

5-2021

Imaging Based Prediction of Pathology in Adult Diffuse Glioma with Applications to Therapy and Prognosis

Evan Gates

Follow this and additional works at: https://digitalcommons.library.tmc.edu/utgsbs_dissertations



Part of the [Diagnosis Commons](#), [Medical Pathology Commons](#), [Neoplasms Commons](#), and the [Radiology Commons](#)

Recommended Citation

Gates, Evan, "Imaging Based Prediction of Pathology in Adult Diffuse Glioma with Applications to Therapy and Prognosis" (2021). *The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access)*. 1083.
https://digitalcommons.library.tmc.edu/utgsbs_dissertations/1083


This Dissertation (PhD) is brought to you for free and open access by the The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences at DigitalCommons@TMC. It has been accepted for inclusion in The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digitalcommons@library.tmc.edu.

IMAGING BASED PREDICTION OF PATHOLOGY IN ADULT
DIFFUSE GLIOMA WITH APPLICATIONS TO THERAPY AND
PROGNOSIS

by

Evan Donald Huckins Gates, M.S.

APPROVED:

DocuSigned by:

43AF3DD9F1B94CC...
David T. A. Fuentes, Ph.D., Advisor

DocuSigned by:

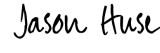
8300BEB031064A1...
Dawid Schellingerhout, M.D., Co-Advisor

DocuSigned by:

72476AA5F011459...
Kristy Brock, Ph.D.

DocuSigned by:

155DCFC2C100401...
John D. Hazle, Ph.D.

DocuSigned by:

F58317857E4A490...
Jason T. Huse, M.D., Ph.D.

APPROVED:

Dean, The University of Texas MD Anderson Cancer Center UTHealth Graduate
School of Biomedical Sciences

IMAGING BASED PREDICTION OF PATHOLOGY IN ADULT
DIFFUSE GLIOMA WITH APPLICATIONS TO THERAPY AND
PROGNOSIS

A
DISSERTATION

Presented to the Faculty of
The University of Texas
M. D. Anderson Cancer Center UTHealth
Graduate School of Biomedical Sciences
in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by
Evan Donald Huckins Gates, M.S.
Houston, Texas

May 2021

©Copyright by Evan Gates, 2021.
All rights reserved.

“To Baker. Your gentle reminders to go outside and take time for myself kept me sane during the pandemic.”

Acknowledgments

I first want to thank my advisory team who guided me on this project: Dr. David Fuentes for showing me how to work smarter and put my best foot forward and Dr. Dawid Schellingerhout for sharing his vision and reminding me what the ultimate goal of our research is. I also thank Dr. Jonathan Lin who laid the ground work for and data collection for this project and Dr. Kristy Brock for her professional mentorship that helped me advance to the next stage of my career.

This work would not have been possible without numerous clinical collaborators from around MD Anderson. I want to thank Dr. Jeffrey Weinberg, Dr. Sujit Prabhu, Dr. Dima Suki, Dr. Gregory Fuller, Dr. Jason Huse, Dr. Kazutaka Fukumura, Dr. Erik Sulman, Dr. Anthony Liu, Dr. Jason Stafford, Dr. John Hazle, Emmanuel Afowowe, Thomas Nguyen, Alicia Ledoux, Riya Nellipallil, and Stephanie Carlon. As well as all the other lab members and medical physics students who have helped me through the graduate school process including Dr. Drew Mitchel, Dr. Jonas Actor, Adrian Celaya, Dr. Megan Jacobsen, Dr. Sara Thrower, Dr. Travis Salzillo, and Dr. Jie Yang.

I am extremely grateful my fellowship from the Gulf Coast Consortia, on the NLM Training Program in Biomedical Informatics and Data Science T15LM007093 which provided stipend support and advanced training. I also acknowledge scholarship support from the American Legion Auxiliary. This work was also enabled by excellent institutional resources: Some data for this work have been obtained through a search of the integrated multidisciplinary Brain and Spine Center Database. The Brain and Spine Center Database was supported in part, by an institutional M. D. Anderson database development grant. Molecular data was provided by The Proactive Program (MD Anderson protocol 2012-0441, PI Dr. John de Groot) The High Performance Computing for research facility at the University of Texas MD Anderson Cancer Center provided computational resources that have contributed to the research results reported in this work.

This acknowledgement would not be complete without a special thanks to my wife, Cami Gates, who always pushes me to work towards my dreams and never settle for less. I am also grateful for my family Austin, Dylan, Bonnie, and Eric who cheered me on unwaveringly as I completed my education.

IMAGING BASED PREDICTION OF PATHOLOGY IN ADULT DIFFUSE GLIOMA WITH APPLICATIONS TO THERAPY AND PROGNOSIS

Evan Donald Huckins Gates, M.S.

Advisory Professor: David Thomas Alfonso Fuentes, Ph.D.

The overall aggressiveness of a glioma is measured by histologic and molecular analysis of tissue samples. However, the well-known spatial heterogeneity in gliomas limits the ability for clinicians to use that information to make spatially specific treatment decisions. Magnetic resonance imaging (MRI) visualizes and assesses the tumor. But, the exact degree to which MRI correlates with the actual underlying tissue characteristics is not known.

In this work, we derive quantitative relationships between imaging and underlying pathology. These relations increase the value of MRI by allowing it to be a better surrogate for underlying pathology and they allow evaluation of the underlying biological heterogeneity via imaging. This provides an approach to answer questions about how tissue heterogeneity can affect prognosis.

We estimated the local pathology within tumors using imaging data and stereotactically precise biopsy samples from an ongoing clinical imaging trial. From this data, we trained a random forest model to reliably predict tumor grade, proliferation, cellularity, and vascularity, representing tumor aggressiveness. We then made voxel-wise predictions to map the tumor heterogeneity and identify high-grade malignancy disease.

Next, we used the previously trained models on a cohort of 1,850 glioma patients who previously underwent surgical resection. High contrast enhancement, proliferation, vascularity, and cellularity were associated with worse prognosis even after controlling for clinical factors. Patients that had substantial reduction in cellularity between preoperative and postoperative imaging (i.e. due to resection) also showed improved survival.

We developed a clinically implementable model for predicting pathology and prognosis after surgery based on imaging. Results from imaging pathology correlations enhance our understanding of disease extent within glioma patients and the relationship between residual estimated pathology and outcome helps refine our knowledge of the interaction of tumor heterogeneity and prognosis.

Contents

Approval Page	i
Title Page	ii
Copyright	iii
Dedication	iv
Acknowledgments	v
Abstract	vi
Table of Contents	vii
List of Figures	xi
List of Tables	xix
Abbreviations and Symbol Names	xxvi
1 Introduction	1
1.1 Dissertation Organization	3
1.2 Background	4
1.2.1 World Health Organization Grading of Gliomas	4
1.2.2 Radiographic and Histologic Heterogeneity of Gliomas	6
1.2.3 Use for Synthetic Pathology Maps in Diagnosis, Surgery, and Ra- diation Therapy	8
1.2.4 The Data Processing Inequality	10
2 Image-Based Prediction of Local Glioma Pathology	12
2.1 Introduction	12
2.2 Methods	14
2.2.1 Image Acquisition	14
2.2.2 Image Registration, Normalization, and Extraction of Values . . .	17
2.2.2.1 Dynamic Contrast Enhanced (DCE) Image Processing . .	17
2.2.2.2 Dynamic Susceptibility Contrast (DSC) Image Processing	18
2.2.2.3 Diffusion Weighted Imaging	19
2.2.2.4 Registration, Normalization, Measurement	20
2.2.3 Biopsy Collection and Pathology Analysis	23

2.2.3.1	Derivation of Normal White Matter Cell Density	25
2.2.4	Predictive Modeling	26
2.2.4.1	Biopsy-Paired Repeated Stratified Cross Validation . . .	27
2.2.4.2	Variable Selection	28
2.2.4.3	Model Types	29
2.3	Results	31
2.3.1	Patient Data, Biopsy Collection, and Pathology Analysis	31
2.3.2	Image Data	32
2.3.3	Variable Selection	32
2.3.3.1	Variable Selection for Predicting Proliferation (Ki67) . .	33
2.3.3.2	Variable Selection for Predicting Cellular Density	33
2.3.3.3	Variable Selection for Predicting Vascularity (ERG) . . .	34
2.3.3.4	Variable Selection for Predicting Local Grade	35
2.3.4	Predictive Modeling	42
2.3.4.1	Predictive Modeling of Proliferation (Ki67)	42
2.3.4.2	Predictive Modeling of Cellular Density	43
2.3.4.3	Predictive Modeling of Vascularity (ERG)	45
2.3.4.4	Predictive Modeling of Local Grade	46
2.3.5	Graphical Synthetic Pathology Maps	47
2.4	Discussion	50
2.4.1	Limitations	52
2.4.2	Future Work	52
3	Neuroimaging Data Curation	54
3.1	Introduction	54
3.2	Methods: Clinical Image Processing and Curation	55
3.2.1	Classification of Image Studies and Series	55
3.2.2	Brain Tumor Segmentation Challenge Data	59
3.2.3	Neuroimaging Pipeline	60
3.2.4	Mode-CSF Normalization	64
3.2.5	Data Quality Assurance with R Shiny	67
3.2.5.1	App Overview	68
3.2.5.2	Data Review Procedure	72
3.3	Results: Processed Data Summary	73
3.3.1	Image Series Classification	73
3.3.2	Image Data Processing	75
3.3.3	Data Quality Assurance	76
3.3.4	BraTS Data Processing and Review	78
3.3.5	Comparison with Reference Values: Intracranial and Tumor Volume	79
3.3.5.1	Intracranial Volume	79
3.3.5.2	Tumor Volume	81
3.4	Discussion	85
3.5	Summary of Curation Effort	88
4	Preoperative Imaging-Pathologic Estimators of Survival	90
4.1	Introduction	90
4.1.1	Summary of Analysis	92

4.1.2	Image Feature Complexity	93
4.2	Methods	94
4.2.1	Clinical Data Sources and Description	94
4.2.1.1	IDH Mutation Data	95
4.2.1.2	Imaging Data	97
4.2.2	Feature Extraction and Mathematical Description of Image Features	98
4.2.3	Survival Analysis	101
4.2.3.1	Concordance Index	101
4.2.3.2	Proportional Hazard Model	101
4.2.3.3	Cross-Validated Binary Thresholds	102
4.3	Results	104
4.3.1	Key Results	104
4.3.2	Patient Cohorts and Clinical Data Summary	105
4.3.3	Clinical Factors, Tumor Volume, and Survival	106
4.3.4	Survival Based on Preoperative Image Features	107
	TC: Maximum T1 Enhanced Intensity	109
	Cell Density: Maximum Cell Density	111
	Shape: Enhancing Tumor Volume	113
	ERG: Maximum ERG Expression	115
	Ki67: Maximum Ki67 Expression	117
	T2: Maximum T2-Weighted Intensity	119
	FLAIR: Pointwise Maximum FLAIR Intensity	121
	T1: Median T1-Weighted Pre-Contrast Intensity	123
	Local Grade: Diameter of Higher-Grade Region	125
	Subset Analysis	127
4.3.5	Summary	128
4.4	Discussion	128
4.4.1	Limitations	130
4.4.2	Future work	131
5	Biologically Based Extent of Resection	132
5.1	Introduction	132
5.1.1	Summary of Analysis	133
5.2	Methods	134
5.2.1	Patient Cohorts and Data	134
5.2.2	Definitions and Mathematical Description of Extent of Resection	135
5.2.3	Curation of Postoperative Data	136
5.2.3.1	Resection Cavity Segmentation Model	137
5.2.4	Survival Analysis	138
5.3	Results	139
5.3.1	Key Results	139
5.3.2	Patient Data Summary	139
5.3.3	Survival Based on Postoperative Image Features	140
	T1: Overall T1-Weighted Pre-Contrast Brightness	141
	FLAIR: Overall Edema Brightness	143
	Shape: Total Residual Volume	144
	TC: Maximum Contrast Enhanced Intensity	146

	T2: Maximum T2-Weighted intensity	148
	ERG: Maximum ERG Expression	150
	Cell Density: Maximum Cell Density	152
	Ki67: Proliferation-Weighted Non-Enhancing Volume	153
	Local Grade: Diameter of Higher-Grade Region	154
	Subset Analysis	157
5.3.4	Clinical Extent of Resection	157
5.3.5	Extent of Resection Based on Image Features	160
	Cell Density: Overall Reduction in Cellularity	161
	T2: Reduction in Median T2-Weighted Brightness	162
	Shape: Fractional Reduction in Non-Enhancing Tumor Volume	164
	ERG: Reduction in Maximum ERG Expression	165
	Ki67: Reduction in Overall Proliferation	167
	FLAIR: Reduction in Median FLAIR Intensity	169
	T1: Reduction in T1 Pre-Contrast Intensity	171
	Subset Analysis	173
5.4	Discussion	174
5.4.1	Limitations	175
6	Discussion	177
6.1	Future Work	180
A	Appendix A	182
A.1	Additional Figures	182
A.2	Reference Tissue and MCSF Normalization Comparison	186
A.3	Stereology and Normal Brain White Matter Cellularity	187
A.4	Survival Analysis on Individual Grade Subsets	188
A.4.1	Preoperative Tumor Volume	188
A.4.2	Preoperative Image Features	190
A.4.3	Postoperative Image Features	196
A.4.4	Extent of Resection	202
A.5	Survival Analysis on Known IDH Mutation Subset	208
A.5.1	Validation on BraTS 2018 Data	209
A.6	Mathematical Methods	209
A.6.1	Proportional Hazard Model and Partial Likelihood Fitting	209
	Bibliography	212
	Vita	236

List of Figures

1.1	Contrast enhanced T1w images of a high-grade glioblastoma. Left: A lobular enhancing tumor component is visible in the posterior portion of the brain. Right: After surgery, the enhancing tumor is completely removed. Therefore, this patient is said to have received a gross total resection	3
1.2	Example of an estimated proliferation map for a non-enhancing WHO III glioma. The blue circle highlights a region of heightened proliferation in an otherwise radiographically homogeneous area. The extra information comes from the inclusion of extra advanced and functional imaging. . . .	8
2.1	Screenshot of the Brainlab cranial navigation software during sample collection. The instrument tip (crosshair) is overlaid on top of a preoperative T1-weighted image that is co-registered to the patient's physical space in the operating room. The coordinates of the instruments are recorded during biopsy.	23
2.2	a) Each patient has 1-3 biopsies (filled) and each each biopsy has a paired virtual (hollow). b) 20% of the biopsies are selected as validation data and the remainder are used for training. Reals are kept in the same set as their paired virtuals.	27
2.3	Metric values for cross-validated prediction of Ki67	43
2.4	Metric values for cross-validated prediction of CD	44
2.5	Metric values for cross-validated prediction of ERG	46
2.6	Metric values for cross-validated prediction of tumor grade	47
2.7	Map of estimated Ki67 in a WHO IV glioblastoma patient alongside a T2-FLAIR image for reference. The map was generated with conventional imaging only and has been smoothed by a 1 mm gaussian kernel.	48
2.8	Map of estimated Ki67 in a WHO II glioma patient alongside a T2-weighted image for reference. The map was generated with conventional and advanced imaging.	48
2.9	Map of estimated Cellular density (CD) in a WHO IV glioblastoma patient alongside a T2-FLAIR image for reference. The map was generated with conventional imaging only.	49
2.10	Map of estimated ETS related gene (ERG) in a WHO IV glioblastoma patient alongside The T1 post-contrast image for reference. The map was generated with conventional imaging only.	49
2.11	Map of estimated tumor grade in a WHO IV glioblastoma patient alongside a T2 FLAIR image for reference. The map was generated with conventional imaging only and is smoothed by a radius 1 median filter. . . .	49
3.1	Image processing pipeline for neuroimaging data.	60

3.2	Registration procedure including bias correction and skull stripping images individually to aid the registration. A dilated mask was used for the cost function as well since it significantly sped up the registration. The other images like T1 and FLAIR were registered in the same way as the T2 is to the T13D (fixed space image).	61
3.3	Sample images for each step of the data processing pipeline. Top left: brain mask, top right: tumor segmentation, bottom left: 4-class tissue segmentation, bottom right: post-processed CSF ROI.	61
3.4	Example of mode-csf normlization. Left: the regions used for normalization are the normal brain (red) and CSF (cyan). Voxels within 1% of the modal intensity are colored extra bright for illustrative purposes. Right: rescaled density plots of the raw intensity values showing the intensity distributions for the normal brain and CSF. The vertical bars show the range of 1% around the modal intensity corresponding to the bright voxels in the ROIs. The mode-csf normalization maps the intensities such that locations of the peaks are 0 and 1 respectively.	65
3.5	Example of K-means clustering with Atropos. The T2w image with ventricles highlighted for reference is shown on the left. After clustering (middle), the yellow class contains mostly CSF. After erosion, the CSF class is broken into several candidate ROIs in the various CSF pools (right). The ROI closest to the brain center is selected.	66
3.6	Shiny app dashboard for data review	68
3.7	Sample CLARA segmentation showing the three-class output for two WHO IV glioblastoma patients. T1 post contrast and FLAIR images are shown with and without segmentation for reference. The segmented regions are edema (yellow), non-enhancing tumor core and necrosis (green), and enhancing tumor (blue). Clinically, the right-hand case has enhancement pattern “both.”	72
3.8	The three most common locations for CSF ROI were the lateral ventricles (left), quadrigeminal cistern (middle), or superior to the brain (right). All give reasonable CSF intensity statistics.	76
3.9	Examples of FLAIR images and ground truth tumor segmentations included in the 2018 Brain Tumor Segmentation Challenge. Left: Brats_2013_0_1. Right: Brats18_2013_6_1. In both cases, the image field of view is so short that the segmentation is partially outside the brain volume. Both of these were caught by data review.	79
3.10	Distributions of preoperative intracranial volume for historical cases. The dashed lines are normal densities from [1], reproduced in Table 3.14. . . .	80
3.11	Comparison of tumor volumes segmented by CLARA and volumes measured previously by neurosurgery. T2 FLAIR volume included the entire FLAIR lesion or total visible tumor volume. Enhancing volume includes both enhancing tissue and tumor necrosis. T1 hypointense volume includes tumor core outside of enhancing volume. The dashed lines indicate agreement.	83
3.12	Preoperative T2 (top) and enhancing tumor volumes (bottom) for historical cases compared to reference values. Each plot shows values for a different level of segmentation quality and the dashed line indicates agreement. “Poor” data is acceptable with minor errors.	84

- 4.1 Flowchart for patient selection in the historical data. Ambiguous imaging time means the imaging and surgery were on the same day. A complete study includes at least one of each: T1-weighted pre-contrast, T1-weighted post-contrast, T2 weighted, and FLAIR images. 97
- 4.2 Illustration of Intensity At Volume features. On the density plot of voxel values the total area of the density plot is equal to the region volume being measured over. The Intensity At Volume feature for a volume V' is the intensity such that the area in the upper tail (red) has area corresponding to the volume V' 100
- 4.3 Survival curves with an optimized cutoff based on preoperative tumor volume. Left: survival stratified by T1 enhancing volume Right: Survival stratified by total T2 visible tumor volume. p-values from log-rank test. . 107
- 4.4 Visualization of the correlation matrix between the features with largest hazard ratios, Table 4.9. Individual Pearson correlations are listed as percentages (i.e. 85 instead of 0.85). The mutual correlations are reduced since we only select features which are correlated less than 0.8 with enhancing or total tumor volume. 109
- 4.5 Best feature for stratifying survival for WHO grades II, III, and IV cases for TC image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. 111
- 4.6 Best feature for stratifying survival for WHO grades II, III, and IV cases using the estimated cell density map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. 113
- 4.7 Best shape feature for stratifying survival for WHO grades II, III, and IV cases. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. 115
- 4.8 Best feature for stratifying survival for WHO grades II, III, and IV cases for the estimated ERG map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. . . 117

- 4.9 Best feature for stratifying survival for WHO grades II, III, and IV cases for the estimated Ki67 map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. . . 119
- 4.10 Best feature for stratifying survival for WHO grades II, III, and IV cases for T2 image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. 121
- 4.11 Best feature for stratifying survival for WHO grades II, III, and IV cases for FLAIR image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. 123
- 4.12 Best feature for stratifying survival for WHO grades II, III, and IV cases for T1 image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. 125
- 4.13 Best feature for stratifying survival for WHO grades II, III, and IV cases for the estimated grade map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. . . 127
- 5.1 Estimated Ki67 maps on preoperative and postoperative imaging overlaid on a T1w contrast enhanced image. The high Ki67 areas on the preoperative image have been removed by surgery and a resection cavity is seen. We hypothesized that surgical removal of high-proliferative activity areas would lead to improved survival. The images shown are of the same patient at roughly the same axial location but appear differently due to slice orientation. 136
- 5.2 T2-weighted postoperative image and illustration of residual disease segmentation. CLARA (top right) provides a segmentation of residual tumor that also falsely segments the resection cavity as disease. A custom-made U-net identifies the postoperative cavity (bottom right) and the cavity is subtracted from the tumor segmentation to more accurately label the residual volume (bottom right). Red: non-enhancing tumor, yellow: enhancing tumor, green: edema, blue: resection cavity. 138

- 5.3 Best postoperativefeature for stratifying survival for WHO grades II, III, and IV cases for T1 image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. . . 142
- 5.4 Best postoperativefeature for stratifying survival for WHO grades II, III, and IV cases for FLAIR image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. . . 144
- 5.5 Best postoperativeshape feature for stratifying survival for WHO grades II, III, and IV. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. 146
- 5.6 Best postoperativefeature for stratifying survival for WHO grades II, III, and IV cases for TC image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. . . 148
- 5.7 Best postoperativefeature for stratifying survival for WHO grades II, III, and IV cases for T2 image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. . . 150
- 5.8 Best postoperativefeature for stratifying survival for WHO grades II, III, and IV cases for ERG map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. . . 151
- 5.9 Best postoperativefeature for stratifying survival for WHO grades II, III, and IV cases for CD map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. . . 153

- 5.10 Best postoperativefeature for stratifying survival for WHO grades II, III, and IV cases for Ki67 map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. . . 154
- 5.11 Best postoperativefeature for stratifying survival for WHO grades II, III, and IV cases for grade map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Note: the survival difference from this feature was non-significant after multiple comparison correction. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. 156
- 5.12 Survival curves for WHO IV cases based on extent of resection for T1 enhancing volume. Left: using automatically segmented tumor volumes. Right: using reference values. The cutoffs are selected to optimize the hazard ratio between groups. (p-value from log-rank test) 159
- 5.13 Survival curves based on T2 FLAIR resection for WHO IV cases with complete resection of enhancing volume. Left: using automatic volume measurements. Right: using reference values. Interestingly, the literature value is not reproduced. Note that the patient populations in each plot are not necessarily the same. 160
- 5.14 Best extent of resectionfeature for stratifying survival for WHO grades II, III, and IV cases for CD map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. . . 162
- 5.15 Best extent of resectionfeature for stratifying survival for WHO grades II, III, and IV cases for T2 image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. . . 164
- 5.16 Best extent of resectionshape feature for stratifying survival for all WHO grades (II III IV). Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line. 165

5.17	Best <u>extent of resection</u> feature for stratifying survival for all WHO grades (II III IV) for ERG map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.	167
5.18	Best <u>extent of resection</u> feature for stratifying survival for all WHO grades (II III IV) for Ki67 map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.	169
5.19	Best <u>extent of resection</u> feature for stratifying survival for all WHO grades (II III IV) cases for FLAIR image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.	171
5.20	Best <u>extent of resection</u> feature for stratifying survival for all WHO grades (II III IV) for T1 image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.	173
A.1	Average predicted and observed Ki67 values	183
A.2	Average predicted and observed ERG values	184
A.3	Average predicted and observed CD values	185
A.4	Modeling results for Ki67 using the original NLM normalized and new brain-CSF mode normalized data. Top: R^2 and RMSE values for 500 rounds of 5-fold cross-validation. The conventional model with MCSF variables performs better than the old conventional model but not as well as the advanced variable model. The predicted and observed values for the three best models are shown on the bottom. The conventional model still has a high Ki67 sample that it struggles with, limiting the R^2 value.	187
A.5	Best features for stratifying survival for WHO grade II cases for each raw image type (T1, TC, T2, FLAIR)	191
A.6	Best features for stratifying survival for WHO grade III cases for each raw image type (T1, TC, T2, FLAIR)	192
A.7	Best features for stratifying survival for WHO grade IV cases for each raw image type (T1, TC, T2, FLAIR)	193
A.8	Best features for stratifying survival for WHO grade II cases for each pathology map estimate (Ki67, CD, ERG)	194
A.9	Best features for stratifying survival for WHO grade III cases for each pathology map estimate (Ki67, CD, ERG)	194

A.10	Best features for stratifying survival for WHO grade IV cases for each pathology map estimate (Ki67, CD, ERG)	195
A.11	Best extent of resection (EOR) measures for stratifying survival for WHO grade II cases for each raw image type (T1, TC, T2, FLAIR)	197
A.12	Best extent of resection (EOR) measures for stratifying survival for WHO grade III cases for each raw image type (T1, TC, T2, FLAIR)	198
A.13	Best extent of resection (EOR) measures for stratifying survival for WHO grade IV cases for each raw image type (T1, TC, T2, FLAIR)	199
A.14	Best extent of resection (EOR) measures for stratifying survival for WHO grade II cases for each pathology map estimate (Ki67, CD, ERG)	200
A.15	Best extent of resection (EOR) measures for stratifying survival for WHO grade III cases for each pathology map estimate (Ki67, CD, ERG)	200
A.16	Best extent of resection (EOR) measures for stratifying survival for WHO grade IV cases for each pathology map estimate (Ki67, CD, ERG)	201
A.17	Best extent of resection (EOR) measures for stratifying survival for WHO grade II cases for each raw image type (T1, TC, T2, FLAIR)	203
A.18	Best extent of resection (EOR) measures for stratifying survival for WHO grade III cases for each raw image type (T1, TC, T2, FLAIR)	204
A.19	Best extent of resection (EOR) measures for stratifying survival for WHO grade IV cases for each raw image type (T1, TC, T2, FLAIR)	205
A.20	Best extent of resection (EOR) measures for stratifying survival for WHO grade II cases for each pathology map estimate (Ki67, CD, ERG)	206
A.21	Best extent of resection (EOR) measures for stratifying survival for WHO grade III cases for each pathology map estimate (Ki67, CD, ERG)	206
A.22	Best extent of resection (EOR) measures for stratifying survival for WHO grade IV cases for each pathology map estimate (Ki67, CD, ERG)	207

