echo-planar images [TR/TE/flip angle = 1250-1500/35-54 milliseconds/30°-35°, 128 × 128 matrix, slice thickness = 3-5 mm, 7-24 slices with 60-80 time points] before, during, and after the arrival of the contrast agent bolus. The temporal resolution was between 1 and 1.5 seconds, with total acquisition time ranging from 1-2 min.

The 3D 1H MRSI were acquired using point-resolved spectroscopic selection for volume localization and very selective saturation pulses for lipid signal suppression [excited volume = 80 × 80 × 40 mm, TR = 1100-1250 ms, TE = 144 milliseconds, overpress factor = 1.5 if lactate edited, otherwise 1.2, field of view = 16×16×16 or 18×18×16 cm, nominal voxel size=1×1×1cm], flyback echo-planar readout gradient in the SI direction, 988 Hz sweep width and 712 dwell points. A dual-cycle lactate-edited sequence<sup>18</sup> was used for 42 patients (83 samples; 11 min), while a standard single-cycle sequence [154] was used for the remaining 38 patients (68 samples; 6 min).

### 7.1.2.3 MR data processing

Anatomic, diffusion, and perfusion data were aligned to the T1 post-contrast image using either FMRIB's FSL Linear Image Registration Tool (FLIRT) [116,117] or Slicer's BRAINSFit tool with B-spline warping [118]. Spherical 5-mm-diameter ROIs were generated at the location of the spatial coordinates recorded during surgery in order to balance the potential error introduced by tissue shift and the need to restrict the ROI to immediate vicinity of the sampled tissue [155]. All locations were then visually verified for accuracy on anatomical imaging using screenshots taken during the surgery, and excluded if there was a mismatch between the coordinate locations and visualized location on imaging.

A pipeline that utilized components of FMRIB's Diffusion Toolkit was applied to estimate relevant diffusion parameters from the DWI and DTI data as previously described [116]. In order to account for differences in acquisition parameters over the 10 year study duration, voxel values for the Apparent Diffusion Coefficient (ADC) and fractional anisotropy (FA) maps were normalized to the mode of intensities in normal-appearing brain tissue (resulting in nADC and nFA maps). nADC increases as average diffusivity of water within a voxel increases and therefore its decrease should act as a marker for glial proliferation while nFA is an index for the amount of directional movement of water resulting from the parallel orientation of axonal fibers in white matter.

From the DSC perfusion data, rCBV maps were first calculated on a voxel-by-voxel basis utilizing a modified gamma-variate function that takes into account leakage of the contrast agent [156]. To generate a single concentration-time curve per sample, unquantifiable voxels of noise were automatically excluded and the percentage of the tissue sample ROI within each perfusion voxel was determined before taking a weighted average of the remaining dynamic curves based

on their percentage overlap with the ROI (Figure 7.1B) [157]. This method helped mitigate inaccuracies in quantification of perfusion parameters from tissue sample ROIs due to the relatively small size of the ROI compared to the low resolution of the perfusion scan and the presence of susceptibility artifacts and necrosis. rCBV for each tissue sample was calculated as the area under the final gamma-variate-fitted single concentration-time curve after leakage correction. An increase in rCBV reflects an increased volume of blood vessels in a given amount of brain tissue, and is therefore expected to be elevated in regions of recurrent tumor as it recruits blood vessels to supply oxygen to an enlarging mass.

Spectroscopic data were reconstructed and postprocessed using in-house software, as previously described [158–160]. To generate a single spectrum centered at the location of each tissue sample, 3D spectral arrays were first shifted in k-space to reconstruct a spectral voxel on the center coordinates of each tissue sample location (Figure 7.1B). Peak heights and areas were determined from baseline-subtracted, frequency- and phase-corrected spectra on a voxel-by-voxel basis [160]. The choline-to-NAA index (CNI) and the choline-to-creatine index (CCRI) were obtained as previously described, using the entire 3D array of spectra in the iterative regression [161]. Normalized total choline (nCho), creatine (nCre), and N-acetylaspartate (nNAA) intensities were calculated using their median value in voxels that had been identified during the CNI calculation as being from the normal brain. nCho increases with increased cellular turnover, nNAA is a neuronal marker, and nCre is thought to be relatively constant regardless of tissue makeup. Using indices such as CNI and CCRI are appealing because they can capture more complex information, e.g. areas of high cellular turnover with low neuron density (CNI), which can be indicative of tumor tissue.

Together, these post-processing steps resulted in 8 different MR imaging parameters: nADC and nFA from DTI; rCBV from DSC perfusion; and CNI, CCRI, nCho, nCre, and nNAA from MRSI.

#### 7.1.2.4 Tissue sampling and histopathological assessment

A minimum of 4 tissue samples at least 1 cm apart were preoperatively planned to maximize heterogeneity within the hyperintense region on the T2 FLAIR image. Edges of cavities or necrotic regions were avoided, and the accessibility of the tissue target to the surgeon was considered during planning. During surgery, an intraoperative navigation system (BrainLab or Surgical Stealth) guided the neurosurgeon to the targeted locations and was used to record target coordinates for excised tissue samples. Samples were immediately formalin fixed and paraffin embedded [162].

Hematoxylin and eosin stained slides from tissue samples were evaluated by a board-certified pathologist (J.J.P.). Slides were assessed for the presence of tumor cells, necrosis, and treatment-related abnormal vasculature. Samples that had both signs of treatment-related changes and zero tumor cells were considered treatment effect. The presence of tumor cells was scored based upon review of H&E-stained sections by a neuropathologist as 0 = no tumor present, 1 = infiltrating tumor with rare cells, 2 = infiltrating cellular tumor, and 3 = highly cellular infiltrating tumor involving >75% of the tissue. Only samples with more than rare infiltrating tumor cells (tumor scores of 2 or greater) were considered recurrent tumor.



#### 7.1.2.5 Statistical analysis

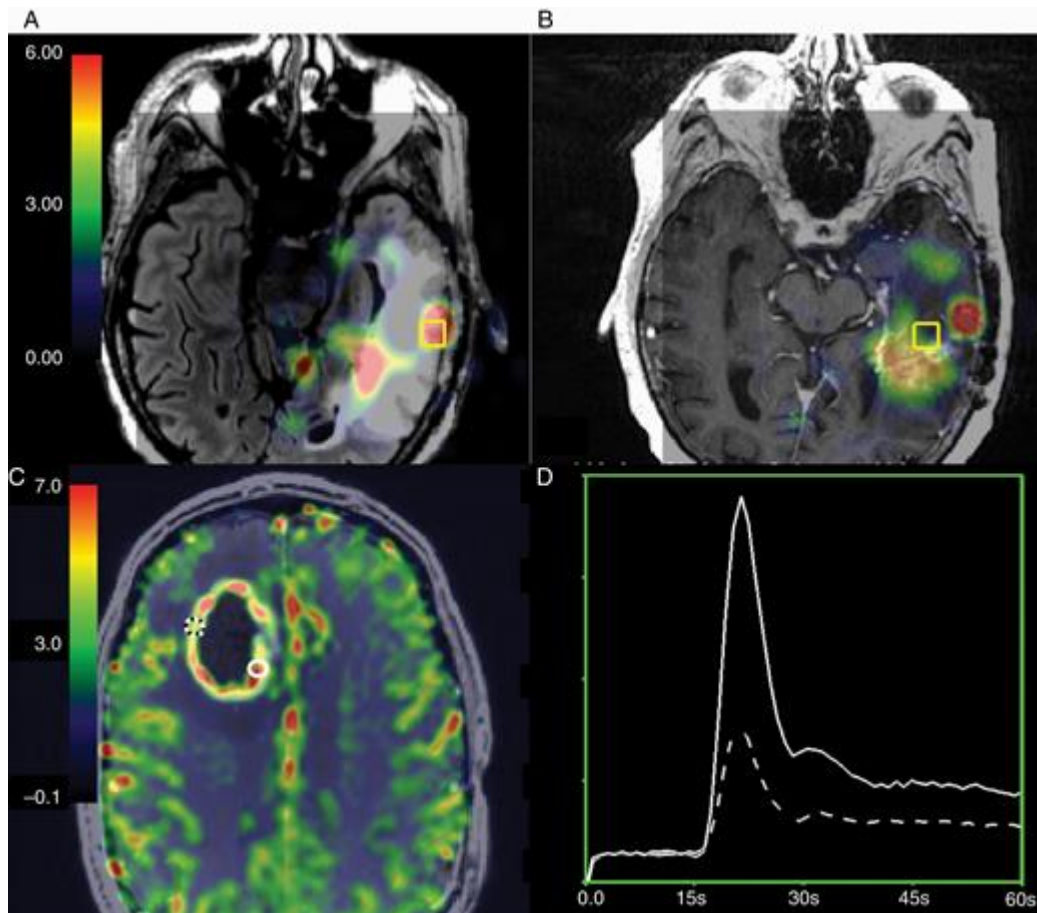
Diffusion, perfusion and spectroscopy parameters summarized from a 5-mm ROI centered on the tissue sample coordinates were tested for association with pathological outcome denoting treatment effect or recurrent HGG. Samples were analyzed both together and separately based on their location in the contrast enhancing or non-enhancing lesion. To account for the potential correlation among multiple samples derived from the same patient, univariate generalized estimating equations (GEE) were fit to the data for each MR parameter to estimate population-average coefficients, conditioning only on the fixed design matrix. The cutoff for determining significance was a p-value  $< 0.05$  after a Benjamini-Hochberg (B-H) correction for multiple testing was applied. Samples that did not have a value for a particular parameter were excluded from the analysis of that parameter. To complement these analyses, univariate GEE was also performed to evaluate the association of parameters from all samples combined irrespective of the presence of contrast enhancement. Finally, beginning with all MR parameters in a model, a backward stepwise GEE with elimination by least statistical significance was performed on all samples, while taking into account the presence or absence of contrast enhancement with an interaction term to evaluate whether individual parameter significance was upheld in a multiparametric setting.

To evaluate whether significant parameters from the previous analysis were able to separate samples into treatment effect and recurrent tumor categories, we used a cross-validation thresholding approach, where the samples were first divided into enhancing and non-enhancing groups based on their location. To create 5-folds for cross validation within each group, samples were assigned to a fold randomly while stratifying by outcome in each fold. All samples from a single patient were included in the same fold in order to ensure the independence of each fold.

To evaluate cutoff values for these imaging metrics, 4 folds were used to calculate the area under the receiver operator characteristic curve (AUC ROC) for all cutoff values and the threshold that yielded the highest AUC was chosen. This value was then applied to separate the fifth fold by outcome, and the sensitivity, specificity and accuracy of this classification were calculated to gain insight into the performance of the cutoff value. This process was repeated 5 times, providing 5 different cutoff value estimates and performance metrics. The mean and standard deviation of all thresholds and metrics derived from all 5 cross-validation experiments was calculated and reported.

In order to evaluate whether a combination of parameters could better predict outcome than thresholds alone, a logistic regression (LR) model was fit using independent significant parameters and all available samples (CEL+NEL) with that parametric information. To ensure that the validation (or 5th) fold remained independent for each cross-validation experiment, standardization was performed on the 4 folds used for training and these normalization parameters were saved for application to the validation fold. This normalization was necessary to be able to compare coefficients among regression models. The accuracy, sensitivity, and specificity of the LR models from training on the 4 folds and subsequent application of the trained model on the 5th fold was recorded. All modeling analyses were performed in R using the caret and pROC packages [163,164]. To verify this procedure we also performed a complementary bootstrapping analysis that randomly selected only one sample per patient 1000 times (200/fold), with priority given to treatment effect samples if present.

### 7.1.3 Results



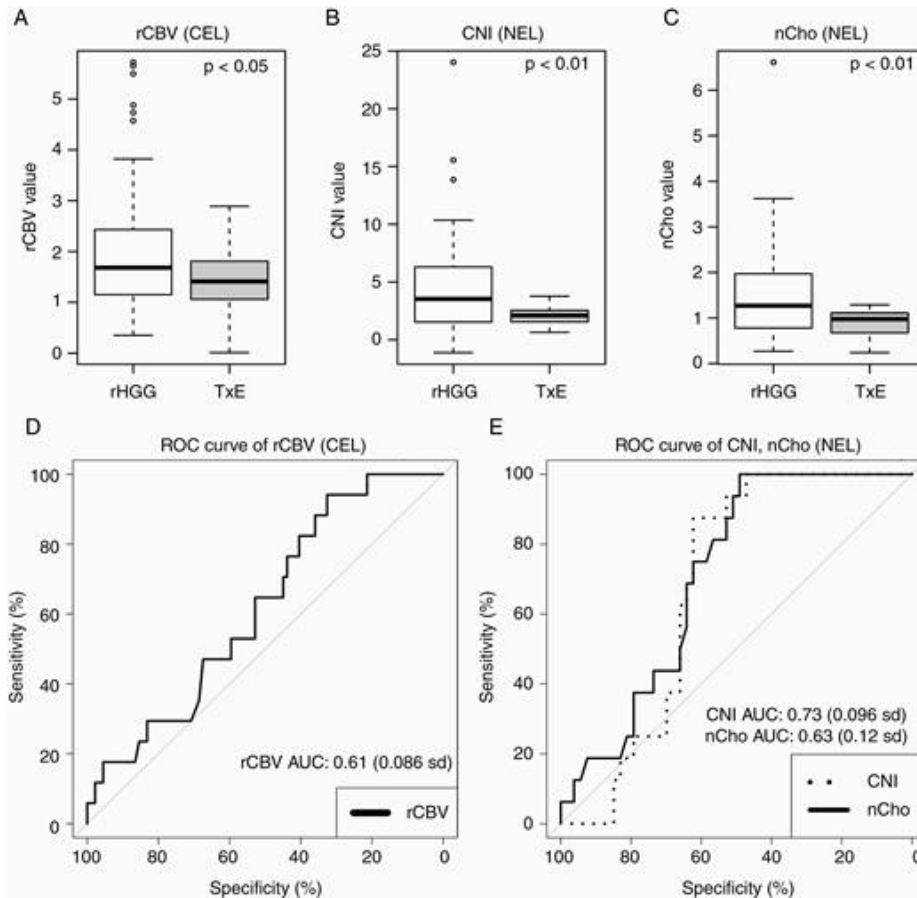
**Figure 7.2. Within-patient imaging differences between treatment-induced injury and recurrent HGG.**

(A) Non-enhancing CNI parameter maps overlaid on the T2 FLAIR image are elevated in HGG samples. (B) Reduced CNI in a sample with treatment-induced injury within the same patient (overlaid on the post-contrast T1-weighted image). (C) Enhancing rCBV parameter maps are elevated in HGG (solid white location) versus treatment-induced injury (dashed location). (D) The corresponding  $\Delta R2^*$  curves.