List of Tables

2.1	Anatomic image acquisition parameters. *One 3D T1C image had voxel size 0.5469 x 0.5429 x 1.8 mm. FSE: Fast Spin Echo, FGRE = Fast Gradient Recalled Echo, ETL: Echo Train Length, MEMP: Multi Echo Multi Planar	15
2.2	Functional Image acquisition parameters. SE: Spin Echo, GE: Gradient Echo, EPI: Echo Planar Image, SPGR: Spoiled Gradient Recalled	16
2.3	List of image measurements collected for each biopsy.	22
2.4	Predictive model types tested for predicting pathology	30
2.5	Sample grades II - IV listed by the final clinical grade of the tumor each sample was collected from. Grade 0 indicates non-tumor.	32
2.6	Number of samples and mean \pm standard deviation for Ki67, cellular density, and ERG for each sample grade. All three roughly increase with increasing sample grade. Grade 0 indicates non-tumor.	32
2.7	Description of the variable subsets identified through variable selection procedures.	33
2.8	p-values for variables significantly associated with Ki67 in all data and the fraction of folds (frequency) where each correlation is significant among training data	34
2.9	Fraction of times each variable was assigned a particular rank relative to other variables in its class for predicting Ki67	35
2.10	p-values for variables significantly associated with CD in all data and the fraction of folds (frequency) where each correlation is significant among training data	36
2.11	Fraction of times each variable was assigned a particular rank relative to other variables in its class for predicting CD	37
2.12	p-values for variables significantly associated with ERG in all data and the fraction of folds (frequency) where each correlation is significant among training data	38
2.13	Fraction of times each variable was assigned a particular rank relative to other variables in its class for predicting ERG	39
2.14	p-values for variables significantly associated with grade in all data and the fraction of folds (frequency) where each correlation is significant among training data	40
2.15	Fraction of times each variable was assigned a particular rank relative to other variables in its class for predicting grade	41
2.16	Panel A: R^2 and Panel B: RMSE values for predicting Ki67, values are mean \pm standard deviation. lm = linear model, nnet = neural network, rf = random forest, rpart = decision tree.	43

2.17	Panel A: R^2 and Panel B: RMSE values for predicting CD, values are mean \pm standard deviation. lm = linear model, nnet = neural network, rf = random forest, rpart = decision tree.	44
2.18	Panel A: R^2 and Panel B: RMSE values for predicting ERG, values are mean \pm standard deviation. lm = linear model, nnet = neural network, rf = random forest, rpart = decision tree.	45
2.19	Panel A: Accuracy and Panel B: Kappa values for predicting grade, values are mean \pm standard deviation. lm = linear model, nnet = neural network, rf = random forest, rpart = decision tree.	47
3.1	Short name and description of each image type identified by the historical data. Image types that were used for survival analysis are indicated by the 'analyzed' column.	57
3.2	Regular expressions for classifying MR studies. Spine, intraop, non-brain, and magnetic resonance angiography (MRA) studies were excluded from further analysis. ABTI: advanced brain tumor imaging.	57
3.3	Regular expressions used to classify MR series by description. The non-anatomical series are not used for further processing.	58
3.4	Total number of studies of each type categorized by the regular expression library for all patients. OSF: outside facility	73
3.5	Total number of series of each type categorized by the regular expression library for all patients with any brain MR images.	74
3.6	Fixed image type for each study. T13D, CubeFLAIR, and WandT2 are usually near 1 mm isotropic resolution or better.	74
3.7	Number of images of each type included in the curated data set. Only cases with a full preoperative or postoperative study were included.	75
3.8	Data processing times for a representative image study (fixed image size 512x512x124). Times are for single-node processing, note that many nodes can run simultaneously. ⁺ Dicom conversion is run on a separate workstation not on a computing cluster. *CLARA inference runs all inference sequentially inside a docker which creates a bottleneck in the processing.	75
3.9	Proportion of cases where the selected CSF ROI included the given location. Multiple locations are possible so the totals exceed 100%	76
3.10	Result of data QA for all 1717 cases. Review is NA if a preop or postop study was not available for that patient.	77
3.11	Proportion of each failure mode in the data processing pipeline. Steps are listed in the order they occur.	77
3.12	Result of data QA for all 285 BraTS cases.	78
3.13	Proportion of each failure mode in the data processing pipeline for the BraTS data. No other failure modes were observed.	78
3.14	Total intracranial volume (TIV) measured using 3D imaging for several different populations. [†] This value was reported in Table 1 of [1] but it is inconsistent with the reported mean and range of values. The true value is likely around 155 based on the other values in that table.	80
3.15	How CLARA segmented volumes are added to compare to reference T1 enhancing and T2 FLAIR volumes. Enhanc is enhancing tumor (blue in Figure 3.7), neh is nonenhancing tumor core and necrosis (green in Figure 3.7).	81

3.16	Root mean square error in cm^3 for preoperative segmented tumor volume (T2-FLAIR and T1 enhancing) compared to reference tumor volume. Separate values are given for each level of data quality.	83
3.17	Summary of curation effort. Three steps that take an appreciable amount of time: A) Downloading raw DICOM data from PACS. This takes took a few weeks since we throttle and only transferred during nights and weekends. B) Processing the image data from raw DICOMs to a final data matrix can be done in about 24 hours on a moderate sized computing cluster. C) Data review takes about 80 human hours to review all studies. This includes breaks and distractions. Additional time to re-processed and review bad data included. ¹ R shiny app . ² Includes to develop and test processing pipeline. ³ Not included in total labor amount.	89
4.1	The Karnofsky performance scale, reproduced from: <i>Oken, M.M., Creech, R.H., Tormey, D.C., Horton, J., Davis, T.E., McFadden, E.T., Carbone, P.P.: Toxicity And Response Criteria Of The Eastern Cooperative Oncology Group</i> . Am J Clin Oncol 5:649-655, 1982. [2]	91
4.2	IDH mutation rates stratified by patient age (above or below 55 years) and WHO grade. The number of cases per group is given as N. IDH mutation status was confirmed by IDH1 R132H immunohistochemical staining or sequencing.	96
4.3	Images, regions (Figure 3.7), and features used to compute image measurements. Shape features Voxel Volume and Max 3D Diameter only depend on the region being measured over not the underlying image. . . .	100
4.4	Definition of concordant and discordant pairs in terms of feature values x and survival times y	101
4.5	Clinical data summary for all 1181 preoperative cases. IDH1 mutation status is listed for cases where IDH1 mutation status was explicitly mentioned in the clinical record. Median overall survival (OS) was not reached for the WHO II group.	105
4.6	Histologic diagnoses by WHO grade for all patients. The histologic grades are mostly consistent with the 2012 WHO grading scale. NOS=Not otherwise specified.	106
4.7	Multivariate cox proportional hazards model hazard ratios and significance for all cases using only clinical factors that strongly influence survival. Confidence intervals are 95%	106
4.8	Cox proportional hazards model hazard ratios and significance for tumor volumes, multivariate ratios include controls for age, KPS and grade. EV = enhancing volume, WT = whole tumor	107
4.9	Best image features from each image type among patients with all WHO grades (II III IV) in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate (Panel A) and multivariate (Panel B) models. p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons.	108

5.1	Clinical data summary for all cases with preoperative and postoperative imaging. IDH1 mutation status is listed for cases where IDH1 mutation status was explicitly mentioned in the clinical record.	140
5.2	Histologic diagnoses by WHO grade for all patients with preoperative and postoperative imaging. NOS=Not otherwise specified.	140
5.3	Best <u>postoperative</u> image features from each image type among patients with all WHO grades (II III IV) in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate (Panel A) and multivariate (Panel B) models. p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample.	141
5.4	Multivariate Cox proportional hazards model for WHO IV cases using CLARA segmented enhancing volume EOR measurements and clinical factors that strongly influence survival. Confidence intervals are 95% . . .	158
5.5	Multivariate Cox proportional hazards model for all WHO grades (II III IV) cases using CLARA segmented enhancing volume EOR measurements and clinical factors that strongly influence survival. Confidence intervals are 95%	158
5.6	Multivariate Cox proportional hazards model for all WHO grades (II III IV) cases using CLARA segmented total tumor volume EOR measurements and clinical factors that strongly influence survival. Confidence intervals are 95%	158
5.7	Best <u>extent of resection</u> features from each image type among patients with all WHO grades (II III IV) in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate (Panel A) and multivariate (Panel B) models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. For local grade (GR) and T1 post-contrast (TC) no remaining features were univariate significant so they are omitted from the table. IS = in sample.	161
A.1	Correlation between NLM and Mode-CSF normalized image intensities. .	186
A.2	Multivariate cox proportional hazards model hazard ratios and significance for grade II cases using only clinical factors that strongly influence survival. Confidence intervals are 95%	189
A.3	Multivariate cox proportional hazards model hazard ratios and significance for grade III cases using only clinical factors that strongly influence survival. Confidence intervals are 95%	189
A.4	Multivariate cox proportional hazards model hazard ratios and significance for grade IV cases using only clinical factors that strongly influence survival. Confidence intervals are 95%	189
A.5	Cox proportional hazards model hazard ratios and significance for tumor volumes among grade II cases, multivariate ratios include controls for age, and KPS	189

A.6	Cox proportional hazards model hazard ratios and significance for tumor volumes among grade III cases, multivariate ratios include controls for age, and KPS	189
A.7	Cox proportional hazards model hazard ratios and significance for tumor volumes among grade IV cases, multivariate ratios include controls for age, and KPS	189
A.8	Best preoperative image features from each image type among patients with WHO grade(s) II in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample	190
A.9	Best preoperative image features from each image type among patients with WHO grade(s) III in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample	190
A.10	Best preoperative image features from each image type among patients with WHO grade(s) IV in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample	191
A.11	Best <u>postoperative</u> image features from each image type among patients with WHO grade(s) II in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample	196
A.12	Best <u>postoperative</u> image features from each image type among patients with WHO grade(s) III in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample	196

- A.13 Best postoperativeimage features from each image type among patients with WHO grade(s) IV in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample 197
- A.14 Best EORimage features from each image type among patients with WHO grade(s) II in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample 202
- A.15 Best EORimage features from each image type among patients with WHO grade(s) III in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample 202
- A.16 Best EORimage features from each image type among patients with WHO grade(s) IV in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample 202
- A.17 Best image features from each image type among patients with WHO grade(s) II III IV with known IDH mutation status in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. No local grade features had significant univariate hazard ratios on this subset. Uncorrected p-values are listed along with corrected significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). 208
- A.18 Best postoperativeimage features from each image type among patients with WHO grade(s) II III IV in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. No local grade, T2, or synthetic pathology features had significant univariate hazard ratios on this subset. Uncorrected p-values are listed along with corrected significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). 208

- A.19 Best EORimage features from each image type among patients with WHO grade(s) II III IV in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. No local grade, Ki67, CD, T1, TC, or FLAIR features had significant univariate hazard ratios on this subset. Uncorrected p-values are listed along with corrected significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). 208
- A.20 Survival stratification using the in-sample cutoffs for raw image features trained on the historical data and applied to the BraTS 2018 data. Results are for univariate analysis. Validation HRs marked * are significantly different from 1 (log-rank $p < 0.05$) 209

Abbreviations and Symbol Names

ADC	Apparent Diffusion Coefficient
CBF	Cerebral Blood Flow
CBV	Cerebral Blood Volume
DCE	Dynamic Contrast-Enhanced
DSC	Dynamic Susceptibility Contrast
DTI	Diffusion Tensor Imaging
DWI	Diffusion-Weighted Imaging
EORfrac	fractional reduction in feature value from preoperative to postoperative images
EORreduc	reduction in feature value from preoperative to postoperative images
ERG	Erythroblast Transformation-Specific (ETS) Related Gene
FL	FLAIR image
HR	hazard ratio from proportional hazard model
K^{trans}	Volume Transfer Constant
Ki67	Cell proliferation marker
KPS	Karnofsky performance status
RF	Random Forest
RMS	Root Mean Square
ROI	Region of Interest
TC	T1-weighted image post-contrast
VOI	Volume of Interest
WHO	World Health Organization

Chapter 1

Introduction

Malignant primary brain and central nervous system (CNS) tumors occur at an incidence rate of approximately 7 cases per 100,000 individuals per year. Gliomas, specifically glioblastomas, are the most common and make up nearly half of all malignant CNS tumors [3]. Glioblastoma has one of the worst prognoses of any cancer and a median overall survival of only 14 months with the best current therapy [4]. However, low-grade glioma patients have survival statistics measured in years or decades especially with good treatment [5, 6]. Patients with gliomas are treated by three main methods: surgical resection and debulking reduces the total tumor burden and provides diagnostic tissue samples. radiotherapy treats microscopic infiltrative disease in the postoperative cavity, and chemotherapy is given systemically to slow the progression of the disease.

Despite a multifaceted treatment approach, gliomas are notorious for ultimately recurring. This is likely due to the well-known tumor heterogeneity of pathologic and molecular factors which complicates treatment [7–13]. While a more complete surgical resection is favorable [14, 15], what constitutes a “total resection” is debated. The margins applied to radiation plans are largely empiric, because tumor borders on imaging are so poorly defined. Only a single chemotherapeutic agent, temozolamide, has proven effective against glioma with an acceptable safety profile, despite much (and ongoing) work [4].

Many of the difficulties of caring for glioma patients arise from imaging problems. Magnetic resonance (MR) is the main modality of imaging for tumor interrogation, but while tumors are known to be highly heterogenous pathologically, this is only partly reflected

in current imaging, with different areas of radiographically similar tumor showing divergences in grade [14–17]. There is a **great need** for improved imaging to better delineate and characterize tumor in the brain, specifically dealing with spatial heterogeneity to allow improved biopsy, surgical resection and radiation planning. The goal of this work is to provide: a) an imaging-pathology translation key for glioma, and b) applied this as a prognostic tool and c) as a measure of surgical impact on prognosis or outcome.

We do this using imaging-based machine learning models trained on spatially specific tumor samples extracted during a prospective diagnostic clinical trial (PI: Schellingerhout). Having a complete description of both imaging and pathological findings at selected points in patient brains allowed us to develop models that can translate from imaging to pathological features and generate parametric maps of predicted pathological features. This approach maximizes the extraction of clinically useful information contained in multi-parametric magnetic resonance (MR) imaging, and allows us to more successfully deal with tumor heterogeneity [18]. MR imaging, especially advanced and functional techniques are known to correlate globally with tumor malignancy. However, the point-wise correlation between imaging and pathology is not fully understood and represents a knowledge gap that we address in this project. This project naturally divides into two aims:

First, we correlated imaging with local pathology and generate predictive models of glioma tumor pathology and grade. Our working hypothesis for this aim is that correlations between imaging and tumor pathology are sufficiently consistent across the glioma patient population to make accurate local predictions of proliferation, vascularity, cellularity, and grade using machine learning models. To overcome the confounding effects of tumor heterogeneity and allow us to estimate grade at a voxel-wise level, we use pathology from spatially precise stereotactic tissue biopsies, acquired in a clinical trial, that have a known location in the brain. By including samples from a diverse pool of overall grades as well as from different regions like enhancing volume or non-enhancing regions, we trained models to make reliable predictions for new patients using only imaging data. This provides the foundation for glioma pathological parametric maps.

Second, we evaluated the effect of biological heterogeneity and extent of resection on prognosis. Current literature establishes the prognostic threshold for gross total resection (GTR) at 98% removal of enhancing tumor [14]. An example of a total resection is

shown in Figure 1.1. However, to preserve brain function, GTR is achieved in only 30% to 50% of cases [14, 15]. Without GTR, the chance that highly aggressive disease remains increases, especially when high grade disease is near the tumor boundary [16]. Our working hypothesis here is that similar survival benefit to GTR will be achieved so long as the estimated highly malignant disease, as estimated by pathological parametric maps, is removed. Predicting disease burden pre- and post- operatively gives us a biologically relevant measure of surgical resection in terms of residual volume of grade disease. We curated data from a very large historical cohort of > 1000 patients at our institution to validate the prognostic effect. Using real clinical data makes our algorithms widely applicable and robust to variations in clinical image quality. As part of the analysis, we constructed multivariate survival models using estimated pathology and known prognostic factors like treatment history, age, and genetic markers [19].

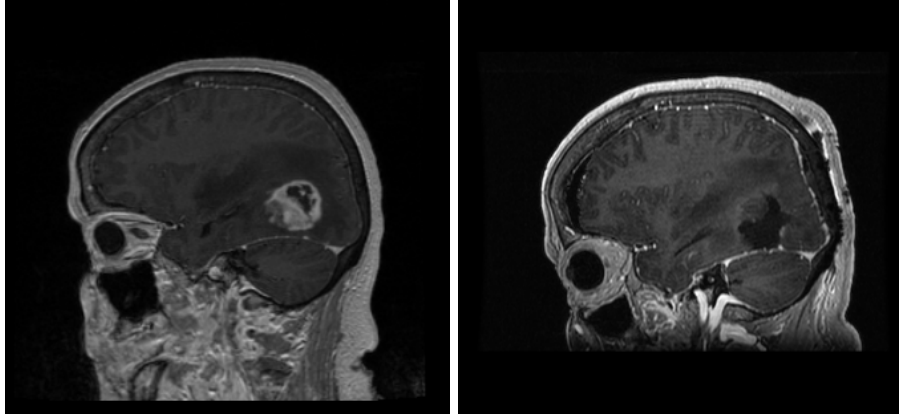


FIGURE 1.1: Contrast enhanced T1w images of a high-grade glioblastoma. Left: A lobular enhancing tumor component is visible in the posterior portion of the brain. Right: After surgery, the enhancing tumor is completely removed. Therefore, this patient is said to have received a gross total resection

1.1 Dissertation Organization

The organization of this dissertation is as follows: Chapter 2 addresses the first aim of this project and establishes image-based predictive models of tumor pathology. Chapter 3 describes the large-scale curation of historical data used for survival analysis. Chapter 4 correlates traditional radiomics image features and measurements from estimated pathology maps with survival. Chapter 5 expands the analysis to postoperative

imaging features and analyzes differences between preoperative and postoperative features with respect to survival. Chapter 6 summarizes the conclusions and future work of this project.

1.2 Background

1.2.1 World Health Organization Grading of Gliomas

This work deals with the diagnosis and treatment of adult diffuse gliomas corresponding to World Health Organization (WHO) grades II, III, and IV. Grade I tumors are normally found in pediatric populations, and are thus not represented in our adult study. The following summary of the WHO grading scale refers to this subset of gliomas in particular in order to contextualize the analysis and terminology used.

Gliomas are family of proliferative central nervous system diseases that have highly variable prognosis depending on the level of malignancy. Some patients live 15 or more years with proper treatment and others have extremely poor prognosis with overall survival around 12 months [4–6]. Capturing this range in prognosis is the one main goal of the World Health Organization (WHO) grading scale [20], with the other goal being the effective and precise classification of similar disease subsets. For decades, glial tumors were classified based on their histologic appearance and supposed cellular lineage (e.g. from astrocytes or oligodendrocytes). This was accomplished primarily using light microscopy and histologic staining. As of the fourth revision of the WHO grading scale in 2007 [21], tumors were classified using an overall grade and a histologic type. In short, cellular atypia indicated a WHO grade II tumor and the presence of anaplasia and mitotic figures (beyond solitary mitoses) indicated a grade III tumor. The presence of microvascular proliferation or necrosis indicated a highly malignant grade IV tumor. Along with these grades there were four main histologic types for diffuse gliomas: astrocytoma, oligodendroglioma, mixed oligoastrocytoma, and glioblastoma. Glioblastomas were exclusively grade IV but the other histologic types could be either grade II (i.e. astrocytoma) or grade III (i.e. anaplastic astrocytoma) [21, 22].

More recently, studies discovered key genotypes associated with specific histologic diagnoses and large differences in prognosis. This led to a major update to the WHO grading

scale in 2016 [20, 22, 23]. Oligodendrogliomas were defined explicitly by the presence of IDH mutation and simultaneous co-deletion of chromosome arms 1p and 19q (1p/19q codeletion). With this change, the mixed histologic diagnosis of oligoastrocytoma was essentially removed since these tumors could be genetically classified as either oligodendroglioma or astrocytoma with the exception of a few rare cases [24]. The integrated diagnosis for most diffuse gliomas was also revised to include IDH mutation status, reflecting the large prognostic differences observed between tumor of the same overall grade. For example: *Glioblastoma, IDH wild-type* versus *Glioblastoma, IDH mutant*. Other genetic mutations like ATRX or p53 commonly occur alongside IDH mutations. These are noted as characteristic but not necessary for diagnosis of IDH mutant tumors.

A new update to the WHO grading scale is anticipated to be released in mid-2021 which will continue to finalize the inclusion of molecular information in the diagnoses of glial tumors. Much of this information has already been released as a series of statements from the Consortium to Inform Molecular and Practical Approaches to CNS Tumor Taxonomy (cIMPACT-NOW) [25–27]. One noteworthy anticipated change is the upgrading of IDH wild-type grade II and III astrocytic tumors to grade IV if they have EGFR amplification, chromosome 7 gain and 10 loss, or TERT promotor mutations [26]. It is also anticipated that the grade nomenclature will change from roman numerals (II, III, IV) to Arabic numerals (2, 3, 4). Since these changes are not official and since all grading data in this work was collected prior to 2021 we do not change any of our grading based on the 2021 WHO classifications. It should be noted that the guiding principal of the WHO grading system has always been to indicate prognosis, and that each successive version of the grading system represents a refinement of the prior version.

Implications for WHO Grades Used in this Work

The data collected in this work includes biopsy data collected prior to 2016 as well as historical diagnoses dating back to the 1990s. Since the WHO grading scale has changed several times in this time frame, we must be precise when referring to tumor grade so that the assignments are compatible between eras. Although the diagnoses themselves changed in 2016, the final WHO grade assigned for IDH mutant and wild-type tumors with the same histology did not [23]. For example, both IDH mutant and IDH wild-type astrocytoma are WHO grade II, despite the wild-type tumor having a poorer prognosis.

This means that the WHO grade (without reference to molecular subtype) prior to and after the WHO 2016 revision are still compatible. In this work, **when we refer to overall WHO grade, we mean the histologic grade alone** without reference to molecular markers like IDH status. This can be thought of as the otherwise specified (NOS) categories of the WHO 2016 classification. In the historical data, we kept the antiquated diagnosis of “mixed oligoastrocytoma” as well, this category having been eliminated only recently, and with technology not available historically.

1.2.2 Radiographic and Histologic Heterogeneity of Gliomas

This work discusses diffuse gliomas which are grades II, III, and IV where IV is the most malignant. Because the behavior of these tumors varies so much, information on the individual histology and genetics of a particular patient’s tumor is extremely important. Most commonly, this is gathered using tissue specimen from biopsy or surgical resection.

Although tissue analysis provides rich histologic and genetic information, it is limited by the well-known intratumoral heterogeneity of gliomas and thus is vulnerable to sampling error [9, 11, 28–30]. For example, high grade glioblastomas often contain regions of both low and grade disease [9]. The overall clinical WHO grade of the tumor (and corresponding estimate of prognosis) is assigned as the maximum grade found in any tissue specimen from that tumor. If the highest grade is not sampled (e.g. if a biopsy is not taken from the most malignant region) there is a risk of “under-grading” which a previous study has shown to occur in as many as 38% of cases [31]. Sampling a tumor often has to be limited due to risks of neurological deficits to the patient [14]. This leads to technical difficulties if the actual grade of un-resected tissue is higher than the samples extracted.

Diagnostic imaging provides very rich information that can aid diagnosis and treatment. Magnetic resonance imaging (MRI) is the superior modality for brain tumor imaging due to strong soft tissue contrast and a plethora of advanced and functional imaging techniques [32]. Relative to obtaining tissue data, MRI is non-invasive, cheap, safe, and efficient. Key imaging findings like contrast enhancement or T2-weighted hyperintensity strongly correlate with tumor grade. This allows a preliminary diagnosis to be rendered before tissue data is available [33]. MRI also provides a foundation to plan interventions: surgery and radiation rely heavily on 3D image guidance to locate disease

inside the brain. High-resolution MR images are used in conjunction with stereotactic cranial navigation to help neurosurgeons locate tumor during craniotomy. Similarly, contrast enhanced and T2-weighted images are used to define gross target volumes for radiation treatment. All three of these procedures: diagnosis, surgery, and radiotherapy operate on the fundamental assumption of correspondence between imaging and the underlying tumor biology.

Radiographic heterogeneity is well established and generally characterized in terms of the visible sub-regions like enhancing volume, T2 bright regions, and necrosis [34]. Much work has been done to study the relationship between radiographic appearance and prognosis. This field is generally referred to as “radiomics” [35, 36]. Radiomics operates on the hypothesis that image heterogeneity, patterns in intensity and contrasts, capture the underlying biological heterogeneity and helps tease out the differences in clinical progression [35]. Imaging elucidates some of the histological heterogeneity but there is still variation with radiographically similar regions [37–39]. Contrast enhancing tumor, the current target for treatment, is not entirely sensitive nor specific for active malignancy. Thus, there has been considerable research attempting to establish links between imaging and the true underlying biology (e.g. radiogenomics).

Radiographic heterogeneity is also the basis for guiding surgical treatment. Neurosurgical literature has established “gross total resection” (GTR) of tumor as a favorable prognostic factor for glioma patients [14, 15]. However, there are several definitions of total resection. The most common is based on removal of disease that demonstrates contrast enhancement of T1w MRI, Figure 1.1. The enhancement comes from compromised blood-brain barrier due to rapid angiogenesis within the tumor. Thus, enhancement is a surrogate for active tumor tissue. Further definitions of GTR are based on resection of T2w or T2-FLAIR hyper-intensity, which is useful for low grade or non-enhancing gliomas [15]. Essentially, all these radiographically based definitions use imaging as a surrogate to measure the removal of malignant tumor tissue.

Recent work shows that traditional radiomic features based only on preoperative images may have very limited effectiveness on gross total resection (GTR) subpopulations where a majority of tumor is removed [40]. The removal of most of the tumor appears to invalidate predictions made on pre-surgical imaging. This is a major limitation for

radiomics studies that use only preoperative data especially considering GTR is recommended for all glioma grades [41]. In Chapter 5 we look at the change in image features between preoperative and postoperative imaging to examine heterogeneity in the context of treatment. Features that are both prognostic preoperatively and whose change correlate with outcome provide possible targets for surgical intervention or new definitions for extent of resection. This furthers the extensive work done with preoperative radiomics to the post-treatment realm. Like the preoperative case, we analyze using both raw image features and estimated histologic heterogeneity.

1.2.3 Use for Synthetic Pathology Maps in Diagnosis, Surgery, and Radiation Therapy

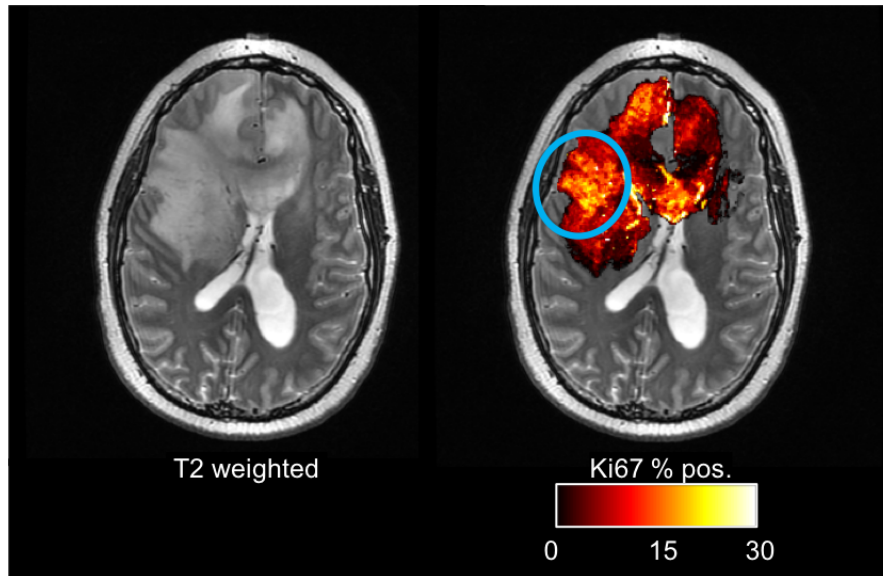


FIGURE 1.2: Example of an estimated proliferation map for a non-enhancing WHO III glioma. The blue circle highlights a region of heightened proliferation in an otherwise radiographically homogeneous area. The extra information comes from the inclusion of extra advanced and functional imaging.

Image-based predictions of glioma pathology can be presented as graphical “synthetic pathology maps” that can be interpreted directly or used alongside conventional diagnostic imaging for treatment planning. An example is shown in Figure 1.2. The first application is in accurately grading gliomas. The difference in overall survival between low-grade gliomas and high-grade gliomas (i.e. glioblastoma) is considerable with median overall survival of 12 months for high grade tumors versus 5 - 11 years for low grade tumors [4–6]. Surgical specimens from tumor resection (preferred treatment for

all grades [14, 41]) provide definitive tumor grading, but for inoperable tumors where only biopsy is possible there is considerable uncertainty in the true tumor grade [31].

After grading, the two therapeutic applications that stand to benefit the most from these maps are surgery and radiation therapy. Surgical resection is recommended for all glioma patients. Modern cranial navigation technology used to guide surgery already overlays conventional MRI with physiologically relevant information from fiber tractography and functional MRI in order to visualize critical structures in three dimensions. Our maps of predicted tumor pathology can integrate seamlessly alongside these other advanced imaging derivatives to guide the surgeon towards highly malignant tumor during resection.

Much of the difficulty with surgical resection is identifying glioma cell infiltration into normal-appearing tissues. There has been some success using diffusion tensor imaging like the sequences in our study to identify white matter infiltration [42]. Combinations of diffusion and perfusion imaging were also recently used to prospectively predict the locations of recurrence within peritumoral edema on a prospectively validated cohort [43]. With these studies in mind, it is reasonable to expect that disease extent can be quantified by machine learning models.

After surgery, nearly all glioma patients also complete a course of radiation therapy. In the radiotherapy planning process, treatment plans are developed using conventional CT and MR imaging to define tumor margins, outline normal tissue structures, and calculate the radiation dose to these areas. We envision our maps would be particularly useful to help define radiation target volumes, ensuring active tumor was adequately covered. Since the maps are image based, they would be easily to integrate with the treatment planning system. Additionally, current radiation doses are already near the limits of normal tissue toxicity. Reducing dose in areas associated with neurological deficits like the hippocampus will provide more flexibility in treatment planning and mitigate these potential neurological deficits. In addition to the highest-grade regions of the tumor, the proposed models can also identify these low-risk areas. Recent work using advanced diffusion tensor MRI to adjust radiotherapy plans has been presented by Jena et. al. [44]. However, at this time planning and evaluation radiotherapy plans is reserved for our future work.

1.2.4 The Data Processing Inequality

The **data processing inequality** states that information content from derived or processed data is less than or equal to the information in the combined raw data [45]. Specifically, the Shannon entropy [46] $\mathcal{H}(y)$ over a series of discrete observations of a random variable $Y = \{y_1, \dots, y_n\}$ is defined as

$$\mathcal{H}(Y) = - \sum_{i=1}^N p(y_i) \log_2 p(y_i).$$

Where p is the probability $p(y_i) = p(Y = y_i)$. From the Shannon entropy, we can compute the mutual information I between Y and another random variable Z in terms of the conditional entropy $\mathcal{H}(Y|Z) = \sum_{j=1}^N p(z_j) \mathcal{H}(Y|z_j)$.

$$I(Y, Z) = \mathcal{H}(Y) - \mathcal{H}(Y|Z) = \mathcal{H}(Z) - \mathcal{H}(Z|Y)$$

The mutual information provides an upper bound on the amount of communicated of information from Y to Z . The data processing inequality states that if $X \rightarrow Y \rightarrow Z$ is a markov chain, then $I(X, Z) \leq I(X, Y)$ [45]. In other words, the information transmitted all the way through the chain is less than the information transmitted at each step. In this work, the steps in the markov process are manifested as transformations of the raw image data (i.e. by random forests to estimate pathology). By the data processing inequality, any prognostic stratification obtained by estimates of proliferative activity etc. are also possible given some combination features from the input images. We can phrase the search for prognostic biomarkers in terms of finding some function using the input images and a binary mask to produce a single numerical risk estimate.

$$f(T1, T2, TC, FLAIR, \text{mask}) : \mathbb{R}^{4N} \times \{0, 1\}^N \rightarrow \mathbb{R}$$

For an image with N voxels, searching for prognostic biomarkers among all possible combinations of image features from T1, T1C, T2, and FLAIR images means searching a space of functions with domain in \mathbb{R}^{4N} . Estimated pathology maps represent an intuitive reduction in the dimensionality of the feature space. A random forest model trained on imaging and pathology data provides a mapping, for example to Ki67, that reduces the dimension of the domain of the functions f .

$$f(\text{Ki67}(T1, T2, TC, FLAIR), \text{mask}) : \mathbb{R}^N \times \{0, 1\}^N \rightarrow \mathbb{R}$$

By using tissue data to estimate Ki67 from T1, T1C, T2, and FLAIR we encode the relevant information into a single 3D image and reduce the search space to functions over \mathbb{R}^N . We can think of the Ki67 model as a biology informed dimensionality reduction. Alternatively, a Ki67 map can be viewed as a synthesis of existing data from multiple images into a single, more clinically useful form.

This biological perspective guiding our analysis is of great importance, as it fights against issues arising from high-dimensional data analysis like the curse of dimensionality and false detection of biomarkers. This is analogous to unsupervised principle component analysis which redefines the basis of a space in terms of orthogonal vectors that describe the greatest proportion of the variance in the data [47, 48]. Examining the prominent features in the first handful of components is one way to identify candidate features or biomarkers. In general, the first components with large variance contain the most information about the target quantities of interest. However, since PCA is unsupervised this is not necessarily the case.

Furthermore, simultaneously assessing multiple images at once is a very challenging task for humans. The ability for physicians to assess disease and plan treatment is much greater when reviewing a synthesis of information like a Ki67 index map versus trying to comprehend all the input images separately. Also, Ki67 index is a familiar and interpretable quantity whereas an arbitrary non-linear $f(T1, T2, TC, FLAIR, \text{mask})$ is unlikely to be meaningfully interpretable.

Chapter 2

Image-Based Prediction of Local Glioma Pathology

2.1 Introduction

The data discussed here and similar results have been previously published in journal publications [49–51] and included in a previous doctoral thesis [52]. While some data reported are the same as previously reported, the analysis methods have been improved and deepened during the production of this work. To summarize: Jonathan Lin initially reported on the collection and analysis of stereotactic biopsy data and presented random forest modeling [52]. Then, in subsequent publications Gates et al. published predictive models trained on the biopsy data using improved, and different, machine learning techniques [49–51]. Now, in this work we further improve on these publications by implementing a new automated normalization scheme and expanding the cross-validation procedure beyond a single five-fold split.

Gliomas are highly malignant primary brain tumors and are the most common type of central nervous system cancer [20]. High-grade gliomas also carry one of the worst prognosis of any cancer type. Standard of care treatment depends somewhat on the specific patient’s demographics and disease state, but generally follows a three-pronged approach. Surgical resection of bulk tumor is now recommended for all glioma patients, even though a “watchful waiting” approach was previously applied to low grade cases [41]. Concurrent radiation and temozolomide chemotherapy are almost always

given: either before surgery to reduce the tumor volume or after surgery to treat microscopic infiltration into the surrounding brain [4]. Despite radical therapy, the increase in patient survival is modest and often measured in months not years. Thus, in addition to novel therapies there is room for improvement in existing radiation and surgical techniques. Such improvements could be in the form of focusing existing therapies on active tumor or reducing damage to normal brain and cognitive deficits [53, 54].

A main contributor to the difficulty in treating gliomas is their heterogeneous and infiltrative nature. Diffuse gliomas subtly invade the surrounding normal brain and it is very difficult, both radiographically and surgically, to define a border between diseased and healthy tissue. Even within the core of a glioma, especially a high grade glioma, there is substantial histologic and molecular heterogeneity. This makes focusing treatment on active tumor areas as opposed to necrotic tissue or peritumoral edema difficult.

Magnetic resonance imaging (MRI) is by far the best imaging modality used to assess the intratumoral heterogeneity and extent of disease in the brain. Currently, clinicians use contrast enhanced T1-weighted MRI alongside T2-weighted and T2-FLAIR MRI to assess this heterogeneity. However, the sensitivity and specificity of these conventional imaging sequences, especially contrast enhancement, is not fully understood [38]. Recently, advanced techniques like diffusion weighted imaging and dynamic contrast enhanced sequences have shown promise in helping to identify highly malignant tumor areas [42, 55–57]. These functional imaging techniques generate signal using physical or physiological properties of the underlying tissues such as restricted diffusion or compromised blood-brain barrier. The complementary information aids in the diagnosis and treatment of glioma patients, but the degree to which these sequences faithfully represent the true underlying biology is an open question [58].

Extensive research has been performed to correlate clinical imaging and brain tumor pathology [35]. This includes the compilation of public repositories from The Cancer Imaging Archive [59, 60] and the organization of international challenges [34]. These data sources focus on advanced image processing or radiomic feature extraction to discover correlates with tumor characteristics. With the exception of characteristics which are known to be homogeneous throughout each glioma like IDH mutation status, image based analysis alone is fundamentally limited by the ability of the images themselves to resolve the intratumoral heterogeneity.

In order to understand these limitations, there is a need to research precise imaging-pathology correlations using spatial specificity to avoid the confounding effects of heterogeneity. To date, several studies have done so using image-guided biopsy sampling of gliomas [38, 43, 61–63]. Although, few have published successful predictive modeling of biological targets. In this work, we follow a similar approach using spatially specific biopsy sampling to estimate local tumor pathology using preoperative MRI data. Our work includes a comprehensive evaluation of conventional and advanced imaging techniques and simultaneously analyzes several key pathological characteristics like proliferation, local grade, cellularity, and vascularity. Furthermore, we use the data collected to train predictive models that can estimate these properties point-wise in new patients. This allows the translation of the imaging-pathological correlations into useful tools for informing prognosis and guiding treatment.

2.2 Methods

Data for this work was collected via a HIPAA-compliant IRB-approved MD Anderson Cancer Center clinical imaging trial (NCT03458676). Adult patients with suspected supratentorial glioma were recruited prior to surgical resection. Exclusion criteria included previous brain tumor treatment; however previous closed biopsy was acceptable. Patients with contraindication to MRI like metal implants or allergy to gadolinium contrast were also excluded. All patients gave informed consent for both the preoperative magnetic resonance imaging study and the collection of biopsy samples during surgery.

2.2.1 Image Acquisition

Each patient was imaged within the three days prior to surgery. The preoperative MR study was acquired on either of two GE 3.0 T MRI scanners: a GE Signa HDxt 3T or GE Discovery MR750 3T (GE Healthcare Technology). The imaging study consisted of both conventional imaging sequence and advanced sequences. Conventional sequences included T1-weighted pre-contrast spin-echo, T1 post-contrast (both spin- and gradient-echo), T2-weighted fast spin-echo, T2 FLAIR, T2 weighted gradient echo, and susceptibility-weighted angiography (SWAN). Both diffusion weighted (DWI) and

diffusion tensor (DTI) image series were acquired. The DWI were processed into parametric maps of apparent diffusion coefficient (ADC, eADC), while DTI were processed to maps of average diffusion coefficient (AvgDC) and diffusion fractional anisotropy (FA). The specific acquisition parameters for the study protocol are listed in Table 2.1 and Table 2.2

	T2	FLAIR	SWAN	3D T1C	T1C SE	T1
Slice Orientation	Axial	Sagittal	Axial	Axial	Sagittal	Axial
Pulse Sequence Name	FSE	3D FSE	3D FGRE	3D FGRE	MEMP	MEMP
TR (ms)	5800	7000	46	5.724 – 8.208	700	700
TE (ms)	77	125	23.1	1.736 – 2.1	11	10
TI (ms)	–	2060 – 2072	–	–	–	–
FA (°)	90-125	90	15	20	90	90
FOV (cm)	19.6 – 23.8	23.04 – 25.6	20	19.2 – 22.4	16.5 – 24	16.5 – 22
Matrix	352 × 224	256 × 256	320 × 224	352 × 224	256 × 192	256 × 192
BW (kHz)	162.773	122.07	244.141	195.312	244.141	244.141
Voxel size (mm)	0.5469 × 0.5469 × 2	0.5 × 0.5 × 1	0.3906 × 0.3906 × 1	0.4688 × 0.4688 × 3.5*	0.9375 × 0.9375 × 5	0.8594 × 0.8594 × 5
ETL	8	140	6	1	1	1
% Phase FOV	70-85	100	100	80	100	75
% Sampling	100	100	69.1964 – 69.7891	100	100	100

TABLE 2.1: Anatomic image acquisition parameters. *One 3D T1C image had voxel size 0.5469 x 0.5429 x 1.8 mm. FSE: Fast Spin Echo, FGRE = Fast Gradient Recalled Echo, ETL: Echo Train Length, MEMP: Multi Echo Multi Planar

Two dynamic acquisitions were acquired each following a bolus of gadolinium contrast. Contrast was given in a bolus of 0.1 mmol/kg either gadopentetate dimeglumine (Magnevist) or gadobutrol (Gadavist; both Bayer Healthcare, Leverkusen, Germany) at 5 mL/sec, followed by 30 mL saline at 5 mL/sec. First, dynamic contrast enhanced images were acquired, then the contrast bolus from the DCE acquisition was used to acquire the T1 weighted post-contrast images. Afterwards, a second bolus was injected and used to acquire a dynamic susceptibility contrast (DSC) series. DCE images were processed into metrics describing the contrast uptake curves and leakage constants K^{trans} , k_{ep} , Peak

	DWI	DTI	DCE	DSC	MRS
Slice Orientation	Axial	Axial	Axial	Axial	Axial
Pulse Seq. Name	SE-EPI	SE-EPI	SPGR	GR-EPI	Probe-P
TR (ms)	8000	10175	3.1	1500	1000
TE (ms)	88	90	1.1	25	144
FA (°)	90	90	15 / 30	60 / 90	N/A
FOV (cm)	22	22	16 – 18	22 / 24	20
Matrix	128 × 128	128 × 128	256 × 160	128 × 160	16 × 16
Voxel size (mm)	0.8594 x 0.8594 x 3	0.8594 x 0.8594 x 3.5	0.7813 x 0.7813 x 2	0.9375 x 0.9375 x 3.5	12.5 x 12.5 x 12.5
Spacing Between Slices (mm)	3.5	3.5	2 – 5	3.5 / 5	N/A
No. of Slices/Volume	24	36	28	16	1
Total No. of Slices	48 – 192	1008 – 1452	576 – 1700	960 – 1440	1
ETL	1	1	1	1	1
% Phase FOV	100	100	75	100	100
% Sampling	100	100	100	100	100
NEX	See below	1	1	1	2
b-values (s/mm ²) (NEX)	0(1), 150(1), 1000(1), 2000 (2)	1200 (1) (N=27 encoding directions)	–	–	–
No. of Phases	–	–	36 – 60	60	–

TABLE 2.2: Functional Image acquisition parameters. SE: Spin Echo, GE: Gradient Echo, EPI: Echo Planar Image, SPGR: Spoiled Gradient Recalled

enhancement, mean transit time (MTT), time to peak (TTP), using NordicICE (Nordic Neuro Labs). DSC was processed into cerebral perfusion metrics for blood flow (CBF) and relative volume (rCBV) as well as contrast fractions v_p and v_e using NordicICE. Both DCE and DSC were processed using arterial input function (AIF) deconvolution using semi-automatically selected pixels. The AIF was measured in the middle cerebral artery ipsilateral to the lesion when possible. The anterior cerebral artery was used if no suitable AIF in the middle cerebral artery could be found. Magnetic resonance spectroscopy within a tumor ROI placed by a neuroradiologist was also acquired during this study but this data is not included in the analysis.

2.2.2 Image Registration, Normalization, and Extraction of Values

2.2.2.1 Dynamic Contrast Enhanced (DCE) Image Processing

DCE imaging uses a temporal series of T_1 -weighted images to observe the passage of a contrast agent throughout the vascular space and across the blood-brain-barrier. In this case, the contrast agent is a gadolinium chelate which creates T_1 shortening effects on the nearby water molecules. The concentration of tracer (contrast) in C_t in tissue and C_p in plasma is modeled by a compartment model [64].

$$\frac{dC_t}{dt} = K^{\text{trans}} (C_p - C_t/v_e) = K^{\text{trans}} C_p - k_{\text{ep}} C_t$$

Where K^{trans} is the transfer constant into the extravascular-extracellular space and k_{ep} is the back-flow transfer constant. When we apply boundary conditions based on the incoming bolus of contrast we get

$$C_t(t) = v_p C_p(t) + K^{\text{trans}} \int_0^t C_p(\tau) e^{\frac{K^{\text{trans}}}{v_e}(t-\tau)} d\tau$$

The contrast agent concentration is linear related to the relaxivity of tissue as $R_1 = r \times C_t | R_{10}$ where $R_{10} = 1/T_{10}$ is the baseline relaxivity and r is the relaxivity of the contrast agent [65].

Dynamic contrast enhanced imaging was processed using Nordic ICE software (v.2.3.14, Nordic Neuro Lab, Bergen, Norway). This software is FDA approved for permeability and perfusion image analysis. The 4D time series was loaded into the DCE module and processed using the following procedure: Signal conversion “1/T1 to SPGR” was selected and the temporal resolution was updated based on the DICOM header tag (0018,1060) (Trigger Time). Spatial smoothing, temporal smoothing, and noise thresholds were not applied to the time series data. Next, vascular deconvolution was enabled and an arterial input function (AIF) was semi-automatically chosen in the middle cerebral artery ipsilateral to the lesion. The AIF was deemed acceptable if the peak signal intensity was at least five times the mean intensity. If no acceptable AIF could be found in the middle cerebral artery, the anterior cerebral artery and basilar artery were searched as well. In this case, an acceptable AIF is one with peak height at least 5 times greater than mean curve and in the range $\Delta R1 \in [20, 40]$ where $\Delta R1$ is the change in relaxation

due to the contrast. Hematocrit correction factor was set to 0.45, T1 set to 1200 ms, and the AIF-tissue delay was automatically calculated. The full set of extended Tofts pharmacokinetic parameters were extracted including K^{trans} , k_{ep} , v_p , v_e as well as additional parameters: Peak enhancement, time-to-peak enhancement, and area under signal-time curve [64, 66].

2.2.2.2 Dynamic Susceptibility Contrast (DSC) Image Processing

Similarly to DCE, DSC uses a serial acquisition during injection of paramagnetic contrast agent to characterize the perfusion, or blood flow, through the tissue. Unlike DCE which uses the T_1 shortening effect, DSC signal takes advantage of differences in magnetic susceptibility in the vascular beds of tissue as the contrast agent passes through. The signal is visible in T_2^* -weighted images. The blood volume (BV) is related to the integral ratio of the contrast agent concentrations in tissue and artery as derived from a compartment model.

$$BV = \frac{k_H}{\rho} \frac{\int_0^\infty C_T(t) dt}{\int_0^\infty C_A(t) dt}$$

Where k_h is a hematocrit scaling factor accounting for vessel size and $\rho = 1.04$ g/mL is the tissue density. The contrast agent concentration for the whole tissue is described by the following.

$$C_T(t) = \frac{\rho}{k_H} \cdot BF \cdot \int_0^t C_A(\tau) R(t - \tau) d\tau$$

Here, $R(t - \tau)$ is a residue function that describes the fraction of contrast remaining. Solving for blood flow (BF) is performed by using the experimental tissue concentration and arterial concentration curves and deconvolving the residue function. To do so, singular value decomposition is the most common [67–69]. Blood volume and blood flow are reported as relative values $rCBV$ and $rCBF$ since the true susceptibility calibration factors are unknown. In addition to blood flow and volume, we also calculate mean transit time $MTT = rCBV/rCBF$ and the delay time for contrast arrival.

Dynamic susceptibility contrast images were processed by the same procedure as DCE with a few differences: Signal conversion was set to “SI to delR2” instead of “1/T1 to SPGR,” temporal resolution was set to 1.5 s, TE set to 25 ms. Furthermore, contrast

agent leakage correction was applied. The Tofts model parameters exported after processing were cerebral blood volume (both corrected and uncorrected), cerebral blood flow, leakage parameter K2 (with cutoff of 0.1), delay time, and mean transit time.

2.2.2.3 Diffusion Weighted Imaging

Diffusion weighted imaging (DWI) and diffusion tensor imaging (DTI) use diffusion sensitizing gradient pulses to detect Brownian motion of water in tissue [70]. Measures of water diffusivity can be a surrogate for architectural disruption in brain structures like white matter or increased cell packing inside tumors [71]. In short, the diffusion encoding gradients de-phase and refocus spins after some short delay. Spins that have moved by diffusion are not perfectly refocused and show reduced signal. The reduction in signal at echo time is related to the b -value of the pulse sequence and the apparent diffusion coefficient (ADC).

$$S(TE) = S_0 e^{-b \cdot ADC}$$

Where the b -value $b = \gamma^2 \delta^2 G^2 (\Delta - \frac{\delta}{3})$ is a function of the gyromagnetic ratio γ , and diffusion gradient duration (δ), amplitude (G), and delay (Δ). The ADC is computed by acquiring several diffusion weighted images and fitting an exponential to the voxel-wise values. By applying the diffusion encoding gradients in many directions, a 3×3 diffusion tensor can be constructed which describes the diffusion in multiple directions [72, 73].

$$\mathbf{ADC} = \begin{bmatrix} ADC_{xx} & ADC_{xy} & ADC_{xz} \\ ADC_{yx} & ADC_{yy} & ADC_{yz} \\ ADC_{zx} & ADC_{zy} & ADC_{zz} \end{bmatrix}$$

From this matrix, the eigenvector decomposition with associated eigenvalues $\lambda_1, \lambda_2, \lambda_3$ represent the diffusion along the three principal axes. These give rise to the Average diffusion Coefficient (AvgDC) and fractional anisotropy (FA).

$$\text{AvgDC} = \frac{ADC_x x + ADC_y y + ADC_z z}{3} \quad (2.1)$$

$$\text{FA} = \sqrt{\frac{3}{2} \frac{\sqrt{(\lambda_1 - \bar{\lambda})^2 + (\lambda_2 - \bar{\lambda})^2 + (\lambda_3 - \bar{\lambda})^2}}{\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}}} \quad (2.2)$$

The fractional anisotropy is 0 when the diffusion is isotropic (e.g. in free water) and 1 for diffusion that occurs only along a single axis.