#### 7.1.3.1 Contrast-enhancing samples

Of the 8 MRI/MRSI parameters evaluated for samples within the CEL, the perfusion parameter rCBV was significantly associated with the binary outcome of treatment effect or recurrent tumor ( $p < 0.03$ ). Figure 7.2C&D and Figure 7.3A demonstrate the elevated levels of rCBV values among tumor samples when compared with treatment effect samples. Table 7.2

reports the number of samples included for each test. The model estimations, standard errors, Wald test statistics and p-values for all tested parameters are reported in Table 7.3.



**Figure 7.3. Boxplots representing distributions of values in recurrent HGG samples and treatment-induced injury samples.**

Visualization of (A) rCBV values from samples in the CEL, (B) CNI, and (C) nCho values from samples in the NEL. In (D) the average ROC curve of rCBV in the CEL and (E) the average ROC curve of CNI and nCho in the NEL samples. Relative CBV was significantly associated with treatment effect versus recurrent tumor ( $P < 0.03$ ) in the CEL, while spectroscopic parameters nCho and CNI were significant in the non-enhancing lesion ( $P = 0.008$ ).

### 7.1.3.2 Non-enhancing samples

In the NEL samples, however, spectroscopic parameters nCho ( $p = 0.008$ ) and CNI ( $p = 0.008$ ) were significantly associated with the presence of recurrent tumor versus treatment effect.

Figure 7.2A&B and Figure 7.3B&C demonstrate the elevated levels of total choline and CNI

values among tumor samples when compared with samples containing purely treatment effect.

Table 7.2 reports the number of samples included for each test. The model estimations, standard errors, Wald test statistics and p-values for all tested parameters are reported in Table 7.3.

**Table 7.2. Number of samples included for each test.**

	<b>Total Patients</b>	<b>Total Samples</b>	<b>Anatomic location</b>	<b>Recurrent tumor</b>	<b>Treatment Effect</b>	<b>Total Samples in</b>
Total	139	291	CEL	128	36	164
			NEL	94	33	127
Diffusion (nADC, nFA)	130	276	CEL	124	32	156
			NEL	88	32	120
Perfusion (rCBV)	101	195	CEL	89	18	107
			NEL	64	24	88
Spectroscopy (nNAA, nCho, nCre, CNI, CCRI)	79	151	CEL	70	13	83
			NEL	52	16	68

### 7.1.3.3 All samples combined

When combining the CEL and NEL regions, elevated nCho ( $p = 0.024$ ), CNI ( $p = 0.0008$ ), and CCRI ( $p = 0.012$ ) values were significantly associated with the presence of recurrent tumor. When parameters were combined in a backward stepwise generalized estimating equation (GEE) that included the anatomical lesion as an interaction term, rCBV ( $p = 0.036$ ), and CNI ( $p = 0.003$ ) remained significant in the final model.

**Table 7.3. Generalized estimating equation (GEE) results from imaging values associated with pathology.**

Parameter	Anatomic Region	Coeff estimate	Std Error	Wald Value	B-H adjusted p-value
<b>Apparent Diffusion Coefficient</b>	CEL	0.04	0.39	0.01	0.918
	NEL	0.29	0.36	0.66	0.56
	CEL+NEL	0.33	0.3	1.18	0.448
<b>Fractional Anisotropy</b>	CEL	0.11	0.45	0.07	0.913
	NEL	-0.25	0.57	0.19	0.66
	CEL+NEL	-0.25	0.37	0.45	0.571
<b>Cerebral blood volume</b>	CEL	0.31	0.11	8.72	0.024*
	NEL	0.29	0.29	0.99	0.512
	CEL+NEL	0.23	0.15	2.49	0.22
<b>N-aspartyl acetate (NAA)</b>	CEL	0.06	0.17	0.12	0.913
	NEL	0.47	0.82	0.33	0.977
	CEL+NEL	-0.37	0.40	0.85	0.48
<b>Choline</b>	CEL	0	0	2.04	0.411
	NEL	1.4	0.41	11.48	0.008**
	CEL+NEL	0.19	0.07	6.76	0.024*
<b>Creatine</b>	CEL	0	0	2.04	1.029
	NEL	0.93	0.46	4.16	0.08
	CEL+NEL	-0.06	0.15	0.19	0.66
<b>Choline-to-NAA index</b>	CEL	0	0.17	0.12	0.211
	NEL	0.29	0.09	9.85	0.008**
	CEL+NEL	0.05	0.01	15.08	0.0008**
<b>Choline-to-Creatine index</b>	CEL	0	0	2.62	0.211
	NEL	0.16	0.08	4.41	0.096
	CEL+NEL	0.06	0.02	8.98	0.012*

\*P < 0.05, \*\*P < 0.001.

#### 7.1.3.4 Cutoff analyses

After dividing the data into 5-folds for cutoff determination, the proportion of treatment effect samples in each fold ranged from 18% to 29%. The mean rCBV cutoff value for distinguishing treatment effect from recurrent tumor within the CEL was 1.62 (0.21 sd). However, when this cutoff value was applied to the fifth validation fold, the mean accuracy was only 50%, with a sensitivity of 0.50 and specificity of 0.59. For samples in the NEL, the best cutoff values to separate tumor and treatment effect were 2.71 for CNI and 1.10 for nCho. This resulted in a mean AUC ROC value of 0.73 (0.10 sd) and 0.63 (0.12 sd), for CNI and nCho respectively. When applied to the validation fold, the mean accuracy for CNI was 68%, while sensitivity and specificity were 0.61 and 0.85, respectively. The cutoff threshold for nCho resulted in an accuracy of 66%, while the mean sensitivity was 0.62 and specificity was 0.65. Table 7.4 reports these values and the standard deviations across all five-fold tests and the corresponding ROC curves are shown in Figure 7.3D&E.

#### 7.1.3.5 Multivariate model

Parameter combination into a LR using all significant MR parameters resulted in a mean AUC of 0.69 (0.09 sd). When tested on the fifth fold, the LR predicted with an average accuracy of 64% and sensitivity and specificity of 0.65 and 0.62, respectively (Table 7.4). The single-sample bootstrapping analysis confirmed these results with a mean AUC of 0.70 (95% CI [0.69, 0.72]) and an average accuracy, sensitivity, and specificity of 68%, 0.64, and 0.79, respectively, providing confidence in our predictions despite the use of multiple samples per patient.

**Table 7.4. Threshold and logistic regression analysis results.**

Values represent the means and parenthetical values are standard deviations.

<b>Modality</b>	<b>Perfusion</b>	<b>Spectroscopy</b>		<b>Perfusion + Spectroscopy</b>
<b>Analysis Type</b>	<b>Cutoff</b>	<b>Cutoff</b>		<b>Logistic Regression</b>
<b>Parameter</b>	<b>rCBV</b>	<b>CNI</b>	<b>nCho</b>	<b>rCBV, CNI</b>
<b>Anatomic Region</b>	<b>CEL</b>	<b>NEL</b>	<b>NEL</b>	<b>CEL+NEL</b>
<b>Threshold</b>	1.61 (0.21)	2.71 (0.06)	1.09 (0.03)	n/a
<b>AUC ROC</b>	0.61 (0.09)	0.73 (0.10)	0.63 (0.12)	0.69 (0.09)
<b>Accuracy %</b>	50% (7%)	68% (9%)	66% (13%)	64% (16%)
<b>Sensitivity</b>	0.50 (0.10)	0.61 (0.07)	0.62 (0.11)	0.65 (0.16)
<b>Specificity</b>	0.59 (0.25)	0.85 (0.22)	0.65 (0.22)	0.62 (0.21)

#### 7.1.4 Discussion

While distinguishing recurrent high-grade glioma from treatment induced effects is an important clinical objective, it has remained an immense challenge in part due to the coexistence of the two phenomena in the same lesion. This study attempts to overcome the problem of lesion heterogeneity by using surgical tissue samples with known coordinates on imaging acquired with neuronavigation tools. To our knowledge, this study includes the greatest number of patients with spatially mapped tissue samples to distinguish recurrent high-grade glioma from treatment-induced effects to date and takes advantage of incorporating metabolic and physiologic derived metrics from DSC perfusion weighted imaging, DTI, and MRSI. Through our statistical and predictive approaches, we demonstrated the importance of MRSI in distinguishing recurrent tumor from the effects of treatment in the non-enhancing lesion and examined whether the combination of vascular and metabolic metrics could lead to the generation of more accurate predictions.



Our results support the use of <sup>1</sup>H-MRSI for identifying regions of abnormal metabolism in the non-enhancing region that are indicative of infiltrative recurrent tumor cells rather than the effects of treatment. Tumor samples demonstrated higher levels of choline-containing metabolites (nCho) and elevated CNI compared with treatment effect samples, consistent with our current understanding of the biological underpinnings of choline and NAA metabolite flux [77]. After employing cross validation to gain insight into how a given cutoff threshold would perform when applied to external data that was not used in determining its value, our cutoffs for nCho (1.1) and CNI (2.7) separated NEL samples into treatment effect or recurrent tumor categories with mean sensitivity of 0.62, 0.61 and specificity of 0.85, 0.65, for nCho and CNI respectively. While many prior studies have used single-voxel spectroscopy in the enhancing lesion [165–167], our results indicate that spectroscopic coverage of the non-enhancing lesion area would benefit in the accurate diagnosis of recurrence. In turn, a multi-voxel spectroscopic approach would provide greater utility in assessing the metabolic lesion than single-voxel sequences, as they are often centered on the contrast enhancing region. These spectroscopic findings are similar to metabolic differences observed between vasogenic edema and enhancing metastatic disease in patients with brain metastases [166–169]. Although recent studies apply machine learning techniques to dynamic contrast enhanced (DCE) perfusion, diffusion tensor imaging (DTI) and anatomic imaging, [83] [170] the innate biological differences between metastatic brain tumors and gliomas prohibit their generalizability. Though more rigorous validation is necessary before incorporation into a clinical workflow, this study lays the groundwork for future investigation into the utility of these parameters in a prospective, independent cohort with image-guided tissue samples using more sophisticated machine learning

algorithms. This has potential to direct surgeons to which part of the non-enhancing lesion contains infiltrating tumor.

Although the recent meta-analysis of ROI-based studies by Van Dijken et al. found that the greatest sensitivity and specificity for distinguishing recurrent tumor from treatment effects lies in the spectroscopic parameters derived from the CEL, we were not able to replicate these results with our analyses [83]. This could be because ROI-based studies do not account for the spatial heterogeneity that exists in metabolism within the lesion. Despite our voxel shifting methods to reconstruct spectra at the center of the tissue sample location to avoid errors from interpolation, it was still possible that pathological heterogeneity existed within the 1cc spectral voxel. Additionally, spectral voxels at the location of CEL tissue samples often overlapped with necrotic regions and non-enhancing regions because of their larger size, potentially affecting the quantification of CNI.

Our findings in tissue samples obtained from within the contrast enhancing lesion suggest that elevated rCBV is significantly associated with recurrent tumor compared to treatment effects. The generalized estimating equations and the cutoff analysis suggest that rCBV is useful for differentiating recurrent tumor from treatment effects in the CEL. These findings are consistent with ROI-based studies that report as high as 87% sensitivity and 86% specificity when differentiating recurrent tumor from treatment-induced effects from the contrast-enhancing ROI. Although these individual smaller studies report higher sensitivity and specificity, their cutoff values were highly variable ranging from 0.71 to 3.7 for rCBV, reflecting the difficulty in recommending a universal cutoff. It was our hope that our analysis would provide clarity to this body of work by mapping local MRI characteristics directly to pathology. Although the cutoff for rCBV in our study was 1.59 (0.21 sd), our sensitivity and specificity were not significantly

better than random chance. These results may be in part attributed to the very large range of rCBV values observed in high-grade tumor tissue samples (min: 0.10, max: 5.72, med: 1.61). This large range, taken in combination with previous rCBV reports, supports the notion that a “signature” rCBV value that can distinguish high-grade tumor from treatment-related injury remains difficult to define and that multiparametric analyses with more advanced machine learning methodologies in larger datasets may be necessary to adequately address this problem.

After analyzing the association of singular MR parameters to outcome, we assessed whether the combination of parameters improves classification of samples into treatment effects or recurrent HGG. Although it does not model the potential correlation among samples derived from the same patients, logistic regression was chosen for its interpretability and reluctance to overfitting, and was further validated by comparing to the result obtained from randomly selecting 1 sample per patient and bootstrapping the data. For this analysis, 5-fold cross validation was used where each fold was separated by patient and stratified by outcome to control for information leakage and optimistic prediction. Only parameters from modalities that were determined to be useful for differentiating treatment effects from recurrent HGG in the univariate analyses were retained in the multivariate models. Combining the results into a logistic regression resulted in a model that, when compared with cut off analyses, had similar sensitivity (0.65) and specificity (0.63) when tested on the 5th fold, suggesting that modeling parameters together may not improve the classification of tissue samples by pathology. Though these results seem counterintuitive at first, it is likely that combining the anatomic regions of CEL and NEL averages out the signal that was present in each separate anatomic region, further substantiating our hypothesis that recurrent tumor in these regions have distinct metabolic and physiologic characteristics. For example, because the appearance of the CEL is driven by the

extravasation of contrast by leaky blood vessels, rCBV values in the CEL have a significantly different value distribution from those in the NEL; therefore, the signal driving the difference in rCBV in the CEL is lost when combined with samples in the NEL.

The low sensitivity in predicting pathology from MR parameters and reliably classify sub-regions of a lesion in our dataset can be attributed to three main causes. First, the parameters that were determined as being the most important for prediction of treatment effects and tumor were only obtained in 52% (perfusion) or 67% (MRSI) of patients (Table 7.2) because they were not part of routine clinical evaluation. This, along with imbalanced classes, limited our ability to build a predictive model that could be tested on an unseen dataset with these parameters. Still, using a 5-fold cross-validation approach allowed us to instead estimate the predictive value of our MR metrics, whereby we could iterate over all available treatment effect samples and observe the stability of the prediction. Although we removed samples that were largely composed of necrosis, our modest results could also be explained by the possibility of some necrosis co-existing within tissue samples that were mostly tumor or treatment effect. Although a recent review article summarizing 25 studies of brain shift reported maximum shifts between 4-31mm during the course of resection [170] the vast majority of our tissue samples were acquired with a biopsy needle before opening the dura and resecting the tumor tissue, where most reported shifts have been between 2-5mm, with maximum shifts of <10 mm. These results informed our rationale for using a 5mm ROI around the center of the biopsied sample, even though the diameter of the excised tissue was 2 mm. This is less of an issue for MRSI than other imaging metrics, because the voxel size is 1 mm<sup>3</sup> and we shift the reconstruction of the spectra voxel in k-space so that it is centered on the location of the tissue sample coordinates. Despite our efforts to further correct for errors due to tissue shift by performing extensive quality control through

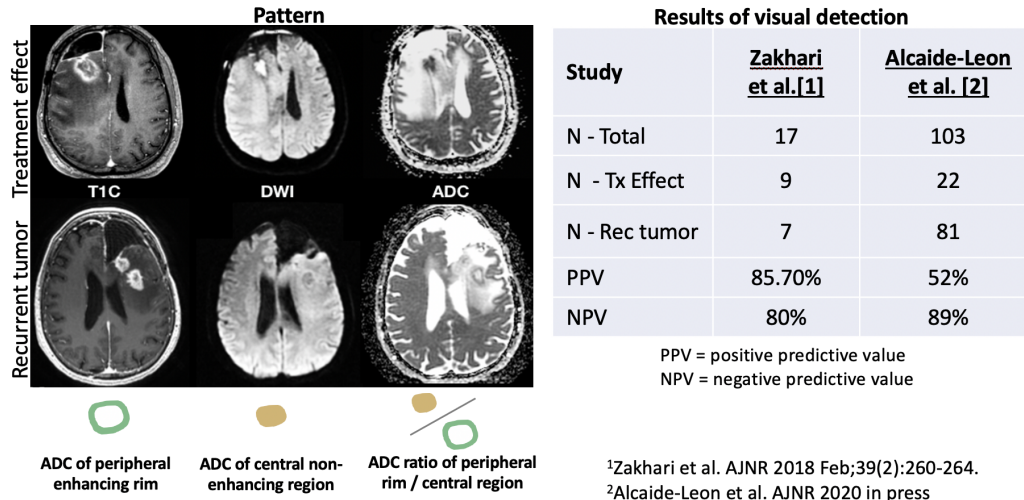
manual visualization of each tissue sample location and exclusion of samples that were structurally inaccurate, it is highly likely that this shift is the main contributor to reducing the accuracy of our results, especially when lesion heterogeneity is pronounced.

In conclusion, this study attempts to overcome the heterogeneity inherent in treated high-grade glioma lesions by mapping pathological findings directly to MR parameters. Our results suggest the need for separate MR markers of recurrent tumor for enhancing and non-enhancing lesions, highlight the potential utility of using 3D MRSI to obtain CNI maps that include the non-enhancing region in the recurrent high-grade setting, and support previous studies that suggest rCBV should be used to differentiate treatment effects from recurrent tumor within the contrast enhancing region. These findings lay the foundation for a larger, multi-institutional investigation that includes MRSI of the non-enhancing region and multiparametric MRI, along with machine learning for differentiation of treatment induced injury from true recurrent tumor.

## 7.2 Using anatomic and diffusion MRI with deep convolutional neural networks to distinguish treatment-induced injury from recurrent glioblastoma

### 7.2.1 Introduction

It is estimated that 25% to 35% of patients with high-grade glioma experience treatment-associated injury that can mimic recurrence[32,145], posing a significant diagnostic challenge for radiologists. It is critical to distinguish these phenomena in order to accurately assess patient response to therapy and ensure that the most effective therapies are used. However, previous attempts to solve this problem have been limited by sample size, using single tissue samples to determine the ground truth outcome of a lesion, and focusing on MR parameters summarized only from the contrast-enhancing area of the lesion. In fact, centrally restricted diffusion has recently shown promise in assessing the presence of treatment-associated injury (Figure 7.4)[171,172], providing evidence of additional relevant MR signal from the necrotic region. In this study, we hypothesize that we can overcome some of the limitations of prior work by a) have the largest patient sample size to date; b) using the entire lesion image; and c) using multiple tissue samples per patient to establish a rigorous ground truth. Our work aims to demonstrate the feasibility of exploiting this pattern using convolutional neural networks to provide faster classification and enhanced sensitivity and specificity compared with visual assessment of the presence of this phenomenon.



**Figure 7.4. Summarizing the centrally restricted diffusion sign.**

T1 post-contrast (A, D), trace DWI image (B, E) and ADC map (C, F) of a patient with a treatment-induced lesion (A, B, C) and a patient with GBM recurrence (D, E, F). In the lesion associated with treatment, DWI and corresponding ADC maps show restricted diffusion in the central enhancing portion of the lesion. In the recurrence, DWI and ADC show restricted diffusion along the peripheral enhancing rim and facilitated diffusion in the central necrotic region.

## 7.2.2 Methods

### 7.2.2.1 Subjects and Image Acquisition

A total of 174 patients with suspected recurrent glioblastoma were scanned on a 3T scanner with an 8-channel head coil. T2-weighted FLAIR (T2 FLAIR) and 3D T1-weighted IR-SPGR imaging before (T1) and after (T1C) the injection of a gadolinium-based contrast agent were acquired. Diffusion-tensor images (DTI) were obtained in the axial plane with either 6 directions and 4 excitations or  $\geq 24$  directions and 1 excitation [TR/TE = 1000/108 ms, voxel size =  $1.7 \times 1.7 \times 3$  mm, b=1000 or 2000 s/mm] and apparent diffusion coefficient (ADC) maps were generated using FMRIB's Diffusion Toolkit [116].

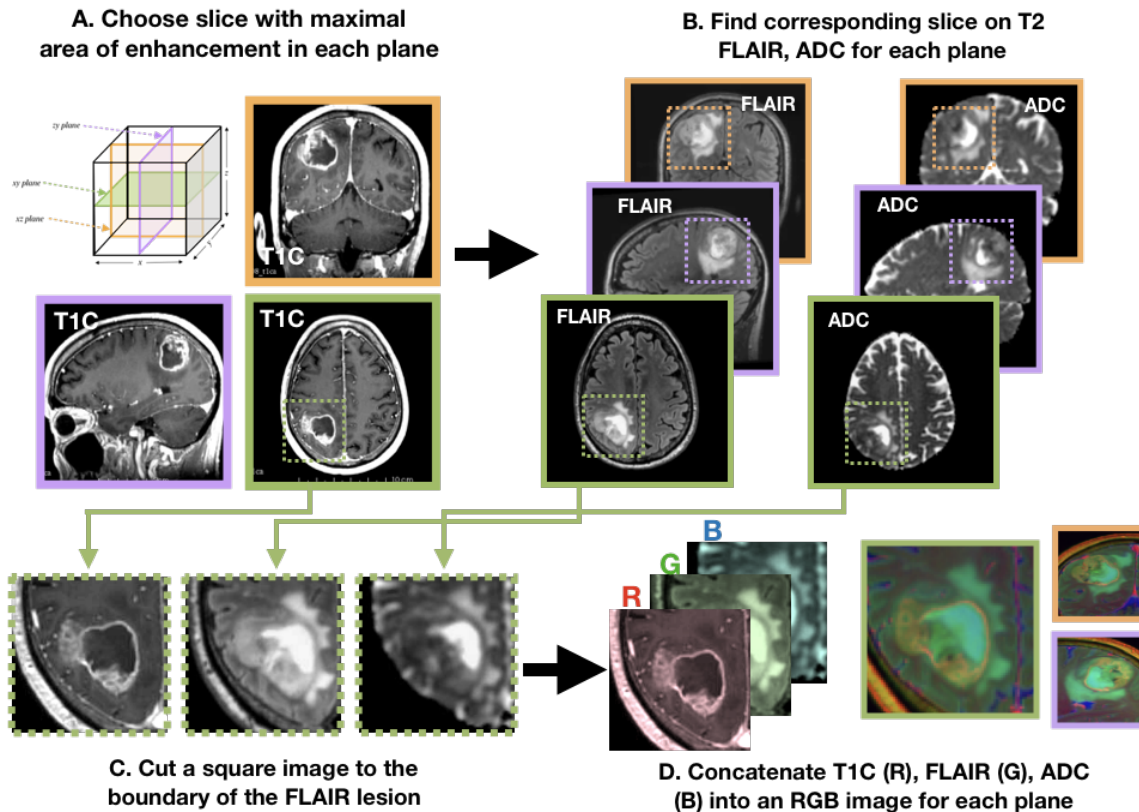
### 7.2.2.2 Immunohistochemistry

A board-certified pathologist evaluated multiple tissue samples per patient and scored them on the presence of tumor cells. In order to have a designation of treatment effect, the lesions must have had at least two samples containing no viable tumor and with signs of treatment effect (e.g. gliosis, hyalinized blood vessels); additionally, they could not have a tissue sample containing tumor in the newly enhancing lesion, as visualized by neuronavigation software during surgery. Patients with the designation of true recurrence must have had a clinical diagnosis at the time of surgery as glioblastoma, while also at least one research tissue sample with evidence of recurrent tumor. With this approach, 32 patients were classified as having treatment-induced effects, and 142 patients were classified as having recurrent glioblastoma.

### 7.2.2.3 Image processing

Figure 7.5 details the image processing pipeline. Each image was multiplied by a brain mask and z-score normalization was performed. The maximum area of the contrast enhancing lesion (CEL) or nonenhancing lesion (NEL) as determined by the ROI area was used to select the slice of interest in each plane. A square box padding the bounds of the NEL was used to extract a patch from the T1C, T2 FLAIR, and ADC images. These patches from each MR modality were concatenated to generate a single RGB-color image for each plane (coronal, sagittal and axial), for each patient.





**Figure 7.5. Image processing pipeline.**

A) Slice with maximal contrast enhancing lesion (CEL) region of interest is chosen in each direction (axial, green; coronal, orange; sagittal, purple). If no CEL, the nonenhancing lesion (NEL) is used. B) The corresponding slice on T2 FLAIR and ADC images are chosen. C) A square bounding box of the NEL on the T2 FLAIR image is created and used on ADC, T1C; repeated for all planes. D) Concatenating the T1C, T2 FLAIR and ADC images together forms an RGB image. These choices were repeated every 5 mm away from the maximal slice if using more than three slices per patient.

#### 7.2.2.4 Model and computational framework

Patients were first split into training and testing in an 80/20 split, such that 144 patients (27 treatment effect, 117 glioblastoma) were in the training set and 30 patients (5 treatment effect, 25 glioblastoma). From the training data, patients were split into 5 folds and stratified based on outcome. A variety of model architectures, slice selection strategies, training strategies, and hyperparameters were searched through in order to find the most appropriate strategy (Figure 7.6). During training, the minority (TxE) class was oversampled 4x through data

augmentation to account for the fewer number of patients in the TxE group. Data was augmented using cropping, rotating, and flipping both horizontally and vertically. Cosine differential learning rates (LR) and cross-entropy loss functions were used during training. The logit output of the final network layer was averaged across the slices directions for each patient and a sigmoid function was applied. All model building, training, and testing were implemented using PyTorch 1.0.0 on a Tesla V100-PCIE-32GB GPU (Nvidia).

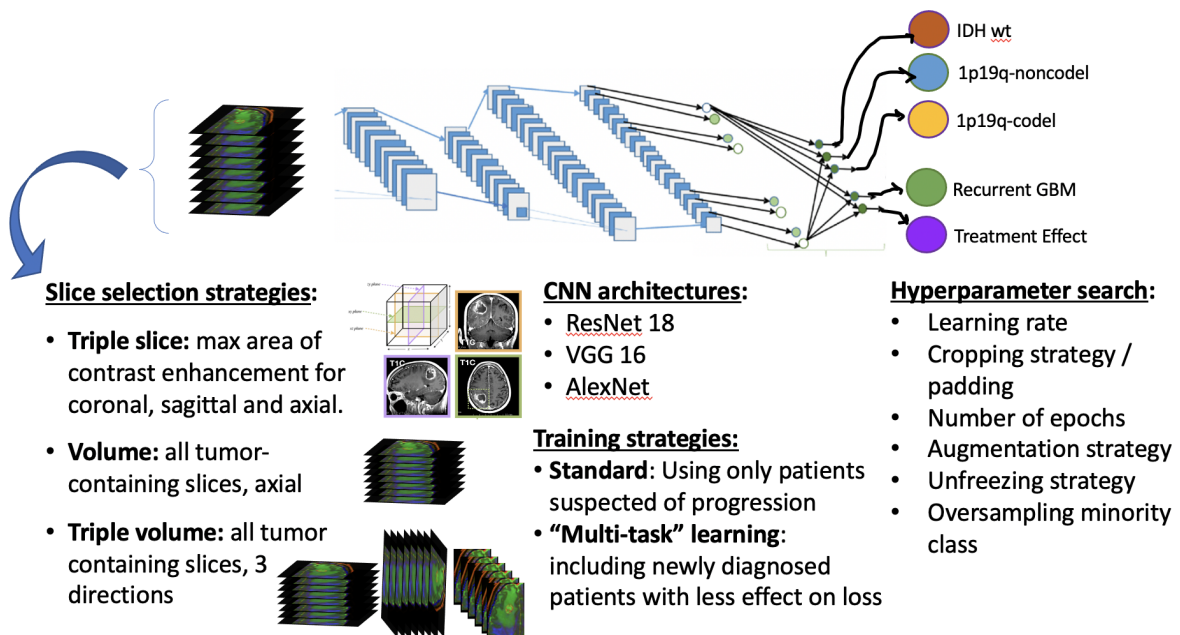


Figure 7.6. Modeling approaches.