2.2.2.4 Registration, Normalization, Measurement

First, DICOM files were converted to NIfTI¹ file format for readability by software packages. Non-brain tissues were removed from each image using image-specific brain masks. The brain masks were initialized using the brain extraction tool (bet2) from the FMRIB software library [74] and further refiled in Amira3D (v6.0, FEI, Hillsboro, OR). The skull-stripped brain volumes for each patient were co-registered that patient's T2-weighted image. Image registration was performed using ANTs [75, 76]. Specifically, A 6 degree-of-freedom (DOF) rigid registration followed by a 12 DOF affine registration using mutual information. After registration, the images were resampled to the same matrix size as the fixed image using nearest-neighbor interpolation. Results of the registration procedure have been previously reported with the 12 DOF affine registration providing the lowest average target registration error [77]. The reported FLAIR to T2w image registration error was 1.55 ± 0.46 mm and T1 post-contract to T2w image registration error was 1.06 ± 0.16 mm. Other sequences have similar magnitude errors.

Maps derived from diffusion weighted and diffusion tensor imaging were registered to the T2w image by skull stripping and registering the b0 image first then applying the resulting transformation the maps. For DCE, the maximum intensity projection of the time series was used to perform registration and for DSC the average intensity image was used.

Intensity normalization was performed on the anatomical MR sequences to provide comparable intensity values for quantitative analysis and predictive modeling. Rather than normalizing using intensity histogram statistics only (e.g. Z-score), we normalized each image using the intensities of reference tissues. The specifics are inspired by similar

¹nifti.nimh.nih.gov

work in literature [63, 78]. First, a normal tissue mask was derived from the common space brain mask by subtracting any visible lesion associated with the tumor. Then, a 3D region-of-interest in the cerebrospinal fluid (CSF) was defined using mixture model clustering. The CSF ROIs were visually inspected to ensure they were completely contained in CSF on all images. Then, the intensities of each image were linearly scaled such that the modal intensity of the CSF and the modal intensity of the normal brain are 0 and 1. To prevent inverting the image contrast, the CSF is mapped to 0 intensity for T1-weighted, FLAIR, and SWAN images and mapped to 1 on T2 and T2-star weighted images. Note, our previously published results used a slightly different normalization method that scaled using mean intensities in hand-drawn gray matter, white matter, and CSF ROIs [49, 50]. Given the high similarity between these normalization schemes, the predictive modeling results are nearly identical. The comparison of values is listed in appendix Section A.2. Quantitative maps from DWI, DTI, DSC, and DCE were not normalized using reference tissues and the values were extracted as provided by the scanner or processing software.

For all images, a 5-mm spherical ROI was placed at the location of each biopsy site recorded during surgery. The size of the ROI accounts for the physical extent of the tissue sample, approximately 10x2x2 mm, and the small amount of uncertainty in the sampling location from the cranial navigation software [79, 80]. For each biopsy ROI, a paired “virtual” ROI was placed in contralateral normal appearing white matter (NAWM) to serve as a control and capture the image characteristics of normal brain. The ROIs were placed in the closest, most ideal, NAWM and reviewed by board-certified neuroradiologist to ensure they were entirely contained in NAWM. The average intensity for each biopsy ROI and contralateral ROI was extracted for each of the 25 total co-registered and normalized images. In other words, each biopsy had 25 independent image measurements associated with it. A full list of parameters is given in Table 2.3

When converting to NifTI format for further processing, scaling factors from the DICOM headers were explicitly accounted for in order to translate pixel values to quantitative measurements. For maps derived from DSC and DCE imaging, tag (0077_1001) or (0077_1101) was used and diffusion imaging a constant scale factor of 10^{-6} was applied.

Image Family	Measurement	Units	Description
Conventional	T1	CSF - brain mode	T1-weighted image intensity
	T2	brain mode - CSF	T1-weighted image intensity
	TC SE	CSF - brain mode	Spin-echo T1 weighted post-contrast image intensity
	TC SPGR	CSF - brain mode	Gradient-echo T1 weighted post-contrast image intensity
	FLAIR	CSF - brain mode	T2-FLAIR image intensity
	T2*	CSF - brain mode	T2*-weighted image intensity
	SWAN	CSF - brain mode	SWAN image intensity
Diffusion	ADC	mm^2/s	Apparent Diffusion Coefficient from DWI
	AvgDC	mm^2/s	Average Diffusion Coefficient from DTI: , Equation 2.1
	eADC	unitless	Exponential Apparent Diffusion Coefficient from DWI: $\exp(-b \cdot ADC)$
	FA	unitless in $[0, 1]$	Fractional Anisotropy from DTI, Equation 2.2
DSC Perfusion	CBV	Arbitrary	Relative cerebral blood volume
	CBF	Arbitrary	Relative cerebral blood flow
	MTT	Seconds	Average time for blood to pass through tissue $MTT = CBV/CBF$
	Delay	Seconds	Time delay to peak signal intensity
	K2	Arbitrary	Blood vessel leakage parameter
DCE Permeability	K^{trans}	1/min	Transfer constant from vascular to tissue compartment.
	K_{ep}	1/min	Backflow transfer constant from tissue to vessel compartment.
	v_p	Percent	Plasma space voxel fraction
	v_e	Percent	extravascular extracellular voxel fraction
	Wash-In	Arbitrary	slope of time-signal curve from contrast entry to peak intensity
	Wash-Out	Arbitrary	slope of time-signal curve after peak intensity
	TTP	Seconds	Time to peak signal intensity
	AUC	Arbitrary	Area under time-signal curve
	Peak	Arbitrary	Magnitude of peak signal intensity

TABLE 2.3: List of image measurements collected for each biopsy.

2.2.3 Biopsy Collection and Pathology Analysis

A neuroradiologist and neurosurgeon reviewed the preoperative imaging data to select targets for biopsy. For each patient, at least one biopsy target was first selected using conventional imaging like contrast enhancement and T2 hyperintensity while blinded to the advanced imaging data. Then, the advanced imaging was used to select another biopsy target in areas of high K^{trans} , high rCBV, or restricted diffusion. Up to five total targets were selected per patient and trajectories to these targets were planned using the Brainlab IPlan software (Brainlab AG, Feldkirchen, Germany) and subsequently loaded into the cranial navigation system in the operating room.

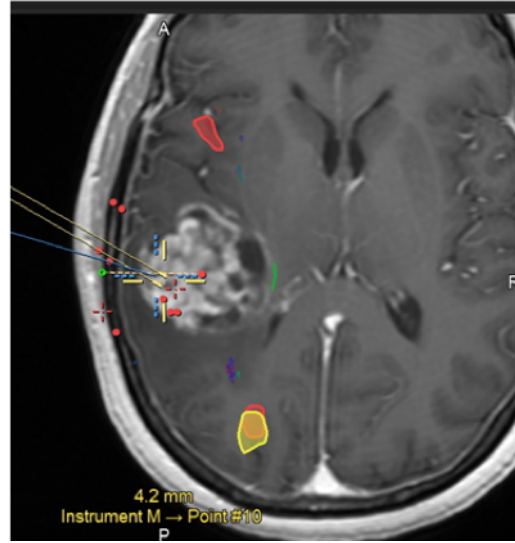


FIGURE 2.1: Screenshot of the Brainlab cranial navigation software during sample collection. The instrument tip (crosshair) is overlaid on top of a preoperative T1-weighted image that is co-registered to the patient’s physical space in the operating room. The coordinates of the instruments are recorded during biopsy.

A primary goal during tissue sample collection was to maximize spatial fidelity between the cranial navigation software and the physical space of the operating suite. Specifically, collecting biopsy samples prior to opening of the skull minimized brain shift which corrupts the spatial accuracy of the cranial navigation system as the brain deforms from its preoperative position [81]. First, the patient was placed in a Mayfield head frame (Integra LifeSciences Corp., Plainsboro, NJ). and the stereotaxy was established with the cranial navigation system. Next, the scalp was retracted and a burr-hole was drilled in the skull to provide access to the brain. A needle trocar was passed trans-durally into the brain towards the biopsy pre-planned biopsy targets. When possible, biopsies were

also collected en-route to the target up to a total of five biopsies per patient. As the tissue samples were collected, the location of the needle tip and trajectory were recorded [79, 80]. An example is shown in Figure 2.1. From this data, the location of the side cutting window was calculated and this exact sampling location, which is near but not exactly coincident with the original target location, was used for further analysis. For some cases, the surgeon opted to use a pituitary probe resembling forceps to collect tissue samples. The location of the instrument’s tip was similarly recorded as samples were collected.

The tissue samples were paraffin embedded and sectioned at 4 μm thickness for histologic analysis. The sections were stained with hematoxylin and eosin (H&E) and individually assessed for tumor presence grade by a board-certified neuropathologist who was blinded to imaging data. The World Health Organization histologic glioma grading scale with grades II, III, and IV was used [21]. Note, an individual sample grade is not necessarily the same as the patient’s clinical tumor grade.

Immunohistochemical (IHC) staining against the proliferation marker Ki67 was applied the tissue samples using the MIB1 antibody. Ki67 is a nuclear antigen expressed by actively dividing cells and not expressed by normal brain [82–84]. High Ki67 labelling index is strongly associated with poor outcomes [85]. After staining, the fraction of positively stained nuclei was semi-automatically measured to record a Ki67 labeling index. Another IHC stain using monoclonal antibody 9YF (Biocare Medical, Concord, CA) was used to stain against the Erythroblast Transformation Specific (ETS) related gene, called ERG. ERG is a vascularity marker expressed more-so in angiogenic tumors than in normal brain tissues. After staining, the fraction of slide area positively stained for ERG was recorded. ERG staining was also used to measure cellular density with Aperio ImageScope software (Leica Biosystems, Inc., Buffalo Grove IL).

Isocitrate dehydrogenase (IDH) mutation status for the biopsy samples was assumed to be consistent with the overall IDH mutation status of the tumor since IDH is not heterogeneously expressed in gliomas [86, 87]. Tissue samples that were normal appearing were assumed to be IDH wild-type.

To serve as controls for each of the real biopsy samples, we extracted image intensity from contralateral VOIs in normal appearing white matter. Since no real tissue data was available, we assumed the pathology target values using literature values. Ki67 is

not expressed in normal brain so we assume the labeling index of virtual samples to be identically 0% [82–84]. ERG expression was randomly imputed using by drawing from a normal distribution with mean and standard deviation 0.199 ± 0.114 % positive area. [88]. Virtual biopsies were assumed to be IDH wild-type.

2.2.3.1 Derivation of Normal White Matter Cell Density

Normal white matter cell density can be estimated using a few methods present in literature including stereology, isotropic fractionator, and histology. A brief summary of these methods is given in the Appendix Section A.3. We found the most appropriate method to be histology based on work by Roetzer et al. who calculated cellularity in whole-brain sections of brain tumor patients [89]. Cell density is measured in nuclei per square millimeter meaning it depends on the thickness of the stained section. Translating between cell density measure at $6 \mu\text{m}$ thickness in Roetzer [89] and our measurements at $4 \mu\text{m}$ thickness can be done using Abercrombie’s formula [90, 91]. Below, we briefly present a derivation of the formulas and correction of values.

Consider the measured (per area) cell density for nuclei of z -axis height H with no “lost caps error” sectioned at thickness t . The relation between observed cell nuclei in the section and the actual number is given by equation (1) in [91].

$$N = n \cdot \frac{t}{t + H} \quad (2.3)$$

Where N is the true number of objects in the sectioned volume ($V = A \cdot t$, A is the area), and n is the observed number of nuclei. Note, $n > N$ since nuclei whose centers are not inside the sectioned volume can still be stained and counted. If we write equation (2.3) for two thicknesses t_1 and t_2 and divide we get:

$$\frac{N_1}{N_2} = \frac{n_1}{n_2} \cdot \left(\frac{t_1}{t_1 + H} \right) / \left(\frac{t_2}{t_2 + H} \right) \quad (2.4)$$

The left-hand side, assuming constant cell volume density in the area of interest, is just the ratio of section thicknesses $N_1/N_2 = t_1/t_2$. Simplifying gives a formula for the ratio of observed planar cell densities n_2/n_1 .

$$\frac{n_2}{n_1} = \frac{t_2 + H}{t_1 + H} \quad (2.5)$$

This gives us the ability to translate between cell density measurements at different section thicknesses. As expected, for $H \ll t_1, t_2$ the observed cell density is just proportional to the thickness. We use $n_1 = 3581 \pm 828$ nuclei/mm² in white matter and $t_1 = 6 \mu\text{m}$ from [89] and a nuclear diameter of $H = 4.7 \mu\text{m}$ from [92]. This measurement for H was from the control arm of an Alzheimer aging study with patients mean age of 88 years. This is a good value to use since the white matter is approximately 80% oligodendrocytes [93]. The next largest fraction is astrocytes which have nuclear sizes comparable to neurons [94]. These measurements give a correction factor of 0.813 at $t_2 = 4 \mu\text{m}$, slightly larger than the naïve correction factor $t_2/t_1 = 2/3$ which does not take into account the nuclear size. Using this, the corrected cell density we used for virtual biopsies is $n_2 = 2912 \pm 673$ nuclei/mm² via equation (2.5). For reference, Roetzer et al. also measured cell density in cortex and tumor regions. Corrected tumor cell density measurements are 4646 ± 1452 nuclei/mm² which is a good comparison for the real biopsy cellularity values. We previously used these corrected normal cell density values in published work [51].

2.2.4 Predictive Modeling

Modeling was performed in R version 3.4.3 [95]. The goal of predictive modeling is to estimate the tissue histologic characteristics of proliferation, cellularity, grade, and cellularity using the imaging inputs. The image features were normalized prior to extraction of values so no further normalization is needed.

2.2.4.1 Biopsy-Paired Repeated Stratified Cross Validation

Cross-validation is the procedure meant to estimate a model's performance on unseen data. In the most basic form, the data is divided into two disjoint partitions. One set is used to fit model parameters (training) and the other set is used to evaluate the accuracy of predictions (testing). An important assumption of cross-validation is that there is data homogeneity between the partitions. In this case, homogeneity means having the same distribution of target values (e.g. grades) in each partition. Without homogeneous partitions, the training data for some folds is not representative of the testing data which will cause an underestimation of true model performance. To avoid this issue, we used stratified resampling to generate folds. Instead of dividing the cases randomly among partitions, equal numbers are sampled from each quartile of the target variable. Stratification has been shown to reduce bias and variance in cross-validation accuracy metrics [96, 97]. The specific procedure is implemented in the `caret` package (v6.0) [98].

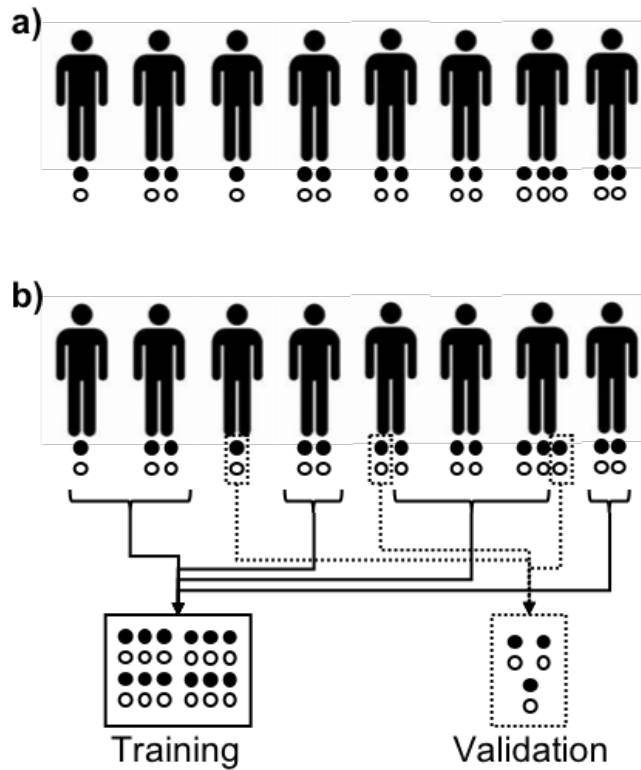


FIGURE 2.2: a) Each patient has 1-3 biopsies (filled) and each each biopsy has a paired virtual (hollow). b) 20% of the biopsies are selected as validation data and the remainder are used for training. Reals are kept in the same set as their paired virtuals.

Data splitting was first applied to real biopsy samples only. Then, each sample’s paired virtual biopsy was placed in the same partition. This balanced the number of real and virtual samples within each partition and prevented data leakage between training and testing sets. In summary: we performed a biopsy-level stratified resampling scheme to generate folds for cross validation. We did not however enforce a patient-level split and allow separate real biopsy sites from the same patient to be split between folds. A schematic of the data splitting is shown in Figure 2.2. Finally, we repeated the five-fold cross validation procedure 500 times and tabulated the performance across multiple rounds. This allowed us to examine the sensitivity of the model performance to fluctuations in data splitting and calculate the range of expected performance metrics.

2.2.4.2 Variable Selection

The first part of the training procedure was to perform variable selection, reducing the total number of variables used for prediction. This removed non-informative variables and reduced the size of the final parameter space. We tested two standard variable selection techniques: First, remove no variables and simply use all variables as inputs. The other was univariate testing, where variables not significantly associated with the outcome were discarded. For continuous outcome variables (Ki67, cell density, ERG) the test statistic was computed based on the Pearson product moment correlation coefficient. This is equivalent to the performing linear regression between the output and a single predictor and testing the significance of the coefficient. All p-values were corrected for multiple comparisons using the Bonferroni method [99].

Additionally, we applied a heuristic variable selection procedure based on random forest variable importance metrics. Each tree in the random forest is trained on a subset of total training data. This means each tree has “out-of-bag” data not used in training that can estimate the out-of-sample performance. As described in Breiman’s original paper [100], variable importance for a particular input m is estimated as follows: For each tree, randomly permute the values of m among the out-of-bag data. This effectively removes any information about the output quantity given by variable m . Then, run the out-of-bag data down the tree and compute the prediction using the shuffled values for variable m . The average change in the error, or misclassification rate for classification problems, across all trees is the variable importance for variable m . A large increase in

error indicates that particular variable was important to the overall performance of the forest.

We performed variable selection as follows: First, we fit a random forest model to all training data using all variables. Then we selected the most important variable from each category of imaging (conventional, diffusion, DSC perfusion, and DCE permeability). The choice to select one predictor from each category is based on the prior knowledge that these different sequence types provide complementary and orthogonal information. For models based on conventional imaging only, the top four most important variables were selected. Using the combined variable selection across all folds risks “leaking” information between training and testing sets because all the data was used at some point in the procedure. To see if this was the case, we compared the predictive ability of this four-variable fixed set was against the performance when using the fold-specific top ranked variables. Other selection methods like principal component analysis, least absolute shrinkage and selection operator (LASSO), and Boruta method were considered but ultimately not tested.

2.2.4.3 Model Types

We tested several types of predictive models to estimate pathology using imaging. For Ki67 and cell density we used Random Forest [100] with parameters: 500 trees, square root of number of predictors tried per split, no max tree size, root-mean-square error metric. We also tried linear models and small dense networks. For predicting tumor grade we used the same random forest parameters except for: number of predictors divided by 3 tried per split, Cohen’s kappa metric [101]. We also tested logistic regression models and neural networks with linear/softmax activation. The description of these models is summarized in Table 2.4.

As an aside, convolutional neural networks are recently the de facto architectures for biomedical image analysis tasks [102, 103]. However, they are not appropriate for this task. The strength of the convolutional network is its ability to use contextual information from kernels to evaluate region segmentations or whole-image classifications. Here, we are interested only in the local image characteristics within a few millimeters of biopsy sites and have distilled the image data into single average intensity values. Thus,

Model	Description	Hyperparameters
Linear Regression (lm)	simple linear regression model	fixed intercept: FALSE
Decision Tree (rpart)	Single decision tree which makes estimates on new data based on the average of training data in terminal nodes	complexity (tuned)
Random Forest (rf)	ensemble of decision trees	500 trees, variables tried per split (tuned)
Neural Network (nnet)	small dense neural network	size (tuned), weight decay (tuned)

TABLE 2.4: Predictive model types tested for predicting pathology

the spatial information has been removed and the data is better suited to the methods described.

2.3 Results

2.3.1 Patient Data, Biopsy Collection, and Pathology Analysis

Thirty-one patients were recruited to the trial. For five patients, tissue harvest was unsuccessful due to surgical complexity. Among the remaining 26 patients, 64 tissue biopsy samples were collected. Further patient exclusions occurred due to a lack of DCE imaging (3 samples), insufficient histology data (2 samples), or lack of samples grade (2 biopsies). Finally, 5 more biopsies were excluded from the final analysis due to VOIs that were outside the imaging field of view (3 samples), lack of analyzable tissue (1 sample), and no tumor grade information (1 sample). For the 23 patients included in the final analysis, 14 were female and 9 were male. The patients' age was 44 ± 17 years (range 21 - 80). The reported ethnicities were 17/2/4/0 white/black/hispanic/asian respectively. After exclusions, 52 real tissue samples from 23 unique patients remained in the final analysis. The clinical grades of the patients were fairly evenly distributed with 7 WHO II, 9 WHO III, and 7 WHO IV. Four of the 7 WHO II gliomas were IDH mutant, as were 7 of 9 WHO III glioma and 2 of 7 WHO IV gliomas. Lastly, 7 of the 52 biopsy samples were collected from contrast enhancing regions. Of those 7 samples, 3 were graded as WHO II, 1 as WHO III, and 3 as WHO IV. Each tissue sample was independently graded by a board-certified Neuropathologist. A few biopsies were graded as II/III based on the pathologist's assessment that the malignancy exceeded a regular WHO II [30]. The sample grades for each biopsy were less than or equal to each tumor's final clinical grade, Table 2.5. However, there was one exception of a WHO grade IV biopsy collected from a clinical WHO grade III tumor. We noted this patient's clinical grade was upgraded to WHO grade IV shortly after the procedure. The proliferation, cellularity and ERG expression roughly increased with sample grade which is expected. The specific values are listed in Table 2.6.

Due to the low number of grade III and II/III samples in the data. We elected to pool the grades III and IV samples into a "higher" grade category with a total of 7 samples. The grade II and II/III samples were pooled into a "lower" grade category with a total of 42 samples, and the virtual biopsies were combined with the three histologically normal samples to form a "normal" category.

Sample Grade:	0	II	II/III	III	IV
Clinical Grade II:	1	12	0	0	0
Clinical Grade III:	2	16	2	0	1
Clinical Grade IV:	0	11	1	2	4

TABLE 2.5: Sample grades II - IV listed by the final clinical grade of the tumor each sample was collected from. Grade 0 indicates non-tumor.

Sample Grade	N	Ki67	Cell density	ERG
0	3	2.417 ± 1.176	1776.714 ± 370.825	1.225 ± 1.444
II	39	4.773 ± 4.346	5789.831 ± 2667.037	2.943 ± 1.598
II/III	3	9.860 ± 2.631	6085.366 ± 2022.448	3.087 ± 1.810
III	2	16.444 ± 12.366	8608.723 ± 8520.914	3.215 ± 0.432
IV	5	25.655 ± 16.185	11546.761 ± 4251.853	4.799 ± 1.647

TABLE 2.6: Number of samples and mean \pm standard deviation for Ki67, cellular density, and ERG for each sample grade. All three roughly increase with increasing sample grade. Grade 0 indicates non-tumor.

2.3.2 Image Data

All sequences were successfully acquired for most patients. One patient had no DCE imaging and three patients had no SWAN imaging as mentioned in Section 2.3.1. All images were successfully skull stripped, co-registered to the T2w image space, and normalized using brain mode and CSF intensity values. Five biopsies had VOIs that were outside the field of view on the SWAN image sequences. For these cases the SWAN values were imputed using the median value among the remaining biopsy samples. After measurement, all biopsies used in further analysis had all 25 measurements in Table 2.3 available.

2.3.3 Variable Selection

Variable selection identified the subset of possible imaging predictors that were significantly associated with each pathology outcome. Then, the predictive power of each subset was tested through modeling and cross-validation. In total, seven variable subsets were tested. Due to the redundancy between ADC from DWI and AvgDC from DTI, we elected to use the ADC measurements only in variable selection.

Subset name	abbreviaton	number of variables	selected within each fold
All	all	24	NO
univariate dynamic	univ.dynamic	varies	YES
univariate fixed	univ.fixed	varies	NO
random forest dynamic	rf.dynamic	4	YES
random forest fixed	rf.fixed	4	NO
conventional TCSE	conv.TCSE	4	NO
conventional TCSPGR	conv.TCSPGR	4	NO

TABLE 2.7: Description of the variable subsets identified through variable selection procedures.

2.3.3.1 Variable Selection for Predicting Proliferation (Ki67)

Univariate analysis showed several imaging variables that were significantly correlated with Ki67. Using the entire data set, the following correlations were significant with Ki67 (Table 2.8): From conventional imaging: T1, TC SE, T2, TC SPGR, T2*. From Diffusion imaging: ADC, eADC, FA. From DSC: CBV, CBF, K2. From DCE: K^{trans} , k_{ep} , v_p , v_e , Wash-In, Wash-Out, AUC, and Peak. These are not surprising given such MR sequences have been developed to help identify regions of heightened malignancy for clinical practice. For use in predictive modeling, the significant of correlation was determined on only the training data for each of the $500 \times 5 = 2500$ folds of cross validation. We found that the same variables were generally significant across folds.