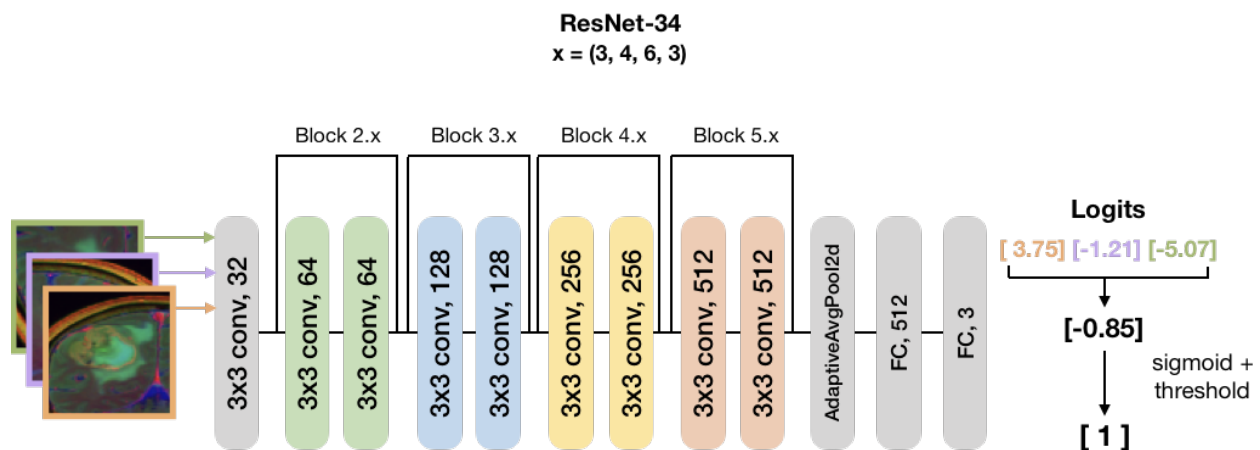
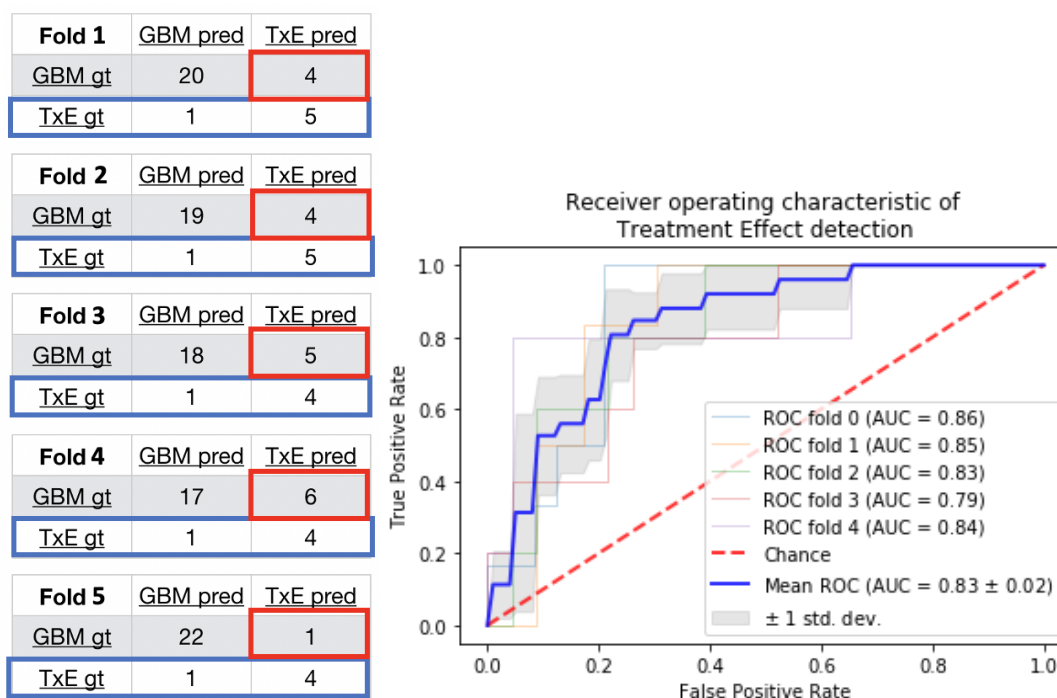


Figure 7.7. ResNet-34 CNN architecture and 3-direction averaging technique utilized.

Before performing a sigmoid function, the raw logit is saved for the coronal (orange), sagittal (purple) and axial (green) planes. An average logit is taken and a sigmoid function is applied to create a single probability of treatment effect per patient.

### 7.2.3 Results and Discussion

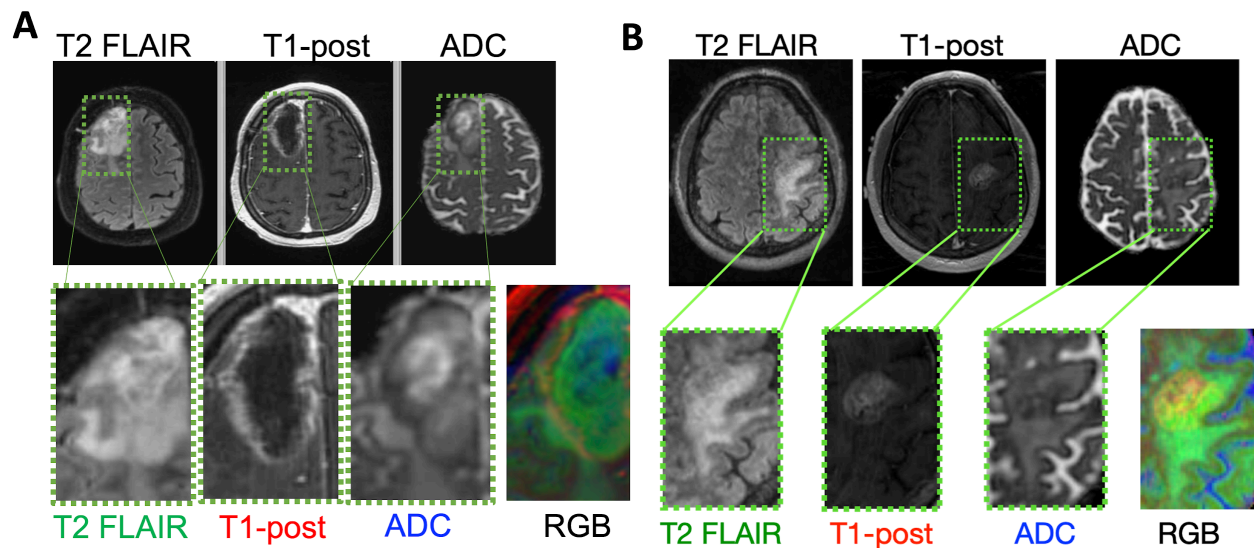
The best result was achieved using a ResNet-34 with just three slices derived from the maximal area of contrast enhancement in the coronal, sagittal and axial plane. Our pre-trained ResNet-34 model resulted in a 5-fold cross validation average AUC ROC of 0.83 +/- 0.02 for the classification of patients into their respective categories (Figure 7.8). In addition, we report the confusion matrices achieved when choosing a threshold corresponding to the top left point on the AUC ROC curve. Our average sensitivity and specificity, 81.3% and 82.7%, are similar to that reported by Zakhari et al. (Figure 7.4); however, our cohort includes over 8 times as many patients, and 3 times the number of TxE lesions.



**Figure 7.8.** ROC curve for each fold using the probabilities derived from the sigmoid value of the average logit.

We examined the misclassification of both treatment-effect and recurrent tumor patients and found that treatment effect patients that were misclassified had patterns that had enhanced diffusion within the necrotic region, i.e. that is usually typical of recurrence (Figure 7.9A). We

found this to be true in most misclassified treatment effect cases. When examining misclassified glioblastoma cases, we also found that this was sometimes due to a correctly predicted treatment-effect-like appearance with ADC uniformly reduced with restricted central diffusion (Figure 7.9B). These results could suggest that within these cross validation folds, the network was correctly learning the centrally restricted diffusion pattern.

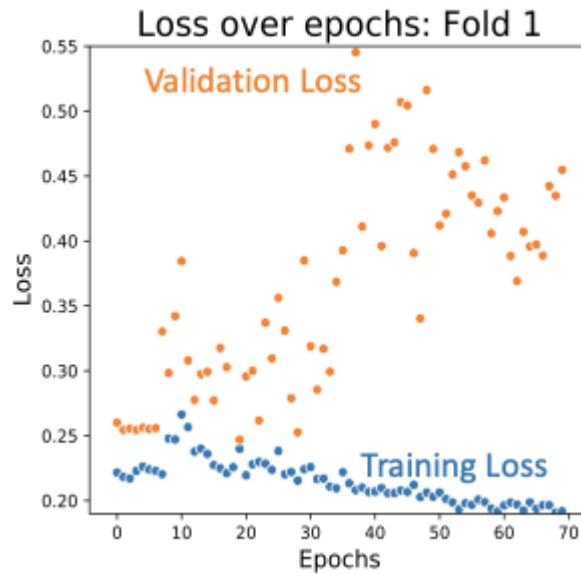


**Figure 7.9. Examples of misclassified patients.**

(A) Example of a treatment-effect lesion that was misclassified as a recurrent glioblastoma. It exhibits a glioblastoma-like appearance with elevated diffusion within the central necrotic region. (B) Example of a recurrent glioblastoma misclassified as a treatment-effect lesion. We observe centrally restricted diffusion as shown by the ADC.

Finally, we tested our models on the holdout testing set. No model was able to generalize to the test set with results comparable to that during five-fold cross validation. We surmise that despite having the largest dataset tackling this problem to date, we still did not have enough data to densely cover the variability in recurrent glioblastoma and treatment-induced effect lesions that exist in the population. We investigated the training and validation set loss curves over the training epochs (Figure 7.10) and discovered an expanding validation loss of the validation fold

during training. Compounding these problems, each “best” model from each of the five folds was trained for a differing number of epochs, and widely varying cutoff points from the AUC ROC implied that it would be difficult to know how long to train and what cutoff to use to predict treatment effect or tumor.



**Figure 7.10. Increasing validation loss while overfitting to the training fold.**

We surmise that the overfitting was partially due to the small sample size for deep learning compounded with the need to oversample the minority class. We tried many strategies to reduce overfitting and regularize the models, but it is likely that our dataset is simply not representative of the full spectrum of heterogeneity in these lesions.

The vast majority of prior efforts in this field use radiomics or calculate the median diffusion-, perfusion-, or spectroscopic- derived parameter value from an anatomical ROI to obtain a threshold that can distinguish treatment effects from true tumor recurrence [83,173]; however, these methods suffer from either requiring manual intervention for selecting regions of interest, lack of pathological confirmation, low sample size, or assessing associations rather than

prediction ability. Our approach is unique in that it leverages: 1) multiple tissue samples for generating a single pathological outcome in the analysis, 2) transfer learning with the merging multi-contrast images into one input, 3) a substantially larger dataset than prior studies, and 4) a fully automated pipeline.

#### 7.2.4 Conclusion

In this study, we lay the groundwork and report results from an initial attempt at leveraging a promising new centrally restricted diffusion pattern together with modern advances in deep learning to create a novel strategy for detecting treatment-associated injury in the context of suspected recurrent glioblastoma. Through combining clinical diffusion-weighted imaging with standard anatomical imaging and transfer learning, it was possible for the network to learn the restricted diffusion patterns that are characteristic of recurrent glioblastoma and treatment related injury. Although similarly high accuracy was found for all 5-folds in validation (0.79-0.86), the network was not able to generalize to a separate dataset of similar proportions, despite hyperparameter optimization. This could be due to the small size of the TxE cohort in training, even though 27 pure treatment effect larger than what has been reported in prior studies or the potential presence of treatment effect within the heterogeneous recurrent tumor lesions influencing the characteristics of the restricted diffusion pattern.

## 8. Conclusions and future directions

### 8.1 Conclusions

This dissertation investigated three clinically-relevant questions whose answers lead toward improving the diagnosis of glioma:

- Upon a new diagnosis of glioma, can we presurgically identify the genetic alterations using deep learning and MR imaging? (Chapter 5)
- While monitoring a patient over time, can we reliably deliver longitudinal brain images of the same MR contrast to a clinician? (Chapter 6)
- Once a suspected recurrence is observed on MR imaging, can we reliably identify whether it is a true recurrence or a treatment-induced lesion? (Chapter 7)

Chapters 2, 3, and 4 dive into the fundamentals of brain tumor biology, of MR for brain tumor imaging, and machine learning in order to provide a foundation for understanding the methods used in Chapters 5, 6, and 7.