We also used random-forest variable importance based measures to determine the most important variables from each class. This selection was performed independently for each of the 2500 rounds of cross validation. In general, the same few variables were selected as the most important in their class or were near the top. The fraction of times each variable was given a specific rank is listed in Table 2.9. Within each variable class (conventional, diffusion, DSC, DCE), the dominant variables were rank 1 a majority of the time. They are T2 (selected in 89% of folds), FA (99% of folds), CBF (68% of folds), and K^{trans} (61% of folds). These four predictors were selected as a final set of fixed inputs.

2.3.3.2 Variable Selection for Predicting Cellular Density

The variables that were significantly correlated with CD in the entire data set were (Table 2.10): From conventional imaging: T1, TC SE, T2, TC SPGR, FLAIR, T2*.

variable	p value	frequency
T2	1.173E-10	1.00
T1	5.430E-04	0.49
TC SE	2.412E-10	0.99
TC SPGR	2.668E-04	0.62
T2*	3.124E-05	0.97
FLAIR	3.274E-02	0.02
SWAN	8.782E-02	0.00
ADC	4.063E-06	1.00
eADC	1.908E-06	1.00
FA	6.980E-07	1.00
CBV	1.869E-09	0.97
CBF	1.602E-13	1.00
MTT	4.893E-01	0.00
K ₂	3.942E-04	0.67
Delay	4.323E-02	0.01
K^{trans}	2.274E-20	1.00
k_{ep}	4.496E-17	1.00
v_p	8.533E-08	0.95
v_e	1.397E-13	1.00
Wash In	2.021E-10	0.94
Wash Out	5.278E-06	0.96
TTP	5.776E-02	0.01
AUC	2.269E-18	1.00
Peak	2.058E-20	1.00

TABLE 2.8: p-values for variables significantly associated with Ki67 in all data and the fraction of folds (frequency) where each correlation is significant among training data

From Diffusion imaging: ADC, eADC, FA. From DSC: rCBV, CBF, K₂. From DCE: K^{trans} , k_{ep} , v_p , v_e , Wash-In, Wash-Out, AUC, and Peak.

Using random forest variable selection, the top variables in each family were T2 (60% of folds), FA (67% of folds), CBF (67% of folds), and AUC (56% of folds). These four variables were used in the fixed variable set. Table 2.11 lists the specific proportions of ranks for each variable.

2.3.3.3 Variable Selection for Predicting Vascularity (ERG)

The variables that were significantly correlated with ERG in the entire data set were (Table 2.12): From conventional imaging: T1, T2, FLAIR, T2*. From Diffusion imaging: ADC, eADC, FA. From DSC: rCBV, CBF, K₂. From DCE: K^{trans} , k_{ep} , v_p , Wash-In, Wash-Out, and Peak.

variable	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank >6	Average rank
T2	0.85	0.13	0.01	0.00	0.00	0.00	0.00	1.16
T1	0.01	0.10	0.31	0.24	0.17	0.11	0.05	3.99
TC SE	0.14	0.64	0.13	0.05	0.03	0.02	0.01	2.27
TC SPGR	0.00	0.04	0.20	0.25	0.22	0.18	0.11	4.64
T2*	0.00	0.05	0.15	0.19	0.21	0.26	0.15	4.94
FLAIR	0.00	0.04	0.18	0.22	0.25	0.20	0.11	4.72
SWAN	0.00	0.01	0.02	0.05	0.12	0.23	0.57	6.27
ADC	0.00	0.14	0.86	0.00	0.00	0.00	0.00	2.86
eADC	0.01	0.85	0.14	0.00	0.00	0.00	0.00	2.13
FA	0.99	0.01	0.00	0.00	0.00	0.00	0.00	1.01
CBV	0.06	0.38	0.30	0.18	0.08	0.00	0.00	2.84
CBF	0.71	0.15	0.07	0.04	0.04	0.00	0.00	1.54
MTT	0.08	0.16	0.25	0.29	0.22	0.00	0.00	3.40
K ₂	0.03	0.07	0.12	0.24	0.53	0.00	0.00	4.16
Delay	0.12	0.23	0.26	0.26	0.13	0.00	0.00	3.05
K^{trans}	0.61	0.27	0.06	0.03	0.01	0.01	0.00	1.59
k_{ep}	0.26	0.54	0.12	0.05	0.02	0.01	0.01	2.07
v_p	0.00	0.00	0.00	0.00	0.04	0.13	0.83	7.50
v_e	0.01	0.01	0.04	0.08	0.33	0.36	0.17	5.53
Wash In	0.00	0.00	0.00	0.02	0.08	0.12	0.77	7.50
Wash Out	0.03	0.02	0.06	0.11	0.27	0.25	0.26	5.54
TTP	0.00	0.00	0.00	0.00	0.01	0.05	0.93	8.09
AUC	0.00	0.02	0.14	0.54	0.21	0.07	0.02	4.23
Peak	0.09	0.13	0.57	0.17	0.04	0.01	0.00	2.97

TABLE 2.9: Fraction of times each variable was assigned a particular rank relative to other variables in its class for predicting Ki67

Using random forest variable importance, the highest ranking variables from each family were T2 (99% of folds), ADC (69% of folds), CBV (52% of folds), and AUC (95% of folds). These four variables were used in the fixed variable set. Table 2.13 lists the specific proportions of ranks for each variable.

2.3.3.4 Variable Selection for Predicting Local Grade

The variables that were significantly correlated with grade in the entire data set were (Table 2.14): From conventional imaging: T1, TC SPGR, T2, FLAIR, T2*, SWAN. From Diffusion imaging: ADC, eADC, FA. From DSC: rCBV, CBF, K₂. From DCE: K^{trans} , k_{ep} , v_p , v_e , AUC, and Peak.

Using random forest variable importance, the highest ranking variables from each family were T2 (99% of folds), ADC (99% of folds), CBV (49% of folds), and k_{ep} (52% of folds).

variable	p value	frequency
T2	3.503E-10	1.00
T1	2.684E-04	0.69
TC SE	2.782E-09	1.00
TC SPGR	3.809E-05	0.80
T2*	1.385E-05	0.99
FLAIR	1.884E-03	0.24
SWAN	5.293E-01	0.00
ADC	2.650E-07	1.00
eADC	2.006E-05	1.00
FA	6.340E-06	1.00
CBV	2.168E-08	1.00
CBF	1.548E-09	1.00
MTT	8.575E-02	0.00
K ₂	3.282E-07	0.98
Delay	6.554E-01	0.00
K^{trans}	1.141E-14	1.00
k_{ep}	2.160E-13	1.00
v_p	3.794E-08	0.98
v_e	1.681E-10	0.99
Wash In	9.920E-06	0.87
Wash Out	6.092E-04	0.51
TTP	1.096E-01	0.00
AUC	5.068E-15	1.00
Peak	3.074E-16	1.00

TABLE 2.10: p-values for variables significantly associated with CD in all data and the fraction of folds (frequency) where each correlation is significant among training data

These four variables were used in the fixed variable set. Table 2.15 lists the specific proportions of ranks for each variable.

variable	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank >6	Average rank
T2	0.60	0.34	0.05	0.00	0.00	0.00	0.00	1.45
T1	0.00	0.00	0.08	0.29	0.23	0.22	0.18	5.12
TC SE	0.30	0.38	0.26	0.04	0.01	0.00	0.00	2.10
TC SPGR	0.09	0.27	0.50	0.10	0.03	0.01	0.00	2.73
T2*	0.00	0.00	0.03	0.17	0.24	0.27	0.27	5.57
FLAIR	0.00	0.00	0.01	0.14	0.22	0.26	0.37	5.83
SWAN	0.00	0.01	0.06	0.26	0.27	0.24	0.17	5.20
ADC	0.10	0.32	0.58	0.00	0.00	0.00	0.00	2.48
eADC	0.23	0.49	0.28	0.00	0.00	0.00	0.00	2.06
FA	0.67	0.20	0.14	0.00	0.00	0.00	0.00	1.47
CBV	0.03	0.26	0.44	0.21	0.06	0.00	0.00	3.02
CBF	0.67	0.25	0.06	0.02	0.00	0.00	0.00	1.43
MTT	0.00	0.02	0.09	0.32	0.57	0.00	0.00	4.44
K ₂	0.27	0.35	0.23	0.12	0.03	0.00	0.00	2.29
Delay	0.04	0.11	0.18	0.34	0.33	0.00	0.00	3.82
K ^{trans}	0.05	0.24	0.37	0.29	0.04	0.01	0.00	3.05
k _{ep}	0.35	0.36	0.19	0.08	0.01	0.00	0.00	2.06
v _p	0.00	0.00	0.01	0.01	0.18	0.30	0.50	6.61
v _e	0.00	0.01	0.01	0.04	0.41	0.25	0.28	5.90
Wash In	0.00	0.01	0.01	0.02	0.20	0.18	0.57	6.82
Wash Out	0.00	0.00	0.00	0.00	0.04	0.11	0.85	7.90
TTP	0.00	0.00	0.00	0.00	0.06	0.15	0.79	7.53
AUC	0.56	0.21	0.15	0.07	0.01	0.00	0.00	1.76
Peak	0.04	0.17	0.25	0.48	0.05	0.01	0.00	3.37

TABLE 2.11: Fraction of times each variable was assigned a particular rank relative to other variables in its class for predicting CD

variable	p value	frequency
T2	1.691E-15	1.00
T1	6.595E-07	1.00
TC SE	5.314E-03	0.17
TC SPGR	8.498E-02	0.00
T2*	1.179E-05	0.96
FLAIR	6.251E-09	1.00
SWAN	2.967E-02	0.00
ADC	3.926E-13	1.00
eADC	3.065E-10	1.00
FA	9.118E-14	1.00
CBV	1.451E-05	0.96
CBF	1.856E-04	0.77
MTT	2.307E-01	0.00
K ₂	2.462E-06	1.00
Delay	1.715E-01	0.00
K ^{trans}	1.860E-03	0.25
k _{ep}	5.217E-07	1.00
v _p	1.312E-04	0.79
v _e	7.359E-03	0.04
Wash In	1.528E-04	0.87
Wash Out	1.499E-04	0.82
TTP	6.617E-01	0.00
AUC	2.111E-03	0.19
Peak	6.229E-06	1.00

TABLE 2.12: p-values for variables significantly associated with ERG in all data and the fraction of folds (frequency) where each correlation is significant among training data

variable	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank >6	Average rank
T2	0.99	0.01	0.00	0.00	0.00	0.00	0.00	1.01
T1	0.00	0.03	0.08	0.28	0.25	0.21	0.15	4.99
TC SE	0.00	0.00	0.05	0.22	0.28	0.27	0.18	5.31
TC SPGR	0.00	0.00	0.02	0.14	0.19	0.26	0.39	5.85
T2*	0.01	0.73	0.22	0.03	0.01	0.00	0.00	2.31
FLAIR	0.00	0.22	0.58	0.14	0.04	0.01	0.00	3.05
SWAN	0.00	0.01	0.05	0.18	0.22	0.24	0.28	5.48
ADC	0.69	0.26	0.05	0.00	0.00	0.00	0.00	1.36
eADC	0.22	0.56	0.22	0.00	0.00	0.00	0.00	2.00
FA	0.10	0.18	0.73	0.00	0.00	0.00	0.00	2.63
CBV	0.52	0.35	0.11	0.02	0.00	0.00	0.00	1.64
CBF	0.07	0.23	0.52	0.18	0.00	0.00	0.00	2.83
MTT	0.01	0.05	0.18	0.70	0.05	0.00	0.00	3.72
K ₂	0.40	0.37	0.18	0.04	0.00	0.00	0.00	1.87
Delay	0.00	0.00	0.00	0.05	0.94	0.00	0.00	4.94
K ^{trans}	0.00	0.02	0.14	0.28	0.27	0.20	0.08	4.76
k _{ep}	0.00	0.01	0.09	0.17	0.26	0.28	0.20	5.37
v _p	0.00	0.14	0.42	0.22	0.12	0.07	0.04	3.68
v _e	0.00	0.04	0.18	0.22	0.22	0.20	0.13	4.76
Wash In	0.00	0.00	0.00	0.00	0.01	0.02	0.97	8.43
Wash Out	0.00	0.00	0.01	0.02	0.03	0.07	0.86	7.72
TTP	0.00	0.00	0.01	0.03	0.07	0.16	0.72	6.95
AUC	0.95	0.05	0.00	0.00	0.00	0.00	0.00	1.05
Peak	0.05	0.73	0.14	0.05	0.02	0.01	0.00	2.30

TABLE 2.13: Fraction of times each variable was assigned a particular rank relative to other variables in its class for predicting ERG

variable	p value	frequency
T2	9.198E-16	1.00
T1	1.455E-04	0.75
TC SE	2.638E-03	0.20
TC SPGR	3.356E-05	0.91
T2*	7.077E-09	1.00
FLAIR	2.539E-11	1.00
SWAN	1.931E-04	0.70
ADC	9.601E-16	1.00
eADC	3.587E-11	1.00
FA	2.008E-14	1.00
CBV	3.541E-04	0.51
CBF	5.359E-04	0.41
MTT	1.621E-01	0.00
K ₂	1.577E-06	1.00
Delay	7.209E-01	0.00
K ^{trans}	6.716E-09	1.00
k _{ep}	6.639E-10	1.00
v _p	2.215E-07	1.00
v _e	9.228E-09	1.00
Wash In	4.540E-02	0.02
Wash Out	4.949E-03	0.06
TTP	3.038E-01	0.00
AUC	2.416E-11	1.00
Peak	4.870E-10	1.00

TABLE 2.14: p-values for variables significantly associated with grade in all data and the fraction of folds (frequency) where each correlation is significant among training data

variable	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank >6	Average rank
T2	0.99	0.01	0.00	0.00	0.00	0.00	0.00	1.01
T1	0.00	0.00	0.00	0.02	0.10	0.45	0.42	6.28
TC SE	0.00	0.00	0.05	0.26	0.50	0.14	0.04	4.85
TC SPGR	0.00	0.00	0.26	0.44	0.23	0.05	0.02	4.12
T2*	0.00	0.03	0.66	0.25	0.06	0.00	0.00	3.36
FLAIR	0.01	0.96	0.03	0.00	0.00	0.00	0.00	2.02
SWAN	0.00	0.00	0.00	0.03	0.11	0.35	0.52	6.35
ADC	0.99	0.01	0.00	0.00	0.00	0.00	0.00	1.01
eADC	0.00	0.66	0.34	0.00	0.00	0.00	0.00	2.33
FA	0.01	0.33	0.66	0.00	0.00	0.00	0.00	2.66
CBV	0.49	0.39	0.11	0.01	0.00	0.00	0.00	1.64
CBF	0.04	0.20	0.63	0.12	0.02	0.00	0.00	2.87
MTT	0.00	0.02	0.11	0.59	0.28	0.00	0.00	4.12
K ₂	0.47	0.38	0.13	0.02	0.00	0.00	0.00	1.71
Delay	0.00	0.01	0.03	0.26	0.70	0.00	0.00	4.65
K ^{trans}	0.24	0.34	0.22	0.13	0.06	0.00	0.00	2.45
k _{ep}	0.52	0.28	0.12	0.06	0.02	0.00	0.00	1.76
v _p	0.00	0.00	0.00	0.02	0.10	0.80	0.07	5.92
v _e	0.04	0.10	0.17	0.29	0.34	0.06	0.00	4.00
Wash In	0.00	0.00	0.00	0.00	0.00	0.02	0.98	7.87
Wash Out	0.00	0.00	0.00	0.00	0.00	0.00	1.00	8.52
TTP	0.00	0.00	0.00	0.00	0.00	0.05	0.95	7.53
AUC	0.17	0.22	0.34	0.21	0.06	0.00	0.00	2.78
Peak	0.03	0.06	0.14	0.30	0.41	0.06	0.00	4.17

TABLE 2.15: Fraction of times each variable was assigned a particular rank relative to other variables in its class for predicting grade

2.3.4 Predictive Modeling

For each target variable: Ki67, cell density, ERG, and local grade, We performed 500 repetitions of five-fold cross validation with stratified sampling. Within each round, each model was trained on four folds and used to predict on the fifth fold. The R^2 correlation and root mean square error (RMSE) metrics were then calculated between the predicted and observed values.

2.3.4.1 Predictive Modeling of Proliferation (Ki67)

The 2500 metric values are summarized in Tables 2.16. A higher R^2 and lower RMSE indicate better predictions. The random forest trained on four fixed variables (T2, FA, K^{trans} , CBF) had the best overall performance with an R^2 value of 0.709 and RMSE of 3.78 Ki67 percentage points. Using the variables selected with each fold lead to slightly reduced performance ($R^2 = 0.655$, RMSE = 4.15). Selecting variables within each fold removed the possible bias from overfitting which comes from using all the data in the variable selection. This means that a final model using the fixed variables probably has a true error of between 3.72 and 4.11 percentage points. The distribution of the metric values are shown in Figure 2.3. Interestingly, when we selected univariate significant predictors using the entire data set or the fold-specific training data we observed the opposite trend. Most models had better performance using the fold-specific variable selections rather than the globally significant variables. There was a non-significant decrease in R^2 of about 0.05.

Overall, these results show high predictability of cellular proliferation (Ki67) using imaging data. The high R^2 value around 0.65 means that the combination of conventional, diffusion, perfusion, and permeability imaging (T2, FA, CBF, K^{trans}) provides a lot of the requisite information. The RMSE of 4.15 is also much smaller than the total range of Ki67 values observed in the data of about 40 percentage points, indicating precise predictions. Furthermore, we can see that much of the information comes from the advanced imaging. Using conventional imaging sequences only we found the best performance dropped to $R^2 = 0.558$ using a linear model trained on TC SE, T1, T2, and FLAIR. Using random forest the R^2 was 0.517. While this is still good predictive

performance for a biological target variable, the decreased accuracy must be noted since these conventional-only models are used in subsequent chapters for survival modeling.

Panel A							
Model	all	conv.TCSE	conv.TCSPGR	rf.dynamic	rf.fixed	univ.dynamic	univ.fixed
lm	0.582 ± 0.238	0.558 ± 0.207	0.521 ± 0.192	0.582 ± 0.233	0.642 ± 0.216	0.552 ± 0.244	0.539 ± 0.241
nnet	0.354 ± 0.245	0.461 ± 0.258	0.483 ± 0.251	0.682 ± 0.230	0.707 ± 0.197	0.532 ± 0.225	0.157 ± 0.194
rf	0.668 ± 0.206	0.517 ± 0.225	0.486 ± 0.190	0.655 ± 0.223	0.709 ± 0.179	0.678 ± 0.200	0.663 ± 0.206
rpart	0.397 ± 0.243	0.378 ± 0.243	0.387 ± 0.246	0.485 ± 0.236	0.471 ± 0.239	0.446 ± 0.234	0.397 ± 0.243
Panel B							
Model	all	conv.TCSE	conv.TCSPGR	rf.dynamic	rf.fixed	univ.dynamic	univ.fixed
lm	5.404 ± 2.163	4.891 ± 1.253	5.189 ± 1.439	4.650 ± 1.484	4.263 ± 1.264	5.321 ± 2.468	5.504 ± 2.476
nnet	6.592 ± 2.241	5.936 ± 2.029	5.884 ± 1.891	3.898 ± 1.616	3.655 ± 0.989	4.943 ± 1.586	7.048 ± 2.197
rf	4.200 ± 1.653	5.018 ± 1.745	5.284 ± 1.684	4.151 ± 1.865	3.778 ± 1.455	4.077 ± 1.542	4.217 ± 1.615
rpart	5.999 ± 2.133	5.998 ± 1.842	5.880 ± 1.763	5.514 ± 1.974	5.571 ± 2.016	5.832 ± 2.020	5.997 ± 2.131

TABLE 2.16: Panel A: R^2 and Panel B: RMSE values for predicting Ki67, values are mean ± standard deviation. lm = linear model, nnet = neural network, rf = random forest, rpart = decision tree.

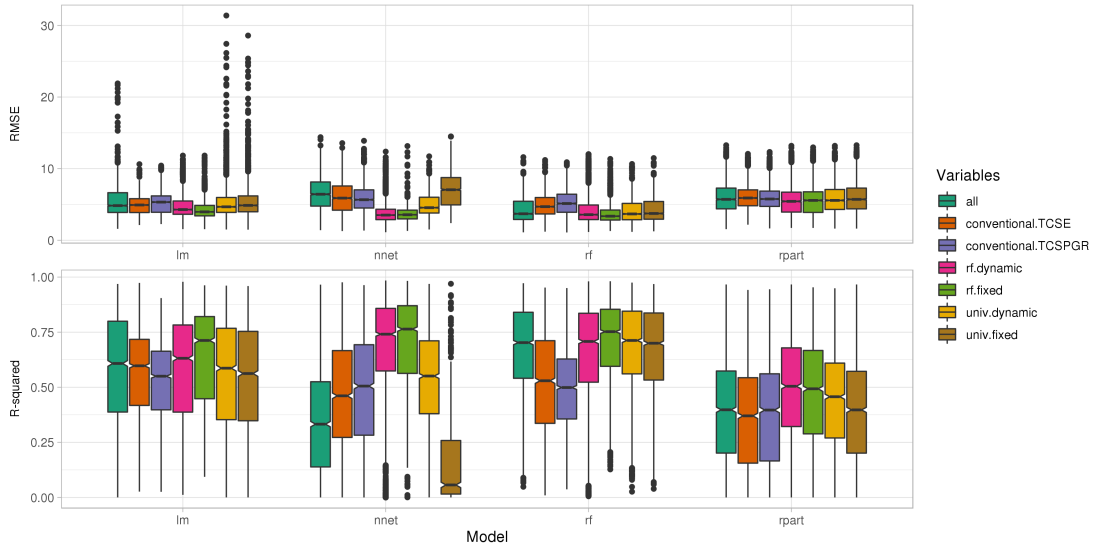


FIGURE 2.3: Metric values for cross-validated prediction of Ki67

2.3.4.2 Predictive Modeling of Cellular Density

All 2500 metric values for predicting cellularity are summarized in Table 2.17. Overall, the random forest trained on the univariate significant variables in each fold had the best performance with $R^2 = 0.592$ and $\text{RMSE} = 1948$ nuclei/ mm^2 . Using a fixed set of 20 univariate variables had comparable performance with $R^2 = 0.585$. We also found acceptable performance with a random forest trained on just four variables: T2, CBF, k_{ep} , and AUC. The R^2 value was 0.567. The small reduction in R^2 is worth using 16 fewer variables in the predictive model since future clinical purposes make requiring 20 inputs infeasible. When the random forest variables were selected within each fold

of cross validation the performance of the random forest decreased considerably from 0.567 to 0.471. This suggests there may be a good deal of overfitting by using all data in variable selection. Plots of the metric values are shown in Figure 2.4.

While not as high as the R^2 value for predicting proliferative index, the R^2 of 0.567 also shows good predictive ability of cell density using imaging data. When we used conventional imaging only we saw almost no decrease in predictive performance. Using T1, T2, TC SE, and FLAIR we found cross-validated R^2 of 0.550. This means that the benefit of the advanced imaging from diffusion, DSC, and DCE sequences is smaller. This makes sense because cellularity is a physical property which is observable using conventional sequences like T2-weighted images.

Panel A							
Model	all	conv.TCSE	conv.TCSPGR	rf.dynamic	rf.fixed	univ.dynamic	univ.fixed
lm	0.452 \pm 0.225	0.508 \pm 0.199	0.527 \pm 0.174	0.486 \pm 0.226	0.554 \pm 0.203	0.460 \pm 0.222	0.482 \pm 0.224
nnet	0.235 \pm 0.173	0.432 \pm 0.204	0.407 \pm 0.197	0.441 \pm 0.197	0.303 \pm 0.196	0.345 \pm 0.144	0.229 \pm 0.157
rf	0.577 \pm 0.179	0.550 \pm 0.171	0.523 \pm 0.161	0.471 \pm 0.198	0.567 \pm 0.165	0.593 \pm 0.174	0.585 \pm 0.173
rpart	0.375 \pm 0.194	0.508 \pm 0.192	0.482 \pm 0.207	0.343 \pm 0.170	0.400 \pm 0.178	0.393 \pm 0.195	0.381 \pm 0.200

Panel B							
Model	all	conv.TCSE	conv.TCSPGR	rf.dynamic	rf.fixed	univ.dynamic	univ.fixed
lm	2332.509 \pm 529.136	2143.360 \pm 391.281	2115.458 \pm 371.488	2154.098 \pm 357.335	1988.118 \pm 331.659	2284.640 \pm 588.994	2222.568 \pm 497.964
nnet	2809.088 \pm 483.159	2390.925 \pm 534.334	2459.946 \pm 507.150	2296.573 \pm 400.057	2616.314 \pm 509.012	2511.036 \pm 416.427	2795.429 \pm 465.630
rf	1989.849 \pm 340.517	2021.873 \pm 333.837	2124.614 \pm 410.242	2192.233 \pm 363.824	1987.769 \pm 299.566	1947.562 \pm 335.103	1986.690 \pm 324.660
rpart	2545.829 \pm 495.808	2192.538 \pm 411.021	2265.952 \pm 462.473	2573.839 \pm 456.202	2438.610 \pm 479.348	2485.580 \pm 465.812	2524.822 \pm 499.977

TABLE 2.17: Panel A: R^2 and Panel B: RMSE values for predicting CD, values are mean \pm standard deviation. lm = linear model, nnet = neural network, rf = random forest, rpart = decision tree.