To our knowledge, the study presented in Chapter 5 is the first of its kind to answer the first question by a) classifying both IDH and 1p19q together using imaging and deep learning; b) investigating the impact of including apparent diffusion coefficient maps; and c) reporting differences between different deep learning strategies complemented by training/validation loss curves and a feature attribution technique. We evaluate many different modeling strategies and hyperparameters in order to create an algorithm capable of predicting IDH mutation status and 1p19q codeletion status. We found that a multiclass one-step classification system performed better than trying to first predict IDH, then 1p19q. We also found evidence that ADC is useful for generalizing to unseen data. Finally, we gain insights into different modeling paradigms that

can potentially inform other researchers' study designs in the future: full slices contain information about the location of the tumor, which is essential for reliable prediction of IDH mutation status.

In Chapter 6, we compare the performance of 1) a rule-based classifier; 2) a machine learning model using metadata alone; 3) a deep learning imaging model; and 4) a machine learning model that combines the outputs of the deep learning model with the metadata ("combined"). All three algorithms vastly improved the performance of the rule-based approach, with the deep learning and combination models performing the best in different scenarios. We found that algorithms containing imaging data (3 and 4) performed the best on poorly annotated, heterogeneous PACS data, but the deep learning algorithm alone (3) was able to correctly identify MR series that were impacted by artifacts.

In Chapter 7, we report two strategies that aim to improve the ability of clinicians to differentiate treatment-induced lesions from true recurrent high-grade glioma. In the first study, we attempt to overcome the heterogeneity inherent in treated high-grade glioma lesions by mapping pathological findings directly to MR parameters. Our results suggest the need for separate MR markers of recurrent tumor for enhancing and non-enhancing lesions, highlight the potential utility of using 3D MRSI to obtain CNI maps that include the non-enhancing region in the recurrent high-grade setting, and support previous studies that suggest rCBV should be used to differentiate treatment effects from recurrent tumor within the contrast enhancing region. In the second, we combine pretrained deep learning algorithms together with a promising new centrally restricted diffusion pattern in order to predict whether a patient is experiencing a treatment-induced lesion or a true recurrence. Though our algorithm doesn't generalize to our test set, we lay the groundwork for investigating this problem with deep learning by reporting



strategies that worked best during five-fold cross validation. We hope that our results can help guide future studies that have larger datasets.

## 8.2 Future directions

In Chapter 5, we observe better prediction of 1p19q codeletion status when we include ADC, and it is possible it would be even better with the inclusion of perfusion maps or other advanced MR modalities. Another valuable future direction would be to include the whole TCGA-TCIA cohort instead of just the BRATS data so that we can include diffusion imaging acquired during patients exams and expand our test set cohort. In addition, larger datasets of IDH mutated patients are becoming more readily available which could greatly improve the accuracy of similar studies. We hope that these results will incite future collaborations allowing for even better model generalization.

In Chapter 6, we compare multiple algorithms to classify MR volumes by their contrast, which serves a crucial step in the UCSF Multiple Sclerosis Bioscreen tool deployed in the neurology clinic. Though we discuss the feasibility of generalizing the algorithm to another disease, we only present results on research cohorts of glioma rather than poorly-annotated PACS glioma data. In order to safely deploy our algorithm in the context of the UCSF Neuro-Oncology clinic, we would likely need to label a poorly-annotated PACS dataset manually as well to ensure that our algorithm would generalize well. We also intend to include segmentation in this pipeline to automatically display volumetric changes over time.

In Chapter 7, we discuss two completely separate attempts to solve the same problem of distinguishing true recurrent glioblastoma from lesions induced by treatment. In Chapter 7.1, our main findings include that MRSI parameters in the non-enhancing region differentiate treatment-

induced effects from recurrent glioma best, which was the first of its kind. Therefore, a very valuable future direction would be a larger, multi-institutional investigation that includes MRSI of the non-enhancing region and multiparametric MRI. In Chapter 7.2, we lay the groundwork for a deep learning study that investigates the same question. However, despite having the most patients and samples of all published studies investigating this question, our sample sizes for both investigations were still not large enough to properly conduct a machine learning study combining the MR parameters. Therefore, multi-institutional data-sharing and collaboration efforts to solve this problem are the most promising future direction for this incredibly important problem.

## References

- [1] M. Neska-Matuszewska, J. Bladowska, M. Sasiadek, A. Zimny, Differentiation of glioblastoma multiforme, metastases and primary central nervous system lymphomas using multiparametric perfusion and diffusion MR imaging of a tumor core and a peritumoral zone-Searching for a practical approach., *PLoS One*. 13 (2018) e0191341. doi:10.1371/journal.pone.0191341.
- [2] Q.T. Ostrom, N. Patil, G. Cioffi, K. Waite, C. Kruchko, J.S. Barnholtz-Sloan, CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2013-2017., *Neuro. Oncol.* 22 (2020) iv1-iv96. doi:10.1093/neuonc/noaa200.
- [3] Q.T. Ostrom, L. Bauchet, F.G. Davis, I. Deltour, J.L. Fisher, C.E. Langer, et al., The epidemiology of glioma in adults: a “state of the science” review., *Neuro. Oncol.* 16 (2014) 896–913. doi:10.1093/neuonc/nou087.
- [4] D.N. Louis, H. Ohgaki, O.D. Wiestler, W.K. Cavenee, P.C. Burger, A. Jouvet, et al., The 2007 WHO classification of tumours of the central nervous system., *Acta Neuropathol.* 114 (2007) 97–109. doi:10.1007/s00401-007-0243-4.
- [5] D.N. Louis, P. Wesseling, K. Aldape, D.J. Brat, D. Capper, I.A. Cree, et al., cIMPACT-NOW update 6: new entity and diagnostic principle recommendations of the cIMPACT-Utrecht meeting on future CNS tumor classification and grading., *Brain Pathol.* 30 (2020) 844–856. doi:10.1111/bpa.12832.
- [6] D.N. Louis, D.W. Ellison, D.J. Brat, K. Aldape, D. Capper, C. Hawkins, et al., cIMPACT-NOW: a practical summary of diagnostic points from Round 1 updates., *Brain*

- Pathol. 29 (2019) 469–472. doi:10.1111/bpa.12732.
- [7] A. Olar, K.M. Wani, K.D. Alfaro-Munoz, L.E. Heathcock, H.F. van Thuijl, M.R. Gilbert, et al., IDH mutation status and role of WHO grade and mitotic index in overall survival in grade II-III diffuse gliomas., *Acta Neuropathol.* 129 (2015) 585–596.  
doi:10.1007/s00401-015-1398-z.
- [8] P.J. Cimino, E.C. Holland, Targeted copy number analysis outperforms histologic grading in predicting patient survival for WHO grades II/III IDH-mutant astrocytomas., *Neuro. Oncol.* 21 (2019) 819–821. doi:10.1093/neuonc/noz052.
- [9] D.E. Reuss, Y. Mamatjan, D. Schrimpf, D. Capper, V. Hovestadt, A. Kratz, et al., IDH mutant diffuse and anaplastic astrocytomas have similar age at presentation and little difference in survival: a grading problem for WHO., *Acta Neuropathol.* 129 (2015) 867–873. doi:10.1007/s00401-015-1438-8.
- [10] C.D. Arvanitis, G.B. Ferraro, R.K. Jain, The blood-brain barrier and blood-tumour barrier in brain tumours and metastases., *Nat. Rev. Cancer.* 20 (2020) 26–41.  
doi:10.1038/s41568-019-0205-x.
- [11] B.W. Chow, C. Gu, The molecular constituents of the blood-brain barrier., *Trends Neurosci.* 38 (2015) 598–608. doi:10.1016/j.tins.2015.08.003.
- [12] S. Ayloo, C. Gu, Transcytosis at the blood-brain barrier., *Curr. Opin. Neurobiol.* 57 (2019) 32–38. doi:10.1016/j.conb.2018.12.014.
- [13] M.S. Thomsen, L.J. Routhe, T. Moos, The vascular basement membrane in the healthy and pathological brain., *J. Cereb. Blood Flow Metab.* 37 (2017) 3300–3317.  
doi:10.1177/0271678X17722436.
- [14] E.K. Nduom, C. Yang, M.J. Merrill, Z. Zhuang, R.R. Lonser, Characterization of the

- blood-brain barrier of metastatic and primary malignant neoplasms., *J. Neurosurg.* 119 (2013) 427–433. doi:10.3171/2013.3.JNS122226.
- [15] N.A. Oberheim Bush, S. Chang, Treatment Strategies for Low-Grade Glioma in Adults., *J. Oncol. Pract.* 12 (2016) 1235–1241. doi:10.1200/JOP.2016.018622.
- [16] M.J. van den Bent, D. Afra, O. de Witte, M. Ben Hassel, S. Schraub, K. Hoang-Xuan, et al., Long-term efficacy of early versus delayed radiotherapy for low-grade astrocytoma and oligodendroglioma in adults: the EORTC 22845 randomised trial., *Lancet.* 366 (2005) 985–990. doi:10.1016/S0140-6736(05)67070-5.
- [17] R. Stupp, W.P. Mason, M.J. van den Bent, M. Weller, B. Fisher, M.J.B. Taphoorn, et al., Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma., *N. Engl. J. Med.* 352 (2005) 987–996. doi:10.1056/NEJMoa043330.
- [18] S.Y. Lee, Temozolomide resistance in glioblastoma multiforme., *Genes Dis.* 3 (2016) 198–210. doi:10.1016/j.gendis.2016.04.007.
- [19] J.R. Wesolowski, P. Rajdev, S.K. Mukherji, Temozolomide (Temodar)., *AJNR Am. J. Neuroradiol.* 31 (2010) 1383–1384. doi:10.3174/ajnr.A2170.
- [20] M.E. Hegi, A.-C. Diserens, T. Gorlia, M.-F. Hamou, N. de Tribolet, M. Weller, et al., MGMT gene silencing and benefit from temozolomide in glioblastoma., *N. Engl. J. Med.* 352 (2005) 997–1003. doi:10.1056/NEJMoa043331.
- [21] M.J. van den Bent, A.A. Brandes, M.J.B. Taphoorn, J.M. Kros, M.C.M. Kouwenhoven, J.-Y. Delattre, et al., Adjuvant procarbazine, lomustine, and vincristine chemotherapy in newly diagnosed anaplastic oligodendroglioma: long-term follow-up of EORTC brain tumor group study 26951., *J. Clin. Oncol.* 31 (2013) 344–350. doi:10.1200/JCO.2012.43.2229.

- [22] J.G. Cairncross, M. Wang, R.B. Jenkins, E.G. Shaw, C. Giannini, D.G. Brachman, et al., Benefit from procarbazine, lomustine, and vincristine in oligodendroglial tumors is associated with mutation of IDH., *J. Clin. Oncol.* 32 (2014) 783–790. doi:10.1200/JCO.2013.49.3726.
- [23] S. Choi, Y. Yu, M.R. Grimmer, M. Wahl, S.M. Chang, J.F. Costello, Temozolomide-associated hypermutation in gliomas., *Neuro. Oncol.* 20 (2018) 1300–1309. doi:10.1093/neuonc/noy016.
- [24] S. Chang, P. Zhang, J.G. Cairncross, M.R. Gilbert, J.-P. Bahary, C.A. Dolinskas, et al., Phase III randomized study of radiation and temozolomide versus radiation and nitrosourea therapy for anaplastic astrocytoma: results of NRG Oncology RTOG 9813., *Neuro. Oncol.* 19 (2017) 252–258. doi:10.1093/neuonc/now236.
- [25] N.A. Oberheim Bush, Y. Yu, J. Villanueva-Meyer, M. Grimmer, S. Hilz, D. Solomon, et al., Temozolomide-induced hypermutation is associated with high-grade transformation, distant recurrence, and reduced survival after transformation in initially low-grade *IDH*-mutant diffuse gliomas., *JCO.* 38 (2020) 2506–2506. doi:10.1200/JCO.2020.38.15\_suppl.2506.
- [26] M. Mehta, P. Wen, R. Nishikawa, D. Reardon, K. Peters, Critical review of the addition of tumor treating fields (TTFields) to the existing standard of care for newly diagnosed glioblastoma patients., *Crit. Rev. Oncol. Hematol.* 111 (2017) 60–65. doi:10.1016/j.critrevonc.2017.01.005.
- [27] A.B. Lassman, A.E. Joanta-Gomez, P.C. Pan, W. Wick, Current usage of tumor treating fields for glioblastoma., *Neurooncol Adv.* 2 (2020) vdaa069. doi:10.1093/noajnl/vdaa069.
- [28] S.A. Toms, C.Y. Kim, G. Nicholas, Z. Ram, Increased compliance with tumor treating

- fields therapy is prognostic for improved survival in the treatment of glioblastoma: a subgroup analysis of the EF-14 phase III trial., *J. Neurooncol.* 141 (2019) 467–473.  
doi:10.1007/s11060-018-03057-z.
- [29] K. Tanimoto, K. Yoshiga, H. Eguchi, M. Kaneyasu, K. Ukon, T. Kumazaki, et al., Hypoxia-inducible factor-1alpha polymorphisms associated with enhanced transactivation capacity, implying clinical significance., *Carcinogenesis.* 24 (2003) 1779–1783.  
doi:10.1093/carcin/bgg132.
- [30] D.H. Gorski, M.A. Beckett, N.T. Jaskowiak, D.P. Calvin, H.J. Mauceri, R.M. Salloum, et al., Blockage of the vascular endothelial growth factor stress response increases the antitumor effects of ionizing radiation., *Cancer Res.* 59 (1999) 3374–3378.
- [31] P.H. Gutin, F.M. Iwamoto, K. Beal, N.A. Mohile, S. Karimi, B.L. Hou, et al., Safety and efficacy of bevacizumab with hypofractionated stereotactic irradiation for recurrent malignant gliomas., *Int. J. Radiat. Oncol. Biol. Phys.* 75 (2009) 156–163.  
doi:10.1016/j.ijrobp.2008.10.043.
- [32] A.W. Abbasi, H.E. Westerlaan, G.A. Holtman, K.M. Aden, P.J. van Laar, A. van der Hoorn, Incidence of Tumour Progression and Pseudoprogression in High-Grade Gliomas: a Systematic Review and Meta-Analysis., *Clin Neuroradiol.* 28 (2018) 401–411.  
doi:10.1007/s00062-017-0584-x.
- [33] Z.-P. Liang, P.C. Lauterbur, Chapter 1: Introduction, in: *Principles of Magnetic Resonance Imaging: A Signal Processing Perspective*, IEEE, 2000: pp. 1–11.  
<https://ieeexplore.ieee.org/document/5264291/authors#authors> (accessed November 25, 2020).
- [34] D. Nishimura, Chapter 1, in: *Principles of Magnetic Resonance Imaging*, 2010.