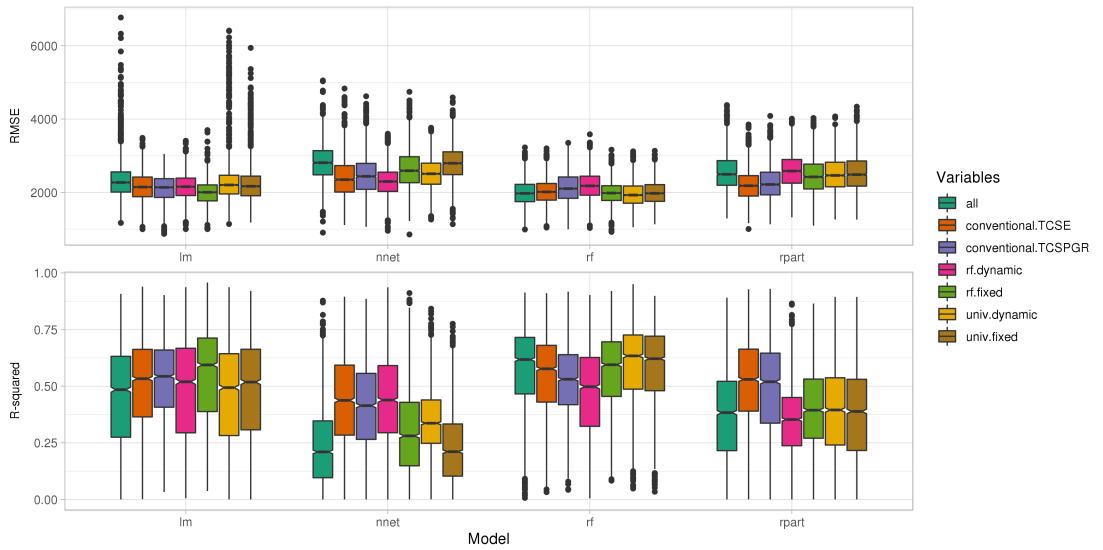


FIGURE 2.4: Metric values for cross-validated prediction of CD

2.3.4.3 Predictive Modeling of Vascularity (ERG)

Similar to Ki67 and CD, the best predictive performance for ERG was the random forest. While the absolute best performance was with a forest using all 23 input variables ($R^2 = 0.677$), we found similar results could be achieved with a smaller four variable set composed of T2, ADC, CBV, and AUC. The R^2 using this set was 0.651. Using a different set of random forest selected variables within each fold had slightly decreased performance ($R^2 = 0.630$ for random forest) compared to the fixed variables which might indicate a small amount of overfitting. Plots of the metric values are shown in Figure 2.5.

Like proliferation, the high R^2 shows how well imaging data can predict for the ERG vascularity marker. Intuitively, we would expect perfusion and permeability imaging to greatly improve the predictability of ERG since they directly image the interaction of vascular contrast agents with tissue. However, we found that we could achieve essentially the same predictive performance using conventional imaging only (T1, TC SE, T2, FLAIR). This is probably due to the use of T1 contrast enhanced image which highlights leaky vasculature. Using the random forest variable importance, we saw the spin-echo contrast enhanced image was ranked as the top conventional variable in 30% of folds.

Panel A							
Model	all	conv.TCSE	conv.TCSPGR	rf.dynamic	rf.fixed	univ.dynamic	univ.fixed
lm	0.428 ± 0.176	0.521 ± 0.110	0.581 ± 0.099	0.509 ± 0.144	0.548 ± 0.154	0.561 ± 0.135	0.578 ± 0.139
nnet	0.450 ± 0.195	0.567 ± 0.092	0.594 ± 0.092	0.665 ± 0.091	0.636 ± 0.130	0.635 ± 0.112	0.548 ± 0.149
rf	0.677 ± 0.075	0.658 ± 0.096	0.644 ± 0.095	0.630 ± 0.099	0.651 ± 0.086	0.650 ± 0.085	0.650 ± 0.084
rpart	0.582 ± 0.120	0.556 ± 0.114	0.552 ± 0.100	0.648 ± 0.106	0.626 ± 0.117	0.582 ± 0.100	0.533 ± 0.113
Panel B							
Model	all	conv.TCSE	conv.TCSPGR	rf.dynamic	rf.fixed	univ.dynamic	univ.fixed
lm	1.687 ± 0.610	1.328 ± 0.184	1.245 ± 0.174	1.398 ± 0.393	1.337 ± 0.399	1.331 ± 0.510	1.319 ± 0.486
nnet	1.434 ± 0.300	1.241 ± 0.171	1.207 ± 0.162	1.091 ± 0.185	1.134 ± 0.217	1.150 ± 0.204	1.282 ± 0.244
rf	1.094 ± 0.177	1.101 ± 0.202	1.121 ± 0.193	1.152 ± 0.190	1.119 ± 0.178	1.115 ± 0.172	1.119 ± 0.172
rpart	1.229 ± 0.220	1.269 ± 0.216	1.270 ± 0.193	1.119 ± 0.201	1.171 ± 0.221	1.225 ± 0.179	1.310 ± 0.199

TABLE 2.18: Panel A: R^2 and Panel B: RMSE values for predicting ERG, values are mean ± standard deviation. lm = linear model, nnet = neural network, rf = random forest, rpart = decision tree.

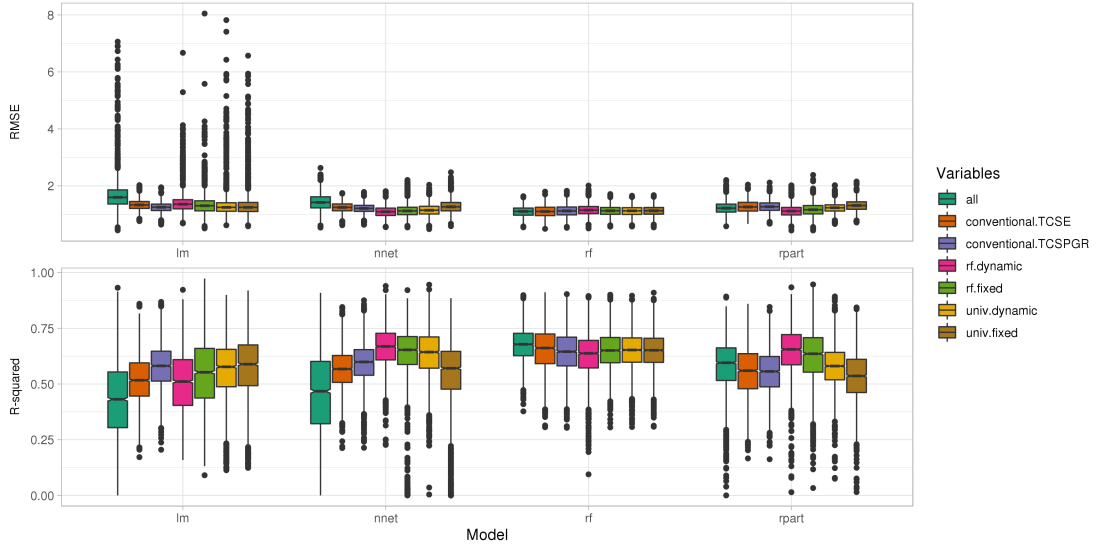


FIGURE 2.5: Metric values for cross-validated prediction of ERG

2.3.4.4 Predictive Modeling of Local Grade

The random forest had very high accuracy and kappa values for predicting tumor grade, binned as normal, lower, and higher grade, using any of the variable subsets. The best performance was achieved using univariate selected variables within each fold (kappa = 0.903) but excellent performance was also achieved using a small set of fixed variables chosen by random forest variable selection. They were T2, ADC, CBV, and k_{ep} and had a performance of kappa = 0.895. The full set of metrics are listed in Table 2.19 and plotted in Figure 2.6.

High accuracy and high kappa values means that imaging has high predictive power over the local biopsy grade even in the presence of dataset imbalance towards lower grade samples. While advanced sequences like diffusion or DCE have been shown to correlate with malignant disease [104], we found that they were not necessarily required to get good prediction of tumor grade. Using just T1, T2, TC SE, and FLAIR we still achieved over 90% overall accuracy and kappa of 0.822. Much of the information related to high grade is likely captured by contrast enhancement whereas the T2 weighted and FLAIR images help separate the normal and low-grade samples.

Panel A							
Model	all	conv.TCSE	conv.TCSPGR	rf.dynamic	rf.fixed	univ.dynamic	univ.fixed
nnet	0.835 \pm 0.076	0.913 \pm 0.058	0.898 \pm 0.060	0.924 \pm 0.046	0.903 \pm 0.053	0.892 \pm 0.054	0.856 \pm 0.069
rf	0.945 \pm 0.041	0.904 \pm 0.053	0.890 \pm 0.058	0.937 \pm 0.046	0.943 \pm 0.045	0.948 \pm 0.041	0.946 \pm 0.041
rpart	0.851 \pm 0.070	0.856 \pm 0.058	0.882 \pm 0.061	0.883 \pm 0.059	0.871 \pm 0.066	0.872 \pm 0.056	0.855 \pm 0.068
Panel B							
Model	all	conv.TCSE	conv.TCSPGR	rf.dynamic	rf.fixed	univ.dynamic	univ.fixed
nnet	0.697 \pm 0.140	0.839 \pm 0.108	0.812 \pm 0.111	0.861 \pm 0.084	0.818 \pm 0.102	0.803 \pm 0.099	0.736 \pm 0.125
rf	0.898 \pm 0.076	0.822 \pm 0.099	0.800 \pm 0.105	0.885 \pm 0.085	0.895 \pm 0.084	0.903 \pm 0.076	0.901 \pm 0.076
rpart	0.736 \pm 0.118	0.739 \pm 0.100	0.790 \pm 0.106	0.786 \pm 0.107	0.771 \pm 0.113	0.765 \pm 0.102	0.742 \pm 0.115

TABLE 2.19: Panel A: Accuracy and Panel B: Kappa values for predicting grade, values are mean \pm standard deviation. lm = linear model, nnet = neural network, rf = random forest, rpart = decision tree.

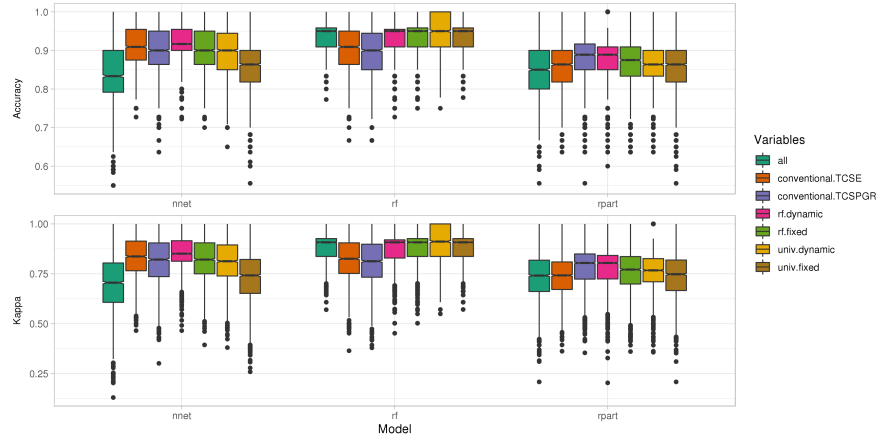


FIGURE 2.6: Metric values for cross-validated prediction of tumor grade

2.3.5 Graphical Synthetic Pathology Maps

The goal of using spatially specific tissue samples and models estimating local pathology is to use those models to guide diagnosis and treatment. Towards this goal, we applied the best performing models voxel-wise throughout the brain to generate “maps” of estimated pathology. Examples are shown in Figure 2.7 and Figure 2.8 where the predictions are masked and shown only within the visible lesion. Qualitatively, we see high proliferative activity is predicted inside the contrast enhancing region which is expected. Although, even within contrast enhancement there is a range of proliferative values as indicated by the orange and yellow colors. This suggests that there is additional information present beyond the presence or absence of enhancement. In a low-grade case (Figure 2.8), there is slightly elevated proliferation inside the core of the tumor relative to the periphery. A similar situation is shown in Figure 2.9 which illustrates pockets of highly cellular disease within the tumor. Figure 2.10 shows the strong correlation

between contrast enhancement and the vascularity marker ERG. ERG values are uniformly elevated in the enhancing region and overall lower elsewhere. Finally, Figure 2.11 illustrates the heterogeneity present within a single tumor. Most of the clinically high grade tumor is estimated as low grade disease. However, there is a small focus of predicted high grade disease within the tumor core. We can also see predictions of normal brain on the periphery of the radiographically visible lesion.

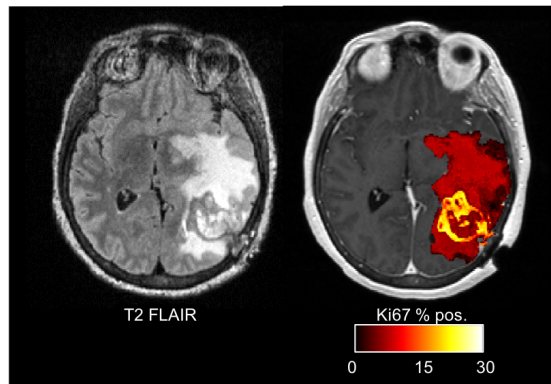


FIGURE 2.7: Map of estimated Ki67 in a WHO IV glioblastoma patient alongside a T2-FLAIR image for reference. The map was generated with conventional imaging only and has been smoothed by a 1 mm gaussian kernel.

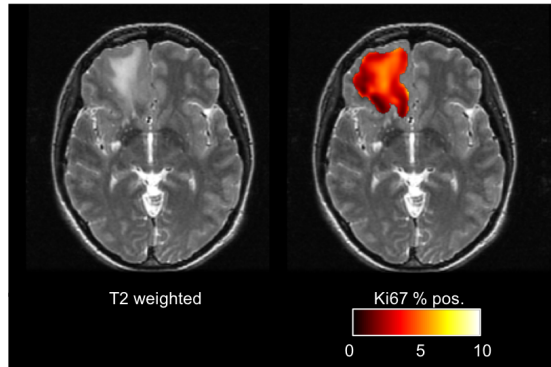


FIGURE 2.8: Map of estimated Ki67 in a WHO II glioma patient alongside a T2-weighted image for reference. The map was generated with conventional and advanced imaging.

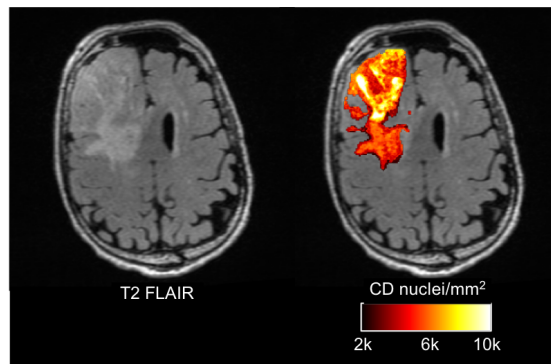


FIGURE 2.9: Map of estimated Cellular density (CD) in a WHO IV glioblastoma patient alongside a T2-FLAIR image for reference. The map was generated with conventional imaging only.

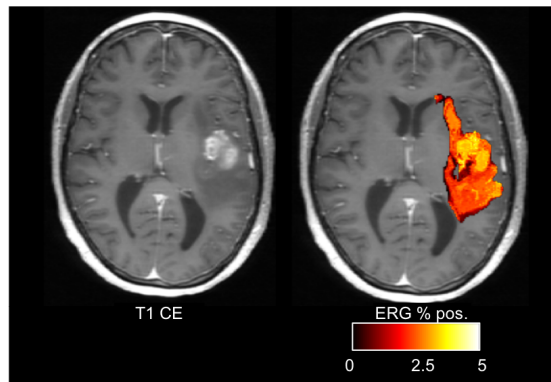


FIGURE 2.10: Map of estimated ETS related gene (ERG) in a WHO IV glioblastoma patient alongside The T1 post-contrast image for reference. The map was generated with conventional imaging only.

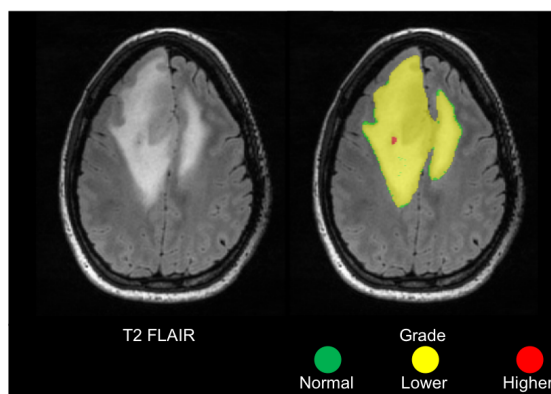


FIGURE 2.11: Map of estimated tumor grade in a WHO IV glioblastoma patient alongside a T2 FLAIR image for reference. The map was generated with conventional imaging only and is smoothed by a radius 1 median filter.

2.4 Discussion

We tested a variety of machine learning methods and variable selection techniques to estimate local tumor pathology in terms of proliferation (Ki67), cellular density (CD), vascularity (ERG), and histologic grade. We achieved good performance using a random forest for predicting Ki67 ($R^2 = 0.709$), ERG ($R^2 = 0.651$), cell density ($R^2 = 0.567$), and tumor grade (Accuracy = 94.3%). In general, the best or comparable performance was achieved using a random forest model trained on the best variable from each sequence family or the best four conventional variables. We used these random forest models moving forward in subsequent chapters of this work.

Previous results using this image guided biopsy data set to estimate Ki67 have been previously reported [49]. However, the results presented here are different and extended in a few key ways: First, the robustness of cross-validation (CV) is improved by using 500 repetitions instead of a single five-fold repetition. This also removes some uncertainty associated with selecting the best predictors with random forest because more than 5 votes are cast for each image type. Although, ultimately the same final variables (T2, FA, CBF, K^{trans}) were selected. Second, the intensity normalization scheme for anatomic imaging in [49] uses hand-drawn ROIs in gray matter, white matter, and CSF whereas the current work used the normal brain mode and automatic CSF ROIs. Overall the predictive results are similar. Previous work reports R^2 values of 0.749 ± 0.137 as the average of a single five-fold CV. Here, we used 500 repetitions of five-fold CV and measured $R^2 = 0.709 \pm 0.179$. While the average is smaller, the ranges of R^2 values still overlap considerably.

Similar studies have also used image guided biopsy data to correlate or estimate imaging and proliferative activity. Barajas et al. [38] found DSC perfusion metrics like CBV correlated with proliferation in contrast enhancing regions. In non-enhancing regions, they also found inverse correlations between proliferation, diffusion weighted imaging, and FLAIR intensity. Our correlation analysis agrees with these results for perfusion metrics, diffusion metrics, and contrast enhancement. But, we did not find a significant correlation between FLAIR intensity and proliferation. A possible explanation for this might be a difference in intensity normalization scheme: brain mode - CSF versus in [38] where the FLAIR intensity was scaled based on white matter. It may also be due to the use of virtual biopsies in our work. In another study, Price et al. [61] found significant

correlation between rCBV metrics (mean, max, 75th percentile, 90th percentile) and proliferation. This agrees with our results as well, as we found both CBV and CBF were significantly associated with Ki67, including in more than 97% of cross-validation folds. More recently, Autry et al. [63] correlated conventional MRI, DSC, and MR spectroscopy with histopathologic parameters in 100 tissue samples from glioblastomas. Surprisingly, they found no difference in proliferation between enhancing and non-enhancing samples, although this may be because the patient population was limited to high-grade gliomas. Correlation between proliferation and other image metrics were not reported.

Several previous studies have examined the relationship between cellularity and local MR image characteristics [38, 43, 51, 62, 105, 106]. Overall, they find correlation between T1 contrast enhancement, DSC, and diffusion metrics with cellularity but no one study other than our on previous work [51] examines conventional, DWI, DSC, and DCE simultaneously. The study by Chang et al. [62] perform predictive modeling for CD using conventional and diffusion metrics and found a reasonable R^2 around 0.55. The study by Durst et al. [105] used conventional, diffusion, and DSC to predict CD with an in-sample $R^2 = 0.49$. These are comparable to the R^2 in this work of 0.567.

One of the best uses for estimating cellularity is to assess tumor infiltration into peritumoral regions [20, 89]. This can be done using predictive models applied voxel-wise to “map” the cellularity across or just outside a visible lesion. Such maps have been shown by Durst et al. [105], Akbari et al. [43], Chang et al. [62], and our own previous work [51]. To our knowledge, none of these maps have been prospectively clinically validated.

In this work we were able to discriminate between normal, lower, and higher grade glioma samples using a combination of conventional and advanced MR imaging. As advanced imaging techniques like diffusion, DSC, or DCE mature, there have been several studies correlation quantitative findings like ADC or K^{trans} with tumor malignancy [104, 107]. However, both image metrics and tissue histology (i.e. grade) are heterogeneous which confounds whole-tumor based analysis. Overall however, there is evidence that diffusion, perfusion, and permeability imaging can indicate overall grade or IDH status [108, 109].

In the study by Barajas et al. spatially specific biopsies were used to correlate local CBV and ADC with histologic markers of high grade disease [38]. As expected, high CBV and low ADC were markers of heightened tumor infiltration. We found the same metrics,

CBV and peak height, were significantly associated with sample grade in addition to several other image features in Table 2.14.

Among our tested model classes, we elected not to include convolutional neural networks. Convolutional neural networks (CNN) have recently become the de-facto model for machine learning in medical imaging [102, 103]. The strength of the CNN comes from the encoding of relevant spatial information via the use of convolutional kernels. This makes them especially powerful for whole-image classification or segmentation tasks. In our work, we use spatially specific biopsy samples where ground truth is only available for a very small volume of interest. While the spatial specificity allows voxel-wise mapping of estimated pathology (Section 2.3.5), it means that neighborhood image characteristics may not necessarily hold useful information. Thus, a CNN is not likely to perform better than a random forest trained on the mean image intensity inside the VOI as we have done here. Furthermore, CNNs add considerable complexity and computational time to train. For these reasons, we elected not to explore CNNs for predicting the biopsy-level pathology.

2.4.1 Limitations

This work estimated the local histologic characteristics of gliomas using magnetic resonance imaging. Recently, the importance of genetic and molecular characteristics has surpassed histology in the diagnosis and classification of gliomas [20]. While mutations like IDH1 [86] would be interesting to analyze using spatially specific sampling, the results may not be as useful seeing as IDH1 mutations are homogeneously expressed throughout a glioma [87]. Analyzing which other molecular markers might be heterogeneously expressed and to what degree they correlate with imaging findings would be highly interesting. Unfortunately, genetic sequencing was not as commonplace at the conception of this study so that data is not available.

2.4.2 Future Work

We found good predictive performance using a consistent modeling approach to estimate key pathological parameters. To estimate the generalizability we implemented a repeated five-fold cross validation scheme. However, the final predictions need to be applied

prospectively to samples collected from new patients in order to fully characterize the accuracy. This can be completed alongside additional data collection.

We found that sampling from a fairly uniform distribution of clinical tumor grades produced an imbalanced distribution of sample grades. Namely, we had far fewer high-grade samples than low grade samples. While we were still able to predict the biopsy characteristics, the performance was reduced for these few high-grade samples. Future data collection efforts should focus in collecting samples that balance the data across the range of pathology (e.g. proliferation, cellularity, local grade) observed in the clinical population.

Other worthwhile future work would also be to examine the effect of image acquisition and processing on the correlation with pathology. These results could be used to improve the sequences themselves in order to maximize the predictability or serve to standardize processing techniques for advanced sequences like DCE. Currently, DCE processing requires several user-dependent steps and empirically determined settings.

Chapter 3

Neuroimaging Data Curation

3.1 Introduction

Chapter 2 described how magnetic resonance imaging can predict local glioma pathology using a random forest trained on imaging and tissue data. The central hypothesis of this work is that the pathological parameters estimated by these predictive models can estimate overall survival and be useful in the clinical care of patients. Testing this requires a large patient population with known outcomes and imaging data to use for predictions. This data needs to be curated, processed, and prepared for further analysis in Chapter 4 and Chapter 5. While the basic framework for neuro-image processing such as registration, segmentation, and normalization are well-established, applying this sort of processing at a large scale is non-trivial. The difficulty is further increased when using clinical diagnostic imaging with its wide variety in image acquisition parameters, scanners, and vendors as is the case in this work.

The goal of this chapter is to describe and demonstrate the automated and semi-automated methods for accomplishing the data curation of our large historical dataset. Systematic curation of high-quality data has been identified as an essential and rate-limiting step for development of artificial intelligence [110, 111]. So, this chapter emphasizes the use of a custom data review dashboard to rate data quality and exclude failures. At its conclusion, the availability and quantity of various image types are tabulated for each patient. We also present the results of data review and stratify the processed cases by overall quality and acceptability for further analysis. For the good cases, we quantify

the success in terms of comparisons between image features and “reference” values like tumor or intracranial volume. These results provide the necessary confidence in the quality of curated data which is necessary to believe further survival results in later chapters. Additionally, the failure rates in automated processing are valuable data for future work developing clinical tools out of these methods.

This chapter is composed of three overall steps: First is the collection and classification of brain MRI studies from the hospital picture archiving and communication system (PACS). The key contribution is a custom library of regular expressions acting on study and series descriptions. Next, we present a data processing pipeline to ferry the raw image data files through registration, segmentation, normalization, pathology estimation, and feature extraction. Finally, the quality of the data processing is manually assessed to ensure only high-quality data move on to further analysis. Data quality assurance of all 3,500+ studies is made feasible by a custom data review dashboard.

3.2 Methods: Clinical Image Processing and Curation

3.2.1 Classification of Image Studies and Series

For the patient cohort we needed to identify relevant imaging to use for processing. Ideally, the inter-patient variability between image sequences would be small. For instance, similar TR/TE, resolution, and slice orientation. Since these patients are from a single institution, it is reasonable to expect some consistency. Identifying studies or images using header metadata is a notoriously difficult problem for PACS administrators and there is no standardized way to label study and series description DICOM tags. However, since the historical data set is finite, it is possible to categorize a majority of the images with iteratively crafted regular expressions.