- [35] F. Bloch, Nuclear Induction, *Phys. Rev.* 70 (1946) 460–474.  
doi:10.1103/PhysRev.70.460.
- [36] D.W. McRobbie, E.A. Moore, M.J. Graves, M.R. Prince, *MRI from picture to proton*, Cambridge University Press, Cambridge, 2006. doi:10.1017/CBO9780511545405.
- [37] H. Schild, *MRI Made Easy*, 1990.
- [38] M.L. Lipton, *Totally Accessible MRI*, Springer New York, New York, NY, 2008.  
doi:10.1007/978-0-387-48896-7.
- [39] A.D. Elster, MD FACR, Image contrast - Questions and Answers in MRI, *Questions and Answers in MRI*. (2021). <https://mriquestions.com/image-contrast-trte.html> (accessed January 18, 2021).
- [40] J.E. Villanueva-Meyer, M.C. Mabray, S. Cha, Current clinical brain tumor imaging., *Neurosurgery*. 81 (2017) 397–415. doi:10.1093/neuros/nyx103.
- [41] J.D. Costabile, E. Alaswad, S. D’Souza, J.A. Thompson, D.R. Ormond, Current applications of diffusion tensor imaging and tractography in intracranial tumor resection., *Front. Oncol.* 9 (2019) 426. doi:10.3389/fonc.2019.00426.
- [42] A.M. Molinaro, S. Hervey-Jumper, R.A. Morshed, J. Young, S.J. Han, P. Chunduru, 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 Oncol.* 6 (2020) 495–503.  
doi:10.1001/jamaoncol.2019.6143.
- [43] Q. Wen, L. Jalilian, J.M. Lupo, A.M. Molinaro, S.M. Chang, J. Clarke, et al., Comparison of ADC metrics and their association with outcome for patients with newly diagnosed glioblastoma being treated with radiation therapy, temozolomide, erlotinib and



- bevacizumab., *J. Neurooncol.* 121 (2015) 331–339. doi:10.1007/s11060-014-1636-6.
- [44] W. Chang, W.B. Pope, R.J. Harris, A.J. Hardy, K. Leu, R.R. Mody, et al., Diffusion MR Characteristics Following Concurrent Radiochemotherapy Predicts Progression-Free and Overall Survival in Newly Diagnosed Glioblastoma., *Tomography.* 1 (2015) 37–43. doi:10.18383/j.tom.2015.00115.
- [45] W.B. Pope, H.J. Kim, J. Huo, J. Alger, M.S. Brown, D. Gjertson, et al., Recurrent glioblastoma multiforme: ADC histogram analysis predicts response to bevacizumab treatment., *Radiology.* 252 (2009) 182–189. doi:10.1148/radiol.2521081534.
- [46] P. Eichinger, E. Alberts, C. Delbridge, S. Trebeschi, A. Valentinitich, S. Bette, et al., Diffusion tensor image features predict IDH genotype in newly diagnosed WHO grade II/III gliomas., *Sci. Rep.* 7 (2017) 13396. doi:10.1038/s41598-017-13679-4.
- [47] Y.W. Park, K. Han, S.S. Ahn, S. Bae, Y.S. Choi, J.H. Chang, et al., Prediction of IDH1-Mutation and 1p/19q-Codeletion Status Using Preoperative MR Imaging Phenotypes in Lower Grade Gliomas., *AJNR Am. J. Neuroradiol.* 39 (2018) 37–42. doi:10.3174/ajnr.A5421.
- [48] M.K. Lee, J.E. Park, Y. Jo, S.Y. Park, S.J. Kim, H.S. Kim, Advanced imaging parameters improve the prediction of diffuse lower-grade gliomas subtype, IDH mutant with no 1p19q codeletion: added value to the T2/FLAIR mismatch sign., *Eur. Radiol.* 30 (2020) 844–854. doi:10.1007/s00330-019-06395-2.
- [49] C. Asao, Y. Korogi, M. Kitajima, T. Hirai, Y. Baba, K. Makino, et al., Diffusion-weighted imaging of radiation-induced brain injury for differentiation from tumor recurrence., *AJNR Am. J. Neuroradiol.* 26 (2005) 1455–1460.
- [50] P.C. Sundgren, X. Fan, P. Weybright, R.C. Welsh, R.C. Carlos, M. Petrou, et al.,

- Differentiation of recurrent brain tumor versus radiation injury using diffusion tensor imaging in patients with new contrast-enhancing lesions., *Magn. Reson. Imaging.* 24 (2006) 1131–1142. doi:10.1016/j.mri.2006.07.008.
- [51] P.A. Hein, C.J. Eskey, J.F. Dunn, E.B. Hug, Diffusion-weighted imaging in the follow-up of treated high-grade gliomas: tumor recurrence versus radiation injury., *AJNR Am. J. Neuroradiol.* 25 (2004) 201–209.
- [52] J.-L. Xu, Y.-L. Li, J.-M. Lian, S. Dou, F.-S. Yan, H. Wu, et al., Distinction between postoperative recurrent glioma and radiation injury using MR diffusion tensor imaging., *Neuroradiology.* 52 (2010) 1193–1199. doi:10.1007/s00234-010-0731-4.
- [53] Q.-S. Zeng, C.-F. Li, H. Liu, J.-H. Zhen, D.-C. Feng, Distinction between recurrent glioma and radiation injury using magnetic resonance spectroscopy in combination with diffusion-weighted imaging., *Int. J. Radiat. Oncol. Biol. Phys.* 68 (2007) 151–158. doi:10.1016/j.ijrobp.2006.12.001.
- [54] E. Matsusue, J.R. Fink, J.K. Rockhill, T. Ogawa, K.R. Maravilla, Distinction between glioma progression and post-radiation change by combined physiologic MR imaging., *Neuroradiology.* 52 (2010) 297–306. doi:10.1007/s00234-009-0613-9.
- [55] A.J. Prager, N. Martinez, K. Beal, A. Omuro, Z. Zhang, R.J. Young, Diffusion and perfusion MRI to differentiate treatment-related changes including pseudoprogression from recurrent tumors in high-grade gliomas with histopathologic evidence., *AJNR Am. J. Neuroradiol.* 36 (2015) 877–885. doi:10.3174/ajnr.A4218.
- [56] A. Jena, S. Taneja, A. Jha, N.K. Damesha, P. Negi, G.K. Jadhav, et al., Multiparametric Evaluation in Differentiating Glioma Recurrence from Treatment-Induced Necrosis Using Simultaneous (18)F-FDG-PET/MRI: A Single-Institution Retrospective Study., *AJNR*

- Am. J. Neuroradiol. 38 (2017) 899–907. doi:10.3174/ajnr.A5124.
- [57] G.A. Alexiou, A. Zikou, S. Tsiouris, A. Goussia, P. Kosta, A. Papadopoulos, et al., Comparison of diffusion tensor, dynamic susceptibility contrast MRI and (99m)Tc-Tetrofosmin brain SPECT for the detection of recurrent high-grade glioma., *Magn. Reson. Imaging*. 32 (2014) 854–859. doi:10.1016/j.mri.2014.04.013.
- [58] X.-F. Wu, X. Liang, X.-C. Wang, J.-B. Qin, L. Zhang, Y. Tan, et al., Differentiating high-grade glioma recurrence from pseudoprogression: Comparing diffusion kurtosis imaging and diffusion tensor imaging., *Eur. J. Radiol*. 135 (2020) 109445. doi:10.1016/j.ejrad.2020.109445.
- [59] J.Y. Kim, J.E. Park, Y. Jo, W.H. Shim, S.J. Nam, J.H. Kim, et al., Incorporating diffusion- and perfusion-weighted MRI into a radiomics model improves diagnostic performance for pseudoprogression in glioblastoma patients., *Neuro. Oncol*. 21 (2019) 404–414. doi:10.1093/neuonc/noy133.
- [60] M. Rogosnitzky, S. Branch, Gadolinium-based contrast agent toxicity: a review of known and proposed mechanisms., *Biometals*. 29 (2016) 365–376. doi:10.1007/s10534-016-9931-7.
- [61] M. Han, B. Yang, B. Fernandez, M. Lafontaine, P. Alcaide-Leon, A. Jakary, et al., Simultaneous multi-slice spin- and gradient-echo dynamic susceptibility-contrast perfusion-weighted MRI of gliomas., *NMR Biomed*. 34 (2021) e4399. doi:10.1002/nbm.4399.
- [62] J.L. Boxerman, D.E. Prah, E.S. Paulson, J.T. Machan, D. Bedekar, K.M. Schmainda, The Role of preload and leakage correction in gadolinium-based cerebral blood volume estimation determined by comparison with MION as a criterion standard., *AJNR Am. J.*

- Neuroradiol. 33 (2012) 1081–1087. doi:10.3174/ajnr.A2934.
- [63] L.S. Hu, L.C. Baxter, D.S. Pinnaduwage, T.L. Paine, J.P. Karis, B.G. Feuerstein, et al., Optimized preload leakage-correction methods to improve the diagnostic accuracy of dynamic susceptibility-weighted contrast-enhanced perfusion MR imaging in posttreatment gliomas., *AJNR Am. J. Neuroradiol.* 31 (2010) 40–48. doi:10.3174/ajnr.A1787.
- [64] J.L. Boxerman, K.M. Schmainda, R.M. Weisskoff, Relative cerebral blood volume maps corrected for contrast agent extravasation significantly correlate with glioma tumor grade, whereas uncorrected maps do not., *AJNR Am. J. Neuroradiol.* 27 (2006) 859–867.
- [65] M. Maeda, S. Itoh, H. Kimura, T. Iwasaki, N. Hayashi, K. Yamamoto, et al., Tumor vascularity in the brain: evaluation with dynamic susceptibility-contrast MR imaging., *Radiology.* 189 (1993) 233–238. doi:10.1148/radiology.189.1.8372199.
- [66] J. Vöglein, J. Tüttenberg, M. Weimer, L. Gerigk, H.-U. Kauczor, M. Essig, et al., Treatment monitoring in gliomas: comparison of dynamic susceptibility-weighted contrast-enhanced and spectroscopic MRI techniques for identifying treatment failure., *Invest Radiol.* 46 (2011) 390–400. doi:10.1097/RLI.0b013e31820e1511.
- [67] E. Essock-Burns, J.M. Lupo, S. Cha, M.-Y. Polley, N.A. Butowski, S.M. Chang, et al., Assessment of perfusion MRI-derived parameters in evaluating and predicting response to antiangiogenic therapy in patients with newly diagnosed glioblastoma., *Neuro. Oncol.* 13 (2011) 119–131. doi:10.1093/neuonc/noq143.
- [68] E. Ozhinsky, D.B. Vigneron, S.J. Nelson, Improved spatial coverage for brain 3D PRESS MRSI by automatic placement of outer-volume suppression saturation bands., *J. Magn. Reson. Imaging.* 33 (2011) 792–802. doi:10.1002/jmri.22507.

- [69] V. Michel, Z. Yuan, S. Ramsubir, M. Bakovic, Choline transport for phospholipid synthesis., *Exp. Biol. Med.* 231 (2006) 490–504. doi:10.1177/153537020623100503.
- [70] J. Urenjak, S.R. Williams, D.G. Gadian, M. Noble, Proton nuclear magnetic resonance spectroscopy unambiguously identifies different neural cell types., *J. Neurosci.* 13 (1993) 981–989.
- [71] D. Galanaud, O. Chinot, F. Nicoli, S. Confort-Gouny, Y. Le Fur, M. Barrie-Attarian, et al., Use of proton magnetic resonance spectroscopy of the brain to differentiate gliomatosis cerebri from low-grade glioma., *J. Neurosurg.* 98 (2003) 269–276. doi:10.3171/jns.2003.98.2.0269.
- [72] X. Li, Y. Lu, A. Pirzkall, T. McKnight, S.J. Nelson, Analysis of the spatial characteristics of metabolic abnormalities in newly diagnosed glioma patients., *J. Magn. Reson. Imaging.* 16 (2002) 229–237. doi:10.1002/jmri.10147.
- [73] J.R. Moffett, B. Ross, P. Arun, C.N. Madhavarao, A.M.A. Namboodiri, N-Acetylaspartate in the CNS: from neurodiagnostics to neurobiology., *Prog. Neurobiol.* 81 (2007) 89–131. doi:10.1016/j.pneurobio.2006.12.003.
- [74] W.G. Negendank, R. Sauter, T.R. Brown, J.L. Evelhoch, A. Falini, E.D. Gotsis, et al., Proton magnetic resonance spectroscopy in patients with glial tumors: a multicenter study., *J. Neurosurg.* 84 (1996) 449–458. doi:10.3171/jns.1996.84.3.0449.
- [75] J.A. Osorio, E. Ozturk-Isik, D. Xu, S. Cha, S. Chang, M.S. Berger, et al., 3D 1H MRSI of brain tumors at 3.0 Tesla using an eight-channel phased-array head coil., *J. Magn. Reson. Imaging.* 26 (2007) 23–30. doi:10.1002/jmri.20970.
- [76] A. Seeger, C. Braun, M. Skardelly, F. Paulsen, J. Schittenhelm, U. Ernemann, et al., Comparison of three different MR perfusion techniques and MR spectroscopy for

- multiparametric assessment in distinguishing recurrent high-grade gliomas from stable disease., *Acad. Radiol.* 20 (2013) 1557–1565. doi:10.1016/j.acra.2013.09.003.
- [77] A.E. Elias, R.C. Carlos, E.A. Smith, D. Frechtling, B. George, P. Maly, et al., MR spectroscopy using normalized and non-normalized metabolite ratios for differentiating recurrent brain tumor from radiation injury., *Acad. Radiol.* 18 (2011) 1101–1108. doi:10.1016/j.acra.2011.05.006.
- [78] K. Ando, R. Ishikura, Y. Nagami, T. Morikawa, Y. Takada, J. Ikeda, et al., [Usefulness of Cho/Cr ratio in proton MR spectroscopy for differentiating residual/recurrent glioma from non-neoplastic lesions]., *Nihon Igaku Hoshasen Gakkai Zasshi.* 64 (2004) 121–126.
- [79] E.A. Smith, R.C. Carlos, L.R. Junck, C.I. Tsien, A. Elias, P.C. Sundgren, Developing a clinical decision model: MR spectroscopy to differentiate between recurrent tumor and radiation change in patients with new contrast-enhancing lesions., *AJR Am. J. Roentgenol.* 192 (2009) W45-52. doi:10.2214/AJR.07.3934.
- [80] P. Weybright, P.C. Sundgren, P. Maly, D.G. Hassan, B. Nan, S. Rohrer, et al., Differentiation between brain tumor recurrence and radiation injury using MR spectroscopy., *AJR Am. J. Roentgenol.* 185 (2005) 1471–1476. doi:10.2214/AJR.04.0933.
- [81] M. Plotkin, J. Eisenacher, H. Bruhn, R. Wurm, R. Michel, F. Stockhammer, et al., 123I-IMT SPECT and 1H MR-spectroscopy at 3.0 T in the differential diagnosis of recurrent or residual gliomas: a comparative study., *J. Neurooncol.* 70 (2004) 49–58.
- [82] M.-T. Chuang, Y.-S. Liu, Y.-S. Tsai, Y.-C. Chen, C.-K. Wang, Differentiating Radiation-Induced Necrosis from Recurrent Brain Tumor Using MR Perfusion and Spectroscopy: A Meta-Analysis., *PLoS One.* 11 (2016) e0141438. doi:10.1371/journal.pone.0141438.

- [83] B.R.J. van Dijken, P.J. van Laar, G.A. Holtman, A. van der Hoorn, Diagnostic accuracy of magnetic resonance imaging techniques for treatment response evaluation in patients with high-grade glioma, a systematic review and meta-analysis., *Eur. Radiol.* 27 (2017) 4129–4144. doi:10.1007/s00330-017-4789-9.
- [84] C. Choi, S.K. Ganji, R.J. DeBerardinis, K.J. Hatanpaa, D. Rakheja, Z. Kovacs, et al., 2-hydroxyglutarate detection by magnetic resonance spectroscopy in IDH-mutated patients with gliomas., *Nat. Med.* 18 (2012) 624–629. doi:10.1038/nm.2682.
- [85] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*, 2nd ed., Springer, New York, NY, 2009.
- [86] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning (Adaptive Computation and Machine Learning series)*, 2nd ed., The MIT Press, 2016.
- [87] P. Diggle, P. Heagerty, K.-Y. Liang, S. Zeger, *Analysis of Longitudinal Data (Oxford Statistical Science) (Oxford Statistical Science Series)*, n.d.
- [88] CS231n Convolutional Neural Networks for Visual Recognition, (n.d.).  
<https://cs231n.github.io/neural-networks-1/> (accessed December 10, 2020).
- [89] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, *Commun ACM.* 60 (2012) 84–90. doi:10.1145/3065386.
- [90] V. Dumoulin, F. Visin, A guide to convolution arithmetic for deep learning, *ArXiv.* (2016).
- [91] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE.* 86 (1998) 2278–2324. doi:10.1109/5.726791.
- [92] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image

- Recognition, ArXiv. (2014).
- [93] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016: pp. 770–778. doi:10.1109/CVPR.2016.90.
- [94] Yuehao Pan, Weimin Huang, Zhiping Lin, Wanzheng Zhu, Jiayin Zhou, J. Wong, et al., Brain tumor grading based on Neural Networks and Convolutional Neural Networks., Conf. Proc. IEEE Eng. Med. Biol. Soc. 2015 (2015) 699–702. doi:10.1109/EMBC.2015.7318458.
- [95] B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imaging. 34 (2015) 1993–2024. doi:10.1109/TMI.2014.2377694.
- [96] C.G. Bangalore Yogananda, B.R. Shah, M. Vejdani-Jahromi, S.S. Nalawade, G.K. Murugesan, F.F. Yu, et al., A novel fully automated MRI-based deep-learning method for classification of IDH mutation status in brain gliomas., Neuro. Oncol. 22 (2020) 402–411. doi:10.1093/neuonc/noz199.
- [97] D.N. Louis, A. Perry, G. Reifenberger, A. von Deimling, D. Figarella-Branger, W.K. Cavenee, et al., The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary., Acta Neuropathol. 131 (2016) 803–820. doi:10.1007/s00401-016-1545-1.
- [98] D.J. Brat, K. Aldape, H. Colman, D. Figarella-Branger, G.N. Fuller, C. Giannini, et al., cIMPACT-NOW update 5: recommended grading criteria and terminologies for IDH-mutant astrocytomas., Acta Neuropathol. 139 (2020) 603–608. doi:10.1007/s00401-020-02127-9.



- [99] G.A. Yeaney, D.J. Brat, What Every Neuropathologist Needs to Know: Update on cIMPACT-NOW., *J. Neuropathol. Exp. Neurol.* 78 (2019) 294–296.  
doi:10.1093/jnen/nlz012.
- [100] G. Cairncross, M. Wang, E. Shaw, R. Jenkins, D. Brachman, J. Buckner, et al., Phase III trial of chemoradiotherapy for anaplastic oligodendroglioma: long-term results of RTOG 9402., *J. Clin. Oncol.* 31 (2013) 337–343. doi:10.1200/JCO.2012.43.2674.
- [101] S. Nobusawa, T. Watanabe, P. Kleihues, H. Ohgaki, IDH1 mutations as molecular signature and predictive factor of secondary glioblastomas., *Clin. Cancer Res.* 15 (2009) 6002–6007. doi:10.1158/1078-0432.CCR-09-0715.
- [102] J.E. Eckel-Passow, D.H. Lachance, A.M. Molinaro, K.M. Walsh, P.A. Decker, H. Sicotte, et al., Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors., *N. Engl. J. Med.* 372 (2015) 2499–2508. doi:10.1056/NEJMoa1407279.
- [103] S.H. Patel, L.M. Poisson, D.J. Brat, Y. Zhou, L. Cooper, M. Snuderl, et al., T2-FLAIR Mismatch, an Imaging Biomarker for IDH and 1p/19q Status in Lower-grade Gliomas: A TCGA/TCIA Project., *Clin. Cancer Res.* 23 (2017) 6078–6085. doi:10.1158/1078-0432.CCR-17-0560.
- [104] M.P.G. Broen, M. Smits, M.M.J. Wijnenga, H.J. Dubbink, M.H.M.E. Anten, O.E.M.G. Schijns, et al., The T2-FLAIR mismatch sign as an imaging marker for non-enhancing IDH-mutant, 1p/19q-intact lower-grade glioma: a validation study., *Neuro. Oncol.* 20 (2018) 1393–1399. doi:10.1093/neuonc/noy048.
- [105] M. Diehn, C. Nardini, D.S. Wang, S. McGovern, M. Jayaraman, Y. Liang, et al., Identification of noninvasive imaging surrogates for brain tumor gene-expression modules., *Proc. Natl. Acad. Sci. USA.* 105 (2008) 5213–5218.

doi:10.1073/pnas.0801279105.

- [106] H. Ding, Y. Huang, Z. Li, S. Li, Q. Chen, C. Xie, et al., Prediction of IDH Status Through MRI Features and Enlightened Reflection on the Delineation of Target Volume in Low-Grade Gliomas., *Technol Cancer Res Treat.* 18 (2019) 1533033819877167. doi:10.1177/1533033819877167.
- [107] R. Jain, D.R. Johnson, S.H. Patel, M. Castillo, M. Smits, M.J. van den Bent, et al., “Real world” use of a highly reliable imaging sign: “T2-FLAIR mismatch” for identification of IDH mutant astrocytomas., *Neuro. Oncol.* 22 (2020) 936–943. doi:10.1093/neuonc/noaa041.
- [108] L. Chen, Z. Voronovich, K. Clark, I. Hands, J. Mannas, M. Walsh, et al., Predicting the likelihood of an isocitrate dehydrogenase 1 or 2 mutation in diagnoses of infiltrative glioma., *Neuro. Oncol.* 16 (2014) 1478–1483. doi:10.1093/neuonc/nou097.
- [109] Z. Li, Y. Wang, J. Yu, Y. Guo, W. Cao, Deep Learning based Radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma., *Sci. Rep.* 7 (2017) 5467. doi:10.1038/s41598-017-05848-2.
- [110] P. Chang, J. Grinband, B.D. Weinberg, M. Bardis, M. Khy, G. Cadena, et al., Deep-Learning Convolutional Neural Networks Accurately Classify Genetic Mutations in Gliomas., *AJNR Am. J. Neuroradiol.* 39 (2018) 1201–1207. doi:10.3174/ajnr.A5667.
- [111] C.-F. Lu, F.-T. Hsu, K.L.-C. Hsieh, Y.-C.J. Kao, S.-J. Cheng, J.B.-K. Hsu, et al., Machine Learning-Based Radiomics for Molecular Subtyping of Gliomas., *Clin. Cancer Res.* 24 (2018) 4429–4436. doi:10.1158/1078-0432.CCR-17-3445.
- [112] K. Chang, H.X. Bai, H. Zhou, C. Su, W.L. Bi, E. Agbodza, et al., Residual Convolutional Neural Network for the Determination of IDH Status in Low- and High-Grade Gliomas

- from MR Imaging., *Clin. Cancer Res.* 24 (2018) 1073–1081. doi:10.1158/1078-0432.CCR-17-2236.
- [113] Y.S. Choi, S. Bae, J.H. Chang, S.-G. Kang, S.H. Kim, J. Kim, et al., Fully automated hybrid approach to predict the IDH mutation status of gliomas via deep learning and radiomics., *Neuro. Oncol.* (2020). doi:10.1093/neuonc/noaa177.
- [114] Y. Cui, L. Ma, X. Chen, Z. Zhang, H. Jiang, S. Lin, Lower apparent diffusion coefficients indicate distinct prognosis in low-grade and high-grade glioma., *J. Neurooncol.* 119 (2014) 377–385. doi:10.1007/s11060-014-1490-6.
- [115] C.M. Heaphy, R.F. de Wilde, Y. Jiao, A.P. Klein, B.H. Edil, C. Shi, et al., Altered telomeres in tumors with ATRX and DAXX mutations., *Science.* 333 (2011) 425. doi:10.1126/science.1207313.
- [116] J.M. Duarte-Carvajalino, G. Sapiro, N. Harel, C. Lenglet, A Framework for Linear and Non-Linear Registration of Diffusion-Weighted MRIs Using Angular Interpolation., *Front. Neurosci.* 7 (2013) 41. doi:10.3389/fnins.2013.00041.
- [117] M. Jenkinson, P. Bannister, M. Brady, S. Smith, Improved optimization for the robust and accurate linear registration and motion correction of brain images., *Neuroimage.* 17 (2002) 825–841. doi:10.1016/s1053-8119(02)91132-8.
- [118] H.J. Johnson, G. Harris, K. Williams, BRAINSFit: Mutual Information Registrations of Whole-Brain 3D Images, Using the Insight Toolkit, (2007). <http://hdl.handle.net/1926/1291> (accessed May 1, 2019).
- [119] R. Kikinis, S.D. Pieper, K.G. Vosburgh, 3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support, in: F.A. Jolesz (Ed.), *Intraoperative Imaging and Image-Guided Therapy*, Springer New York, New York, NY, 2014: pp.

- 277–289. doi:10.1007/978-1-4614-7657-3\_19.
- [120] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, et al., Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features., *Sci. Data*. 4 (2017) 170117. doi:10.1038/sdata.2017.117.
- [121] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, et al., Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge, *ArXiv*. (2018).
- [122] B. Zhang, K. Chang, S. Ramkissoon, S. Tanguturi, W.L. Bi, D.A. Reardon, et al., Multimodal MRI features predict isocitrate dehydrogenase genotype in high-grade gliomas., *Neuro. Oncol.* 19 (2017) 109–117. doi:10.1093/neuonc/now121.
- [123] H. Sajedi, N. Pardakhti, Age prediction based on brain MRI image: A survey., *J Med Syst.* 43 (2019) 279. doi:10.1007/s10916-019-1401-7.
- [124] Pedregosa, F. Varoquaux, G. Gramfort, A. Michel, V. Thirion, B. Grisel, et al., *Scikit-learn: Machine Learning in Python*, *The Journal of Machine Learning Research*. (2011).
- [125] L.N. Smith, A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay, *ArXiv*. (2018).
- [126] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, *ArXiv*. (2014).
- [127] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity Checks for Saliency Maps, *ArXiv*. (2018).
- [128] J.H. Cole, R.P.K. Poudel, D. Tsagkrasoulis, M.W.A. Caan, C. Steves, T.D. Spector, et al., Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker., *Neuroimage*. 163 (2017) 115–124. doi:10.1016/j.neuroimage.2017.07.059.