We started with a database of every diagnostic image available for each patient’s preoperative and immediate postoperative image dates. These studies were first filtered using the “Modality” DICOM tag to remove non-MRI studies like chest x-rays and CT scans. The remaining studies were further categorized using regular expressions (applied in R v.3.6.1). The categories and associated case-insensitive regular expressions are listed in Table 3.2. The regular expressions were tested against each series description one at

a time until a match is found. Studies classified as intra-operative, spine, non-brain, or magnetic resonance angiography (MRA) were excluded from further analysis. Any studies not matching any of these regular expressions were excluded too.

Next, the individual image series in each study were classified with regular expressions. The full list is given in Table 3.3. These were also applied in order with the first match taking priority. So, for example a series labeled “Sag Cube FLAIR” matched “CubeFLAIR” and “FLAIR” regular expressions, but it would have been labeled “CubeFLAIR” because that label is tested first. For full processing, each study needed one of each of the following image contrasts: T1-weighted without intravenous contrast, T1-weighted post intravenous contrast, T2-weighted, and FLAIR.

Most studies contain multiple images of the same contrast with different characteristics. This led us to classify image series at a finer level than just T1, T1C, T2, and FLAIR. The description of each image type is listed in Table 3.1. A benefit of the increased number of image categories is reduced redundancy. For example, many cases had both high resolution gradient-echo 3D T1w images with contrast (T13D) and spin-echo post contrast T1w images (AxT1C). It is useful to retain both since both are advantageous in different cases. Other subtypes like CubeFLAIR versus FLAIR mostly just have different resolution or represent images acquired for surgical planning versus diagnostic purposes. (e.g. WandT2 versus AxT2). When multiple comparable images were present for a single study the best one was chosen as follows: CubeFLAIR was used over FLAIR, WandT2 used over AxT2, T13D used over AxT1C for CLARA segmentation and AxT1C used over T13D for pathology estimation. Images that were not processed and reviewed were still identified and converted to NifTI format.

After inspection, this regular expression dictionary captures a majority of the series. Differentiating these sub-types of the four common image sequences (T1, T1C, T2, FLAIR) allows a fine level of analysis. For example, T13D are almost always gradient echo whereas AxT1C are spin echo. Despite both being contrast enhanced T1-weighted images they have different normal tissue contrast and degrees of tumor contrast enhancement. CubeFLAIR and WandT2 tend to be higher resolution than the FLAIR or AxT2 respectively as well. In some cases, there were still multiple image series of the same type are present in a study. In this case, a series of tie breaks were applied in order

series type	description	analyzed
CubeFLAIR	High resolution spin echo T2-FLAIR image usually with near isotropic voxels and used for surgical planning.	YES
T13D	3D gradient echo T1-weighted image after gadolinium contrast injection, usually used for surgical planning.	YES
AxT1C	Axially acquired spin-echo T1-weighted image post contrast. Usually contains thick slices.	YES
T1C	Spin echo T1-weighted post-contrast.	NO
T1	T1-weighted image without gadolinium contrast. Usually acquired by spin-echo with thick slices.	YES
T2star	T2* weighted image.	NO
FLAIR	spin echo T2-weighted fluid attenuated inversion recovery (FLAIR) sequence. Usually 2 mm or greater slice thickness.	YES
WandT2	High resolution T2-weighted image usually used for surgical planning.	YES
AxT2	Axially acquired T2-weighted image usually with thick slices.	YES
OtherT2	Any other T2-weighted image.	NO
DSC	Dynamic susceptibility contrast time series.	NO
DCE	Dynamic contrast enhanced time series.	NO
Trace	Trace image from diffusion tensor imaging.	NO
eADC	Exponential apparent diffusion coefficient from diffusion weighted imaging.	NO
ADC	Apparent diffusion coefficient from diffusion weighted imaging.	NO
AvgDC	Average diffusion coefficient from diffusion tensor imaging.	NO
DWIDTI	Diffusion weighted image series or diffusion tensor image series.	NO
FA	Fractional anisotropy map from diffusion tensor imaging.	NO

TABLE 3.1: Short name and description of each image type identified by the historical data. Image types that were used for survival analysis are indicated by the 'analyzed' column.

study type	regular expression
Non.Brain	(ORB.FACE)
Intraop	(SUITE SURGERY INTRAOP)
fMRI	(FUNCTIONAL FUNCT HAND\$ Speech fMRI)
MRA	(~MRA angio)
ABTI	ABTI
OSF	(OSF OSI Outside Cor)
MRI.Brain	(MR.*BR NEURO CABI HEAD e\+1 WO? CON)
Spine	SPINE CERVICAL
Other.Brain	BRAIN

TABLE 3.2: Regular expressions for classifying MR studies. Spine, intraop, non-brain, and magnetic resonance angiography (MRA) studies were excluded from further analysis. ABTI: advanced brain tumor imaging.

series type	regular expression
nonAnat	(BRAIN SUITE ^DynaSuite FUNCTIONAL FUNCT HAND\$ FINGER Speech Senten epiRT fMRI HANDS NECK LUMBAR Multiplanar ASSET CAL SCOUT SCREEN BURNED LOC CHEST VEINS SELLA REGISTRATION COLOR Anatomy probe MRS ^ (CAT FAS SENT MOTOR) ->)
CubeFLAIR	(Sag(?!.*REFORMAT).*CUBE.*FL SAG.*FL.*CUBE FLAIR ?WAND)
T13D	(3D.*WAND WAND.*3D (?<!R1[-_])SPGR T1 Wand \+C Ax 3D T1 T1 3D 3D AX.*T1 Stealth STEALTH.*T1.*(POST \+C) (?<!RFMT)SAG 3D T1 Sag T1 GRE ^3D (Sag Ax) T1 \+C CUBE T1 ?\+C)
AxT1C	(AX.*T1.*(POST \+C) POST-AXIAL)
T1C	(\+C.*T1 T1.*C POST SPGR)
T1	[^D]T1
T2star	(T2* T2 * T2[(]?STAR)
FLAIR	^(?! (SAG COR).*REFORMAT C).*FLAIR
WandT2	^(?!.*(\+C POST)).*(WAND.*T2 T2.*WAND T2.*STEALTH)
AxT2	^(?! (STAR \+C)).*(AX.*T2 T2.*AX)
OtherT2	(FSE.*T2 T2.*FSE)
DSC	DSC
DCE	DCE
Trace	(T2 D[WT]I).*TRACE
eADC	^Exponen
ADC	(^APPAR LADC)
AvgDC	(Average DC)
DWIDTI	(DWI\$ ^ (Ax)?DTI DIFFUSION Ax DT1)
FA	(FRACT FA ANISO DTLFA)

TABLE 3.3: Regular expressions used to classify MR series by description. The non-anatomical series are not used for further processing.

to attempt to select the best series of each type for further processing. The tie break procedure was:

1. Voxel size: the series with smaller product of in-plane pixel size \times slice thickness is used. This favors approximately isotropic images (e.g. $0.9 \times 0.9 \times 1.2$ mm instead of $0.5 \times 0.5 \times 5.0$ mm)
2. Slice thickness: If voxel size is the same, the series with thinner slices is used.
3. Repeat or retake: if one series contains RPT, REPEAT, or REPL in the description it is assumed to be a retake of the previous acquisition due to poor image quality. The retake is used.
4. Axial acquisition: if one image has “AX” in the series description it is used.

5. Series time: In the case where two images have otherwise identical metadata, choose the series with a later acquisition time. This captures retakes of corrupted images or images with artifacts.
6. Series time: Use the series with the earliest acquisition time. This captures images as originally acquired rather than reformats which are given the same acquisition time.

If none of the tie breaks worked, one series was selected at random. In our experience, these failed tie breaks were inconsequential and did not needlessly exclude data. Finally, for each study a high-resolution image was chosen as a fixed image to which all other images were co-registered. Just like the series tie break, the high-resolution image was selected based on the product of in-plane pixel width and slice thickness. In the case of a tie: preference was given in order:

$$\text{T13D} > \text{CubeFLAIR} > \text{WandT2} > \text{FLAIR} > \text{AxT2}$$

After the best series from each study had been selected, studies from the same day for each patient were also combined. This happened in cases, for example, where a preoperative and fMRI study were both acquired. In these cases, as many images as possible from the best study (having the highest resolution image) were selected and any missing image types were filled from other studies. This ensured the images used were all from the same study when possible but still allowed studies to be combined and enabled further analysis.

3.2.2 Brain Tumor Segmentation Challenge Data

The first part of the data processing and labeling closely followed other neuroimaging conventions: In particular, the MICCAI Brain Tumor Segmentation Challenge (BraTS) [34, 103, 112]. The training data for the challenge consisted of 210 high-grade and 75 low grade glioma cases. The image data had been co-registered, skull stripped, and tumors had been segmented into three classes: 1) peritumoral edema seen as FLAIR hyperintensity, 2) T1 enhancing tumor, and 3) non-enhancing tumor core and tumor necrosis. These are the same labels we used on the historical data since NVIDIA's

CLARA model is trained on the BraTS data. Examples of the segmentations with the three disjoint labels are shown in Figure 3.3 and Figure 3.7.

Since 163 of the highgrade cases also had overall survival data, they served as an independent test of imaging biomarkers from Chapter 4. The BraTS data was processed in parallel with the historical data using the pipeline in Figure 3.1. Since it was already registered and segmented, it entered the pipeline at the *Segment normal tissue* step. The data was also reviewed using the data QA app described in Section 3.2.5.

3.2.3 Neuroimaging Pipeline

The data processing pipeline took advantage of existing software for individual steps like registration and segmentation. These were either stand-alone applications like ANTS [76, 113] and CLARA¹, or as packages available in python or R programming languages. The goal of constructing the data processing pipeline was to adapt the wide variety of clinical data, including postoperative and oblique data sets, to the formats expected by the various software. An overview of the processing workflow is shown in Figure 3.1. The particular steps roughly in order are:

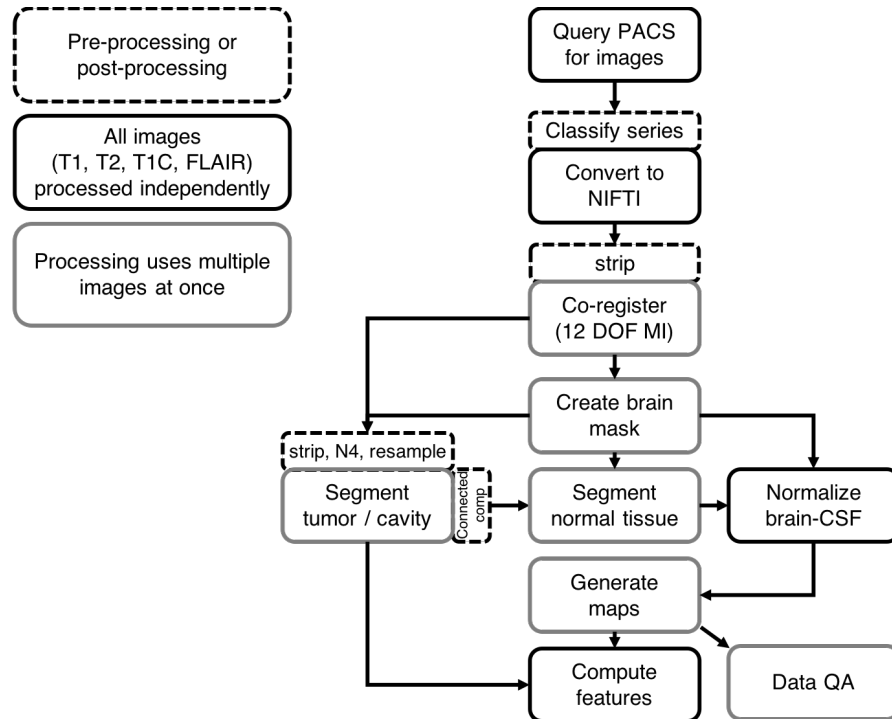


FIGURE 3.1: Image processing pipeline for neuroimaging data.

¹<https://developer.nvidia.com/clara-medical-imaging>

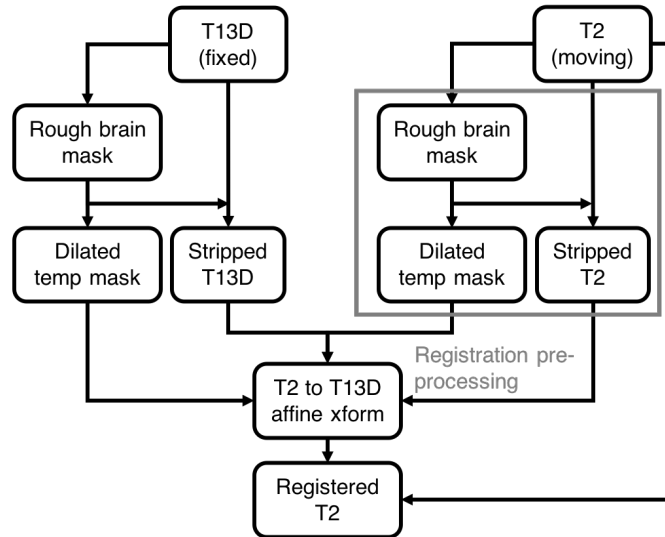


FIGURE 3.2: Registration procedure including bias correction and skull stripping images individually to aid the registration. A dilated mask was used for the cost function as well since it significantly sped up the registration. The other images like T1 and FLAIR were registered in the same way as the T2 is to the T13D (fixed space image).

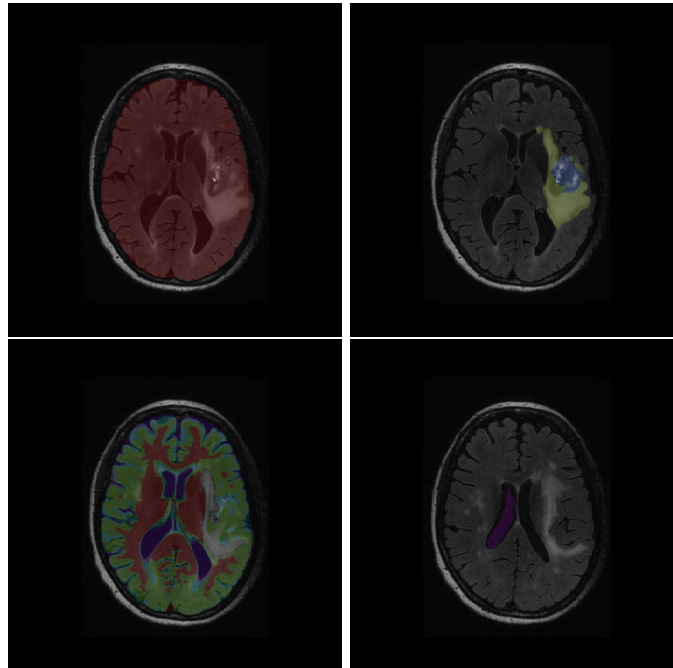


FIGURE 3.3: Sample images for each step of the data processing pipeline. Top left: brain mask, top right: tumor segmentation, bottom left: 4-class tissue segmentation, bottom right: post-processed CSF ROI.

1. Identifying appropriate image series for each patient. First, all imaging data available for each patient's preoperative and postoperative study dates was downloaded from the PACS to a DICOM server. Then, the DICOM header tags for a single image from each series was loaded into an SQL database in order to organize and query the specific series. Individual studies and series (T1, T2, FLAIR, etc) were categorized by a custom library of regular expressions as described in Section 3.2.1. Some further categorization used metadata like magnetic field strength of pixel spacing to identify the best quality imaging.
2. Converting DICOM files to NIFTI format and naming according to the project convention. This also served as an anonymization since patient information in the DICOM metadata was not transcribed in the NIFTI header.
3. Registering images. This was done with a 12 degree-of-freedom affine registration in ANTs. First, images were pre-processed with bias correction over the whole image and individually skull stripped using a brain mask (Example in Figure 3.3). A dilated (1 mm) version of the brain mask was used to mask the cost function and accelerate the registration as well. After registering the skull stripped images, the transform was applied to the non brain-extracted image so that all image voxels were retained. This is described in Figure 3.2. Each image was co-registered to the space of the highest resolution image in the study which was almost always a gradient echo 3D T1 weighted image or a high-resolution FLAIR image.
4. Creating brain masks for each image. For each study, a single brain mask was generated in the space of the highest resolution image using a deformable atlas-based approach [114]. Since multiple image contrasts are available after registration, a more accurate brain mask is achievable than the single-contrast masks generated prior to registration. A sample mask is shown in Figure 3.3. After the brain mask was generated, it was used for all further masking.
5. Segmenting tumor. This step allowed both isolation of normal tissues and feature calculation within the lesion. The best tool for this was a deep neural network, and the current pipeline used NVIDIA's CLARA platform². The particular model used a 3D encoder-decoder architecture that produced the winning entry in the 2018

²Model used: br16 full no automatic mixed precision version 1
https://ngc.nvidia.com/catalog/models/nvidia:med:clara_mri_seg_brain_tumors_br16_full_no_amp

Brain Tumor Segmentation Challenge (BraTS) [115]. The model was pre-trained on the BraTS data and produces segmentations that includes separate regions for enhancing tumor, non-enhancing tumor, and peri-tumoral edema. Sample segmentations are shown in Figure 3.3 and Figure 3.7. Before applying the model, the input images were pre-processed by skull stripping, bias correction using a white matter posterior probability [113, 116], and resampling to 1 mm isotropic resolution. For the T1 post-contrast input, a gradient echo T13D image is used instead of spin echo AxT1C when possible. The resulting masks were resampled back to the original image space via nearest-neighbor interpolation. The predictions from CLARA were post-processed by filtering out all but the largest connected components in the resulting segmentation. For multi-focal tumors, more than one component was retained.

6. Segmenting normal tissue and generating a CSF ROI. In order to normalize using the mode-csf method, each study needed ROIs for the normal brain tissue and cerebrospinal fluid. First, the tumor or resection cavity was subtracted from the whole-brain mask and the resulting non-tumor tissues were clustered using ANTs Atropos [113] into four categories. The T2-bright CSF category was post-processed to define an ROI. Full details are given in Section 3.2.4. Examples of the clustering and CSF ROI are shown in Figure 3.3.
7. Normalize intensities. Anatomic images were linearly scaled such that normal brain and CSF modal intensities had values 0 and 1. Quantitative imaging was already be scaled during DICOM conversion. See Section 3.2.4 for details.
8. Generating maps using the trained random forest model. Each model used T1 pre-contrast, T1 post-contrast, T2-weighted, and FLAIR images. Separate models using AxT1C or T13D image types for a T1 post-contrast image were used depending on availability with the AxT1C model preferred. The specific implementation used R scripts to parse the input files and apply the random forest.
9. Computing measurements of the predicted maps and input images using pyradiomics. First-order statistical measures and shape measurements were extracted. A summary of the exact features used is listed in Section 4.2.2.

Except for the software packages mentioned, the data processing pipeline was implemented in python or R. Processing multiple cases is embarrassingly parallel so we took advantage of a computing cluster. Each compute node had 2 Intel(R) Xeon(R) Gold 6132 CPUs with 14 cores per CPU at 2.60 GHz and 192 GB RAM.

3.2.4 Mode-CSF Normalization

The original method of normalization, called reference tissue normalization, used one set of gray matter, white matter, and cerebrospinal fluid (CSF) ROIs per patient [50, 51, 117]. Each image was linearly scaled such that the darkest tissue ROI had mean intensity 0 and the brightest had intensity 1. For example, of a T2 FLAIR image the intensities were scaled as follows:

$$x \rightarrow \frac{x - \overline{CSF}}{\overline{WM} - \overline{CSF}}$$

Where \overline{CSF} and \overline{WM} are the mean intensities inside the respective ROIs. The benefits of this method are that the manual ROIs won't grossly fail or be misplaced and that the resulting scale is biologically relevant. Most non-contrast-enhanced images generally fall in the intensity range roughly $[0, 2]$ which is pleasant.

However, the reference tissue normalization also has some limitations. First, the manual tissue ROIs are inherently subject to inter-reader variability which adds arbitrariness to the normalization process since there is no definitive way to choose a “best” ROI. Small differences in ROIs perturb the mean intensities inside and this propagates to the resulting normalization. Similarly, there is no automatic way to generate ROIs with similar mean intensities which technically makes a fully-automated pipeline impossible. Lastly, differences in tissue contrast at 1.5 T versus 3.0 T field strengths will clearly affect the normalization. Because of this, it is unclear if this method is sufficient to normalize between studies with different field strengths.

Due to the limitations of the reference tissue approach, we used a modified method, called mode-csf (MCSF), inspired by Autry et al. [63] and the reference tissue method. In this scheme, the modal intensity of the normal brain (whole brain minus tumor) as in [63] and modal intensity of CSF are mapped to 0 and 1 with the order determined by

which is greater. This process is summarized by Figure 3.4. We used CSF since it is the brightest (or darkest) region of MR sequences making it a natural choice to “anchor” the intensity range.

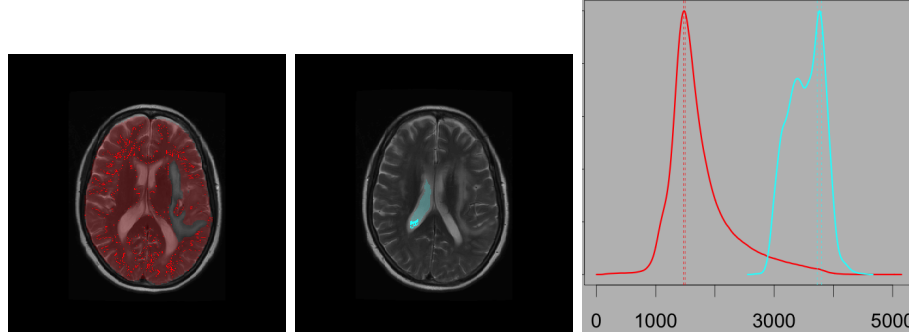


FIGURE 3.4: Example of mode-csf normalization. Left: the regions used for normalization are the normal brain (red) and CSF (cyan). Voxels within 1% of the modal intensity are colored extra bright for illustrative purposes. Right: rescaled density plots of the raw intensity values showing the intensity distributions for the normal brain and CSF. The vertical bars show the range of 1% around the modal intensity corresponding to the bright voxels in the ROIs. The mode-csf normalization maps the intensities such that locations of the peaks are 0 and 1 respectively.

The main advantage is that the mode (compared to the mean) is insensitivity to small errors in masks or ROIs. Specifically, adding or subtracting a relatively small number of voxels does not affect the modal intensity, especially if those voxels have extreme intensity values near the tails of the distribution. In automatic data processing, these small errors are likely and so methods that are insensitive remove those errors as a source of variability.

The specific techniques used in this study for defining whole-brain and CSF ROIs are as follows: A whole-brain mask was generated using a deformable atlas-based approach [114]. Then, any visible tumor was segmented automatically using a pre-trained multi-modality deep learning model from the NVIDIA CLARA platform. The lesion was subtracted from the brain mask resulting in a normal brain mask. More implementation-specific details are give in Section 3.2.3.

Automatically generating a region of pure CSF to normalize with is not a trivial task. The steps used in this study were chosen because they worked empirically and gave values that correlated well with mean CSF intensities from manual ROIs. The steps were:

1. Co-register T1-weighted and T2-weighted images. Using both contrasts helped identify CSF which is bright on T2w images and dark on T1w images.
2. Split the brain into four classes using a Gaussian mixture model. We used ANTs Atropos, initialized by K-means and smoothed by Markov Random Field [113].
3. Isolate the brightest class on the T2w image and erode by 1 mm isotropically. This broke the region into a series of disjoint candidate ROIs (Figure 3.5). It also removed voxels near the boundary of CSF with other tissues that may not be completely inside the CSF and cause partial volume effects.
4. Among the five largest candidate ROIs with more than 500 voxels, choose the one with the unweighted center-of-mass closest to the brain center-of-mass. This almost always selected a ventricular ROI. If none contain > 500 voxels, we used the largest one.

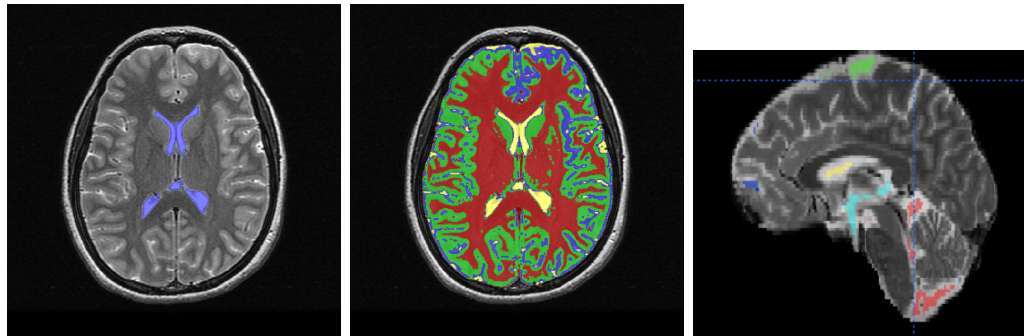


FIGURE 3.5: Example of K-means clustering with Atropos. The T2w image with ventricles highlighted for reference is shown on the left. After clustering (middle), the yellow class contains mostly CSF. After erosion, the CSF class is broken into several candidate ROIs in the various CSF pools (right). The ROI closest to the brain center is selected.

When there are multiple image sequences from the same study, they were mapped to a common space via registration so that the same normal-brain mask and CSF ROI were used for each. A global search over the full range of raw intensity values found the maximum of the density function (i.e. mode). Since the whole brain mask can contain upwards of 10^8 voxels, faster methods based on fast Fourier transforms are preferable to direct computation using Gaussian kernel convolution. The default implementation of `density()` in base R does so.