- [129] N. Bien, P. Rajpurkar, R.L. Ball, J. Irvin, A. Park, E. Jones, et al., Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet., *PLoS Med.* 15 (2018) e1002699.  
doi:10.1371/journal.pmed.1002699.
- [130] S. Qi, L. Yu, H. Li, Y. Ou, X. Qiu, Y. Ding, et al., Isocitrate dehydrogenase mutation is associated with tumor location and magnetic resonance imaging characteristics in astrocytic neoplasms., *Oncol. Lett.* 7 (2014) 1895–1902. doi:10.3892/ol.2014.2013.
- [131] S.R. van der Voort, F. Incekara, M.M.J. Wijnenga, G. Kapas, M. Gardeniers, J.W. Schouten, et al., Predicting the 1p/19q Codeletion Status of Presumed Low-Grade Glioma with an Externally Validated Machine Learning Algorithm., *Clin. Cancer Res.* 25 (2019) 7455–7462. doi:10.1158/1078-0432.CCR-19-1127.
- [132] F. Valente, C. Costa, A. Silv, Content based retrieval systems in a clinical context, in: O.F. Erondü (Ed.), *Medical Imaging in Clinical Practice*, InTech, 2013.  
doi:10.5772/53027.
- [133] E. Schleimer, J. Pearce, A. Barnecut, W. Rowles, A. Lizee, A. Klein, et al., A Precision Medicine Tool for Patients With Multiple Sclerosis (the Open MS BioScreen): Human-Centered Design and Development., *J. Med. Internet Res.* 22 (2020) e15605.  
doi:10.2196/15605.
- [134] S. Ranjbar, K.W. Singleton, P.R. Jackson, C.R. Rickertsen, S.A. Whitmire, K.R. Clark-Swanson, et al., A deep convolutional neural network for annotation of magnetic resonance imaging sequence type., *J Digit Imaging.* 33 (2020) 439–446.  
doi:10.1007/s10278-019-00282-4.
- [135] R. Pizarro, H.-E. Assemblal, D. De Nigris, C. Elliott, S. Antel, D. Arnold, et al., Using

- deep learning algorithms to automatically identify the brain MRI contrast: implications for managing large databases., *Neuroinformatics*. 17 (2019) 115–130.  
doi:10.1007/s12021-018-9387-8.
- [136] R. Gauriau, C. Bridge, L. Chen, F. Kitamura, N.A. Tenenholtz, J.E. Kirsch, et al., Using DICOM metadata for radiological image series categorization: a feasibility study on large clinical brain MRI datasets., *J Digit Imaging*. 33 (2020) 747–762. doi:10.1007/s10278-019-00308-x.
- [137] R. Bitar, G. Leung, R. Perng, S. Tadros, A.R. Moody, J. Sarrazin, et al., MR pulse sequences: what every radiologist wants to know but is afraid to ask., *Radiographics*. 26 (2006) 513–537. doi:10.1148/rg.262055063.
- [138] D.B. Plewes, The AAPM/RSNA physics tutorial for residents. Contrast mechanisms in spin-echo MR imaging., *Radiographics*. 14 (1994) 1389–404; quiz 1405.  
doi:10.1148/radiographics.14.6.7855348.
- [139] B.A. Jung, M. Weigel, Spin echo magnetic resonance imaging., *J. Magn. Reson. Imaging*. 37 (2013) 805–817. doi:10.1002/jmri.24068.
- [140] D. Mason, Scaramallion, Rhaxton, Mrbean-Bremen, J. Suever, Vanessasaurus, et al., pydicom/pydicom: pydicom 2.1.2, Zenodo. (2020). doi:10.5281/zenodo.1291985.
- [141] L. Breiman, Random Forests, Springer Science and Business Media LLC. (2001).  
doi:10.1023/a:1010933404324.
- [142] L. van der Maaten, G. Hinton, Visualizing Data using t-SNE, *Journal of Machine Learning Research*. (2008).
- [143] B.M. Ellingson, T.F. Cloughesy, A. Lai, P.S. Mischel, P.L. Nghiemphu, S. Lalezari, et al., Graded functional diffusion map-defined characteristics of apparent diffusion

- coefficients predict overall survival in recurrent glioblastoma treated with bevacizumab., *Neuro. Oncol.* 13 (2011) 1151–1161. doi:10.1093/neuonc/nor079.
- [144] B.M. Ellingson, C. Chung, W.B. Pope, J.L. Boxerman, T.J. Kaufmann, Pseudoprogession, radionecrosis, inflammation or true tumor progression? challenges associated with glioblastoma response assessment in an evolving therapeutic landscape., *J. Neurooncol.* 134 (2017) 495–504. doi:10.1007/s11060-017-2375-2.
- [145] B.M. Ellingson, P.Y. Wen, T.F. Cloughesy, Modified criteria for radiographic response assessment in glioblastoma clinical trials., *Neurotherapeutics.* 14 (2017) 307–320. doi:10.1007/s13311-016-0507-6.
- [146] P.Y. Wen, D.R. Macdonald, D.A. Reardon, T.F. Cloughesy, A.G. Sorensen, E. Galanis, et al., Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group., *J. Clin. Oncol.* 28 (2010) 1963–1972. doi:10.1200/JCO.2009.26.3541.
- [147] T. Kazda, M. Bulik, P. Pospisil, R. Lakomy, M. Smrcka, P. Slampa, et al., Advanced MRI increases the diagnostic accuracy of recurrent glioblastoma: Single institution thresholds and validation of MR spectroscopy and diffusion weighted MR imaging., *Neuroimage Clin.* 11 (2016) 316–321. doi:10.1016/j.nicl.2016.02.016.
- [148] J. Narang, R. Jain, A.S. Arbab, T. Mikkelsen, L. Scarpace, M.L. Rosenblum, et al., Differentiating treatment-induced necrosis from recurrent/progressive brain tumor using nonmodel-based semiquantitative indices derived from dynamic contrast-enhanced T1-weighted MR perfusion., *Neuro. Oncol.* 13 (2011) 1037–1046. doi:10.1093/neuonc/nor075.
- [149] R. Jain, J. Narang, L. Schultz, L. Scarpace, S. Saksena, S. Brown, et al., Permeability

- estimates in histopathology-proved treatment-induced necrosis using perfusion CT: can these add to other perfusion parameters in differentiating from recurrent/progressive tumors?, *AJNR Am. J. Neuroradiol.* 32 (2011) 658–663. doi:10.3174/ajnr.A2378.
- [150] J.L. Boxerman, B.M. Ellingson, S. Jeyapalan, H. Elinzano, R.J. Harris, J.M. Rogg, et al., Longitudinal DSC-MRI for Distinguishing Tumor Recurrence From Pseudoprogression in Patients With a High-grade Glioma., *Am. J. Clin. Oncol.* 40 (2017) 228–234. doi:10.1097/COC.000000000000156.
- [151] T.-H. Kim, T.J. Yun, C.-K. Park, T.M. Kim, J.-H. Kim, C.-H. Sohn, et al., Combined use of susceptibility weighted magnetic resonance imaging sequences and dynamic susceptibility contrast perfusion weighted imaging to improve the accuracy of the differential diagnosis of recurrence and radionecrosis in high-grade glioma patients., *Oncotarget.* 8 (2017) 20340–20353. doi:10.18632/oncotarget.13050.
- [152] L.S. Hu, L.C. Baxter, K.A. Smith, B.G. Feuerstein, J.P. Karis, J.M. Eschbacher, et al., Relative cerebral blood volume values to differentiate high-grade glioma recurrence from posttreatment radiation effect: direct correlation between image-guided tissue histopathology and localized dynamic susceptibility-weighted contrast-enhanced perfusion MR imaging measurements., *AJNR Am. J. Neuroradiol.* 30 (2009) 552–558. doi:10.3174/ajnr.A1377.
- [153] J.P. Rock, D. Hearshen, L. Scarpace, D. Croteau, J. Gutierrez, J.L. Fisher, et al., Correlations between magnetic resonance spectroscopy and image-guided histopathology, with special attention to radiation necrosis., *Neurosurgery.* 51 (2002) 912–9; discussion 919. doi:10.1097/00006123-200210000-00010.
- [154] I. Park, A.P. Chen, M.L. Zierhut, E. Ozturk-Isik, D.B. Vigneron, S.J. Nelson,



- Implementation of 3 T lactate-edited 3D 1H MR spectroscopic imaging with flyback echo-planar readout for gliomas patients., *Ann. Biomed. Eng.* 39 (2011) 193–204.  
doi:10.1007/s10439-010-0128-x.
- [155] P.J. Kelly, B.A. Kall, S. Goerss, F. Earnest, Computer-assisted stereotaxic laser resection of intra-axial brain neoplasms., *J. Neurosurg.* 64 (1986) 427–439.  
doi:10.3171/jns.1986.64.3.0427.
- [156] R.M. Weisskoff, J.L. Boxerman, A.G. Sorensen, S.M. Kulke, T.A. Campbell, B.R. Rosen, Simultaneous blood volume and permeability mapping using a single Gd-based contrast injection, in: *Proceedings of the 2nd Annual Meeting of SMRM, ISMRM, San Francisco, CA, 1994*: p. 279.
- [157] J.M. Lupo, Q. Wen, S.M. Chang, S.J. Nelson, Weighted-average model curve preprocessing strategy for quantification of DSC perfusion imaging metrics from image-guided tissue samples in patients with brain tumors, in: *Proc. Intl. Soc. Mag. Reson. Med.* 23 , ISMRM, 2015: p. 4377.
- [158] J.C. Crane, M.P. Olson, S.J. Nelson, SIVIC: Open-Source, Standards-Based Software for DICOM MR Spectroscopy Workflows., *Int J Biomed Imaging.* 2013 (2013) 169526.  
doi:10.1155/2013/169526.
- [159] S.J. Nelson, A.K. Kadambi, I. Park, Y. Li, J. Crane, M. Olson, et al., Association of early changes in 1H MRSI parameters with survival for patients with newly diagnosed glioblastoma receiving a multimodality treatment regimen., *Neuro. Oncol.* 19 (2017) 430–439. doi:10.1093/neuonc/now159.
- [160] Y. Li, J.M. Lupo, R. Parvataneni, K.R. Lamborn, S. Cha, S.M. Chang, et al., Survival analysis in patients with newly diagnosed glioblastoma using pre- and postradiotherapy

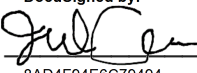
- MR spectroscopic imaging., *Neuro. Oncol.* 15 (2013) 607–617.  
doi:10.1093/neuonc/nos334.
- [161] T.R. McKnight, S.M. Noworolski, D.B. Vigneron, S.J. Nelson, An automated technique for the quantitative assessment of 3D-MRSI data from patients with glioma, *J. Magn. Reson. Imaging.* (2001).
- [162] R.F. Barajas, J.J. Phillips, R. Parvataneni, A. Molinaro, E. Essock-Burns, G. Bourne, et al., Regional variation in histopathologic features of tumor specimens from treatment-naive glioblastoma correlates with anatomic and physiologic MR Imaging., *Neuro. Oncol.* 14 (2012) 942–954. doi:10.1093/neuonc/nos128.
- [163] X. Robin, N. Turck, A. Hainar, Tiberti, Lisacek, Sanchez, et al., pROC, CRAN, 2018.
- [164] M. Kuhn, caret, CRAN, 2019.
- [165] T. Nakajima, T. Kumabe, M. Kanamori, R. Saito, M. Tashiro, M. Watanabe, et al., Differential diagnosis between radiation necrosis and glioma progression using sequential proton magnetic resonance spectroscopy and methionine positron emission tomography., *Neurol Med Chir (Tokyo).* 49 (2009) 394–401. doi:10.2176/nmc.49.394.
- [166] M.M. D’Souza, R. Sharma, A. Jaimini, P. Panwar, S. Saw, P. Kaur, et al., 11C-MET PET/CT and advanced MRI in the evaluation of tumor recurrence in high-grade gliomas., *Clin Nucl Med.* 39 (2014) 791–798. doi:10.1097/RLU.0000000000000532.
- [167] W. Stummer, Mechanisms of tumor-related brain edema., *Neurosurg. Focus.* 22 (2007) E8. doi:10.3171/foc.2007.22.5.9.
- [168] C. Peca, R. Pacelli, A. Elefante, M.L. Del Basso De Caro, P. Vergara, G. Mariniello, et al., Early clinical and neuroradiological worsening after radiotherapy and concomitant temozolomide in patients with glioblastoma: tumour progression or radionecrosis?, *Clin.*

- Neurol. Neurosurg. 111 (2009) 331–334. doi:10.1016/j.clineuro.2008.11.003.
- [169] A. Server, R. Josefsen, B. Kulle, J. Maehlen, T. Schellhorn, Ø. Gadmar, et al., Proton magnetic resonance spectroscopy in the distinction of high-grade cerebral gliomas from single metastatic brain tumors., *Acta Radiol.* 51 (2010) 316–325. doi:10.3109/02841850903482901.
- [170] I.J. Gerard, M. Kersten-Oertel, K. Petrecca, D. Sirhan, J.A. Hall, D.L. Collins, Brain shift in neuronavigation of brain tumors: A review., *Med Image Anal.* 35 (2017) 403–420. doi:10.1016/j.media.2016.08.007.
- [171] N. Zakhari, M.S. Taccone, C. Torres, S. Chakraborty, J. Sinclair, J. Woulfe, et al., Diagnostic Accuracy of Centrally Restricted Diffusion in the Differentiation of Treatment-Related Necrosis from Tumor Recurrence in High-Grade Gliomas., *AJNR Am. J. Neuroradiol.* 39 (2018) 260–264. doi:10.3174/ajnr.A5485.
- [172] P. Alcaide-Leon, J. Cluceru, J.M. Lupo, T.J. Yu, T.L. Luks, T. Tihan, et al., Centrally Reduced Diffusion Sign for Differentiation between Treatment-Related Lesions and Glioma Progression: A Validation Study., *AJNR Am. J. Neuroradiol.* 41 (2020) 2049–2054. doi:10.3174/ajnr.A6843.
- [173] N. Verma, M.C. Cowperthwaite, M.G. Burnett, M.K. Markey, Differentiating tumor recurrence from treatment necrosis: a review of neuro-oncologic imaging strategies., *Neuro. Oncol.* 15 (2013) 515–534. doi:10.1093/neuonc/nos307.

## Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:  
  
8AD4F94E6C79494... Author Signature

3/8/2021  
Date