3.2.5 Data Quality Assurance with R Shiny

A simplified version of the data QA dashboard which can be used for any MR imaging data set is available for download at github.com/EGates1/MRDQED

Large data sets require a way to quickly visualize and check individual data points to catch potential processing failures or misrepresentations. This is especially true with studies like this one where complex images with multiple processing steps are distilled into relatively few image features. Looking at the population is good for catching outliers likely representing processing failures. But, we also needed to avoid the more sinister case where a value is produced that is reasonable with respect to the population range but is actually due to a processing failure. A good example is bright T1 scalp fat being erroneously measured as contrast enhancing tumor.

Loading several 3D images is slow, takes a long time to fully review each slice, and requires stand-alone software to view. So, our solution was a dashboard to quickly and efficiently visualize all the necessary image data for a single case and record if the various images, masks, etc. were sufficient quality to trust the measurements. As a compromise, reviewing just a few perpendicular slices through a segmentation or image was usually sufficient to screen for bad data quality. So, we rendered a handful of key slices as static png image files as part of the data processing pipeline: essentially front-loading the cost to load the data at the expense of the ability to freely scroll through image slices.

Other considerations for a QA tool were usability, especially when other researchers need to contribute to data review, access, how easy it was to launch/interface with the tool, and portability, could it be used with other projects in the future. With these considerations and others in mind we used an R Shiny app. The advantages were:

- Providing visualization and data processing strengths of R and its familiar packages.
- Requiring no more than R and the R shiny package to run.
- Providing an easy framework for reactive programming and user input or interaction.
- Hosting as a web app accessible from anywhere within the institution's firewall by multiple concurrent users.

- Handling page layout and HTML automatically.

The main disadvantage was that the app was solely for identifying bad data and image clean up or editing must be done with outside software. Since it relied on pre-rendered png images, any changes to the underlying data must be re-rendered manually before they would be displayed. Also, the base R shiny package does not come with user authentication built in and must be handled separately if controlled access is desired.

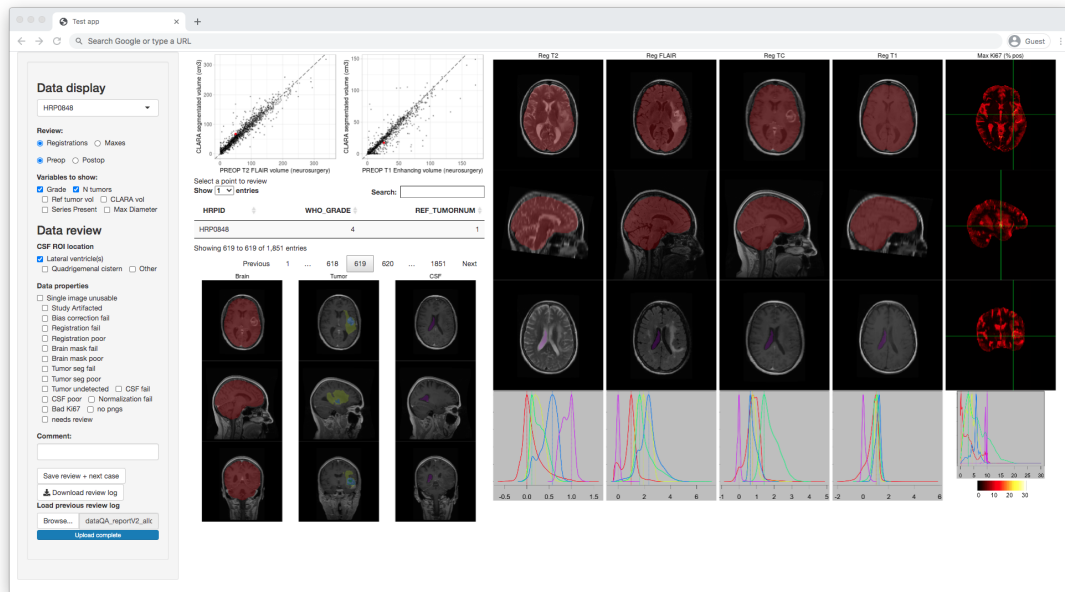


FIGURE 3.6: Shiny app dashboard for data review

3.2.5.1 App Overview

A snapshot of the app is shown in Figure 3.6. First, the left-hand sidebar has two sections: data display and data review. Most of the screen is then taken up by the data display itself. This app was designed to be viewed full-screen on a 1920x1080 resolution 19" monitor. A screen that is much smaller might not look as nice or similar to Figure 3.6. If a png images does not exist, a black rectangle would be shown in its place.

Data display panel The first options in data display control which case is being viewed (HRPID). Each case has a unique identifier of the form HRP####. One way to pick a case to display is with the drop-down box. Clicking a row on the data table to the right also selects that case. The radio buttons for study type change the images displayed between preoperative and postoperative images. Lastly, the check boxes for “variables to

show” determine which columns are displayed in the data table. Information like tumor grade may be useful when reviewing the data and features like maximum diameter can suggest possible processing failures.

Data review panel The next section is titled data review and handles the designation of good and bad data. For each case and study type the app stores the status of the check boxes and comment field. If an image or segmentation has one of these characteristics it can be indicated by checking the corresponding review boxes. In general, each part of the data processing is free from errors, failed (unacceptable), or acceptable with some small errors (poor). The various review boxes are:

1. Study artifacted: All images have large artifacts, are corrupted, or are otherwise not fit for further data processing.
2. Image unusable: One of the images has a large artifact, is corrupted, or is otherwise not fit for further data processing.
3. Bias correction fail: Either there is a strong bias field on one or more images that impacts tumor or CSF segmentation or image features, or the bias correction that was applied corrupted the intensities.
4. Brain mask failure: The brain mask is wrong to such a degree that further processing or features are ruined.
5. Brain mask poor: Brain mask does not perfectly segment intracranial contents but errors are acceptable and not interfere with downstream processing.
6. Tumor segmentation failure: The tumor segmentation misses a large portion of tumor or grossly over-segments normal appearing brain.
7. Tumor segmentation poor: A small amount of normal appearing brain is segmented as tumor or a small amount of obvious lesion is missed, within an acceptable limit.
8. CSF failure: The selected CSF ROI is outside the CSF or captures the wrong modal intensity on one or more images.
9. CSF poor: The CSF ROI is small or closely abutting non-CSF tissues on one or more images.

10. Normalization failure: Even though CSF ROI placement is acceptable, the intensity of the CSF causes normalized image intensities to be outside the expected range.
11. bad Ki67: The maximum intensity Ki67 is outside the visible tumor or the overall values displayed on the map are unreasonable.
12. no pngs: Some pngs are missing unrelated to data errors.
13. needs review: Accuracy is unable to be determined with the images shown. Closer inspection is necessary.

After a case has been reviewed, the “save review + next case” button automatically brings up the next case. After a review session, clicking the “Download review log” button prompts the web browser to save a comma separated value (CSV) file indicating which cases were reviewed, what data qualities were selected, and any comments. Exporting this log is the only way to save the review status and any interruption of the session by navigating away from the app, closing the window, or reconnecting will clear the review history. On subsequent sessions, the CSV review log can be loaded using the file browser right below the download button to import the review data back into the app.

Scatter plots On the top left are two scatter plots comparing the total preoperative segmented tumor volume (shown in three-planes below) and the reference tumor volume. The currently selected case is shown with a red point so it’s location relative to the population can be assessed. Points that are far from the dashed agreement line are suspicious and are good candidates to review first. Hovering your mouse over a point will display the ID of that point below the plots and clicking on a point select that case. Note, these plots are not present in the publicly available version of the dashboard.

Summary data table Below the scatter plots is a table of summary data for each case. The displayed columns can be adjusted in the data display panel and the number of entries shown at a time can be adjusted at the top of the table. The search bar on the top right filters the entries of the table and clicking a row selects that case for review. This makes searching for a particular ID then selecting it easy.

Segmentation views The bottom left of the data display shows the three critical annotations that are generated and used for processing: The brain mask shown as a

single red overlay, tumor segmentation with blue for necrosis and non-enhancing tumor, green for enhancement, and yellow for edema, and CSF ROI in purple. Each one is shown in three orthogonal planes centered on the slices with the largest area. The segmentations and overlaid on the highest resolution image in the study. Most cases will have either a T1 post-contrast image or FLAIR image shown.

Option 1: Corresponding slice views Using the data display radio button, the main display area can be toggled between showing matching slices of all images, which is demonstrated in Figure 3.6, or maximum intensity localizations. The goal of the matches sliced view is to screen for large artifacts and mis-registration. For all images, the same maximal cross-section axial and sagittal whole-brain slices are displayed with brain mask overlaid. Since the slices correspond, salient anatomy (ventricular horns, vascular bifurcations, etc.) should be in the same location across images. If not, mis-registration might be the case. Furthermore, the edge of the brain mask should trace the border of the brain on all images and deviations indicate misalignment. Similarly, the CSF ROI is also displayed in its maximal cross section axial slice for all images. This allows its location in CSF for all images to be verified. Below the maximum views are density plots of the intensity values for each image. The x-axis scale is normalized so that 0 and 1 are the modal brain intensity and modal CSF intensity. Each curve has been scaled to peak height 1 to help with comparison and the colors correspond to the colors shown in the segmentation views (e.g. red for brain, purple for CSF).

Option 2: Maximum intensity (crosshair) views If desired, slice views are replaced by three-plane visualization of the maximum intensity within the segmented tumor region. On each plane, the crosshairs mark the exact location. Maximum image intensity is a potentially useful prognostic feature as well as a location where processing failures are likely. For example, spurious vascular or fat structures segmented as tumor will often have very bright voxels. Note: because the maximum intensity falls in different places on different images the planes shown are not necessarily the same between images nor are they the same as the planes with overlaid segmentations on the segmentation views. The far right columns shows the estimated Ki67 predicted everywhere inside the brain mask and always has the maximum intensity location shown. The color scale is the same as the color bar below the density plot.

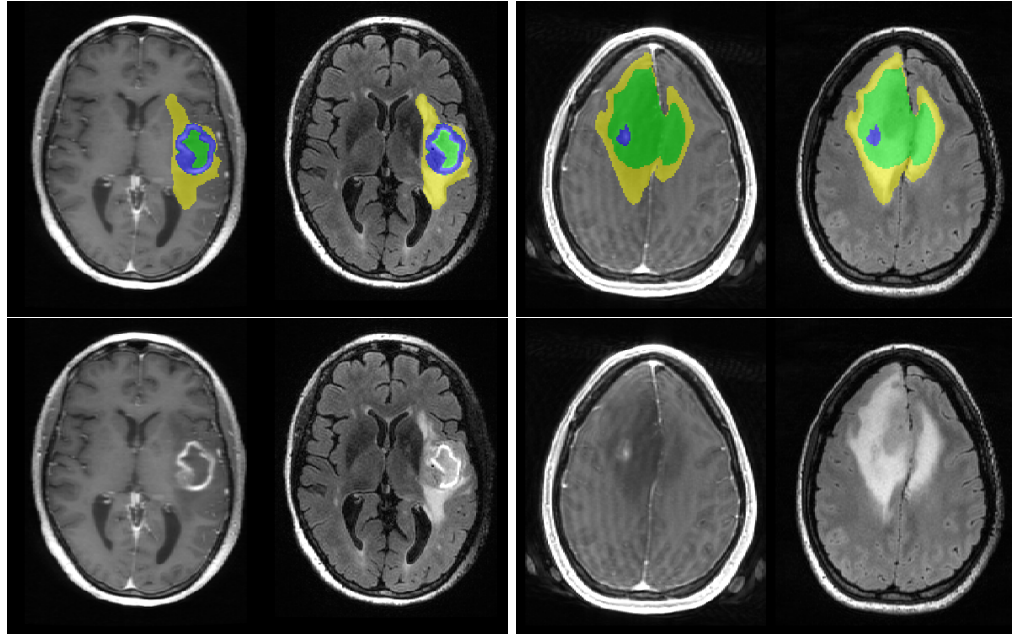


FIGURE 3.7: Sample CLARA segmentation showing the three-class output for two WHO IV glioblastoma patients. T1 post contrast and FLAIR images are shown with and without segmentation for reference. The segmented regions are edema (yellow), non-enhancing tumor core and necrosis (green), and enhancing tumor (blue). Clinically, the right-hand case has enhancement pattern “both.”

3.2.5.2 Data Review Procedure

All cases were reviewed in the shiny app sequentially following these instructions:

1. In a web browser connect to the app.
2. To use a review log started from an existing session, upload it with the “Load previous review log” file browser.
3. Select a case to review with the drop down HRPID menu or by clicking a row of the summary data table. Alternatively, hover over a point on the scatter plots to reveal its HRPID, search that HRPID in the data table search bar, then select the row with that ID.
4. Check the data: do the images in the maximum views look reasonable? Do the brain, tumor, and CSF segmentations look ok? If any are unacceptable check the corresponding boxes under data review. Inspect the values in the plots and table. Do the volumes agree with reference values and does the maximum diameter seem reasonable? If not segmentation errors may be present. Lastly, add any helpful comments about data failures or this specific case in the Comment field.

5. When finished with that case, click the “save review + next case” button. This will bring up the next consecutive HRPID with available data. Alternatively, select a case manually with the table or scatter plots again.
6. Save a review log at the end of the review session by clicking the “Download review log” button.

3.3 Results: Processed Data Summary

3.3.1 Image Series Classification

A full description of the clinical data sources is given in Section 4.2.1. In summary, from 1935 patient records that had either preoperative or postoperative imaging available to download, a majority of the study descriptions were able to be classified by the regular expressions. The exact numbers are tabulated in Table 3.4. A vast majority were in the “MRI Brain” category and only 6% were in the excluded categories: intraoperative, spine, non-brain, or MRA. There were a fairly large number of fMRI studies which usually contained T1w and FLAIR images among others. The functional data was not used for this project.

	Number of studies
ABTI	17
fMRI	234
Intraop	202
MR SPECTROSCOPY	2
MRA	6
MRI Brain	3412
Non-Brain	7
OSF	102
Other Brain	16
Spine	9

TABLE 3.4: Total number of studies of each type categorized by the regular expression library for all patients. OSF: outside facility

A large majority (81.4%) of preoperative studies had a 3D T1w post-contrast image to use as a high resolution. Only 4.7% of studies had to use a low-resolution Axial T2w image as a fixed image as well. For postoperative data, under half of studies had high-resolution postoperative data (39%). The frequency of each image type being used as the fixed image is listed in Table 3.6.

	POSTOP	PREOP
ADC	235	311
AvgDC	810	862
AxT1C	1633	1291
AxT2	1795	1464
CubeFLAIR	41	360
DCE	0	99
DSC	1	94
DWIDTI	4411	3941
eADC	121	43
FA	817	1105
FLAIR	1943	2213
not-Classified	1301	3202
OtherT2	20	206
T1	1997	1946
T13D	716	2501
T1C	2894	2547
T2star	1365	1160
Trace	535	944
WandT2	12	1132

TABLE 3.5: Total number of series of each type categorized by the regular expression library for all patients with any brain MR images.

	POSTOP	PREOP
T13D	669	1116
CubeFLAIR	38	220
WandT2	1	7
FLAIR	372	86
AxT2	648	64

TABLE 3.6: Fixed image type for each study. T13D, CubeFLAIR, and WandT2 are usually near 1 mm isotropic resolution or better.

A total of 1370 adult patients with WHO grade II - IV glioma had a complete preoperative imaging study for analysis. Several had more than the four required series. Among the 1370 cases, 1071 had both gradient echo T1 post-contrast (T13D) and spin-echo T1 post-contrast (AxT1C) images. 123 had only spin-echo contrast enhanced images, and 299 had only gradient-echo. Of all the series identified in the raw data, not all were used for further analysis. The number of each series that were processed are listed in Table 3.7. Note: 1370 cases had preoperative imaging available and 1728 had postoperative imaging available so the entries in Table 3.7 are out of a possible 1728 or 1370.

	AxT1C	AxT2	CubeFLAIR	FLAIR	T1	T13D	T2star	WandT2
POSTOP	1533	1741	39	1746	1749	681	1337	9
PREOP	1234	1265	325	1563	1566	1487	1144	1050

TABLE 3.7: Number of images of each type included in the curated data set. Only cases with a full preoperative or postoperative study were included.

3.3.2 Image Data Processing

On a compute node, registration and brain masking for one case required about one hour. Tumor segmentation inference took only about 30 seconds per case, and the overall processing time was about 2 hours per study at most. The actual computation time varied greatly depending on the image matrix size. The time required for feature extraction after image processing also varied. A summary of the various run times is listed in Table 3.8. Feature extraction varies considerably in computational time as well over all image and mask combinations which varies considerably depending on image size and ranges up to about 20 minutes.

Processing step	Approximate processing time
DICOM to NifTI conversion ⁺	1 minute
pre-processing and registration	62 minutes
brain masking and skull strip	16 minutes
CLARA (inference, pre + post processing)*	5 minutes
Atropos tissue clustering + post processing	3 minutes
Image normalization	5 minutes
Pathology model inference	2 minutes
Pyradiomics feature extraction	17 minutes
png render for data QA	18 minutes
Total:	129 minutes

TABLE 3.8: Data processing times for a representative image study (fixed image size 512x512x124). Times are for single-node processing, note that many nodes can run simultaneously. ⁺Dicom conversion is run on a separate workstation not on a computing cluster. *CLARA inference runs all inference sequentially inside a docker which creates a bottleneck in the processing.

After an initial round of data review, bias correction was noted to have distorted lesion contrast and harmed subsequent segmentations in 172 studies. These cases were re-processed with bias correction disabled. Cases with data processing failures due to failed registrations, failed brain masks, or failed bias correction were also re-processed from raw image files. After re-processing, those cases were reevaluated using the QA app.

The procedure generating cerebrospinal fluid ROIs for normalization performed well. Generally, the ROI was placed in one of three locations: lateral ventricles, quadrigeminal cistern superior to the cerebellum, or in the sulci on the brain periphery. Examples of these locations are shown in Figure 3.8. About 84% of preoperative studies and 78% of postoperative studies had ROIs including the lateral ventricles as desired. When this was not the case, it was often due to mass effect compressing the lateral ventricles and making a clean ROI infeasible. The ROI inclusion rates for the various locations is given in Table 3.9.

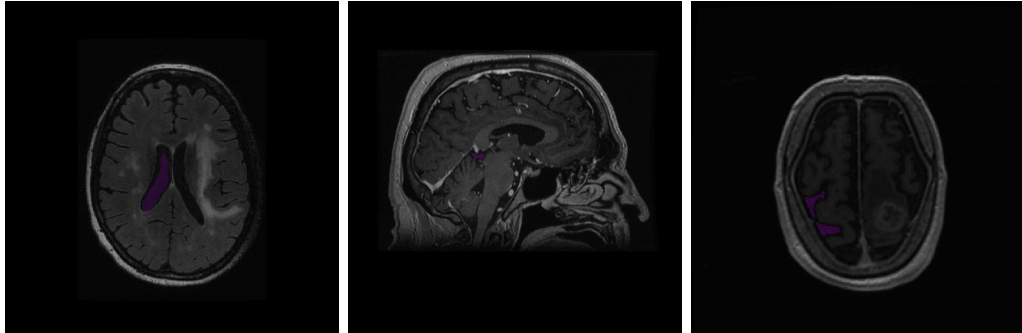


FIGURE 3.8: The three most common locations for CSF ROI were the lateral ventricles (left), quadrigeminal cistern (middle), or superior to the brain (right). All give reasonable CSF intensity statistics.

	studytype	lateral ventricles	quadrigeminal cistern	other	n
1	POSTOP	0.78	0.18	0.17	1721
2	PREOP	0.84	0.18	0.06	1481

TABLE 3.9: Proportion of cases where the selected CSF ROI included the given location. Multiple locations are possible so the totals exceed 100%

3.3.3 Data Quality Assurance

Initial review of the preoperative imaging data (1717) cases took approximately 50 hours of combined reviewing time. This includes reviewing via the web app interface, manually reviewing full 3D images when necessary, and investigating unusual failure modes. An experienced neuroimaging researcher (author EG, 4 years of experience) could review about 70 - 100 cases per hour on average. Qualitatively, the data review procedure had a shallow learning curve associated with it as the reviewer became more familiar with the appearance of the various processing failures. Overall. About 64% of the preop data and 51% of the postop data was “ok” after initial review. The failure rate was greater in

the postoperative data (28%) than in the preoperative data (14%). The exact numbers are given in table 3.10.

		exclude	preop data acceptable	ok	NA
postop data	exclude	74	41	66	19
	acceptable	82	60	128	22
	ok	198	163	459	68
	NA	94	79	164	0

TABLE 3.10: Result of data QA for all 1717 cases. Review is NA if a preop or postop study was not available for that patient.

When counting failure modes, only the most upstream failure was counted. For example, if tumor segmentation failed due to a corrupted image, then only the “bad image” would be counted since the tumor segmentation would likely have succeeded had the images been good quality. Counting all downstream failures would inflate the failure rate for later pieces of the pipeline. However, if two failure modes were judged to have occurred independently in the same case, such as a corrupted image and a failed registration on another image, then both were recorded.

The main unacceptable failure mode was tumor segmentation failure and a majority of those were due to small, undetectable lesions less than 10 cc in total volume. The other source of failures occurred in about equal proportion. Exact proportions are listed in Table 3.11. When corrupted or images or large artifacts were noted, those series were replaced with a similar series if possible. Seventeen image series were replaced in this way. This means the cases excluded for corrupted images or artifacts had no suitable replacement series.

	POSTOP	PREOP
Corrupted image	0.0032	0.0038
Artifact	0.0086	0.0059
Bias correction failure	0.0081	0.0011
Registration failure	0.0324	0.0346
Brain mask failure	0.0411	0.0162
Tumor segmentation failure	0.1583	0.0502
Tumor undetected	0.0238	0.0216
CSF ROI failure	0.0400	0.0130
Normalization failure	0.0249	0.0151
Unreasonable estimated pathology	0.0016	0.0011

TABLE 3.11: Proportion of each failure mode in the data processing pipeline. Steps are listed in the order they occur.

3.3.4 BraTS Data Processing and Review

In the historical data, we differentiated between gradient echo and spin echo post-contrast images and trained separate predictive models for each. The BraTS data was reportedly generated using gradient echo pre-surgical scans [34] so we applied gradient echo models to all cases. However, it was noted in our data review that 18 cases appeared to use spin echo T1 post-contrast images.

Overall the BraTS data was good quality with 219 of 285 cases passing QA, see Table 3.12. However, 43 cases had image quality that was deemed unacceptable. These were almost exclusively due to poor single image quality or improper brain-CSF normalization. Exact numbers are in Table 3.13. While normalization failure is a weakness of our own methods. The base image quality is an issue with the challenge data itself. By far, the most common image quality issue was excessively short field or cropped fields of view. This was observed in 14 separate image studies. In extreme cases like the ones in Figure 3.9 the ground truth tumor segmentation goes beyond the cropped image.

data quality	number of cases
exclude	43
acceptable	23
ok	219

TABLE 3.12: Result of data QA for all 285 BraTS cases.

	proportion of cases
Corrupted image	0.0561
Artifact	0.0000
CSF ROI failure	0.0175
Normalization failure	0.0877

TABLE 3.13: Proportion of each failure mode in the data processing pipeline for the BraTS data. No other failure modes were observed.

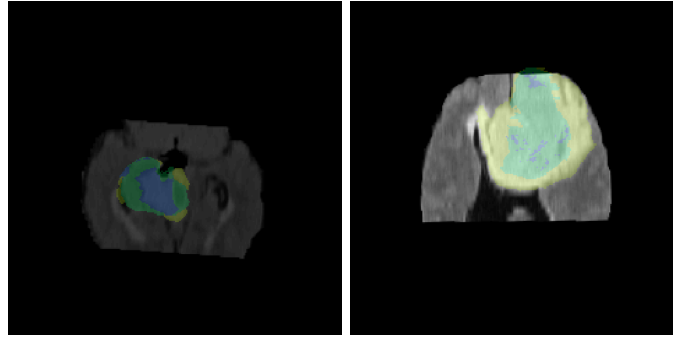


FIGURE 3.9: Examples of FLAIR images and ground truth tumor segmentations included in the 2018 Brain Tumor Segmentation Challenge. Left: Brats_2013_0_1. Right: Brats18_2013_6_1. In both cases, the image field of view is so short that the segmentation is partially outside the brain volume. Both of these were caught by data review.

3.3.5 Comparison with Reference Values: Intracranial and Tumor Volume

The first step in data analysis is a visualization of raw data. Namely, we generated several quantities like ROI volumes that we expect to lie within normal ranges or match independently calculated values. After successful data processing, we expect to reproduce known values summarizing the characteristics of neuroimaging data. One such property is total intracranial volume (TIV), which is the volume of the brain mask. The distribution here should be consistent with values from literature for a similar population. The other value is the patient-specific preoperative tumor volume which has been previously measured by an independent research study [14, 15].

3.3.5.1 Intracranial Volume

First, each patient has a brain mask associated with each of the preoperative and post-operative studies. Using the volume of these masks, we compared these values to the expected population distribution for total intracranial volume (TIV). Note - TIV is the appropriate comparison since the masks used for image processing include cerebrospinal fluid and not just brain tissue. Literature values for TIV are fairly consistent between studies. Lüders 2002 reported a breakdown of brain size (including CSF) by gender in Table 1 of [1]. It listed $1510 \pm 400 \text{ cm}^3$ for men ($n=50$) and $1320 \pm 100 \text{ cm}^3$ for women ($n=50$). All patients were neurologically normal. The standard deviation for males seemed quite high, though. Recalculating from the constituent components gave a more reasonable standard deviation of 155 cm^3 . Table 2 in Jenkins et al. 2000 [118] estimated

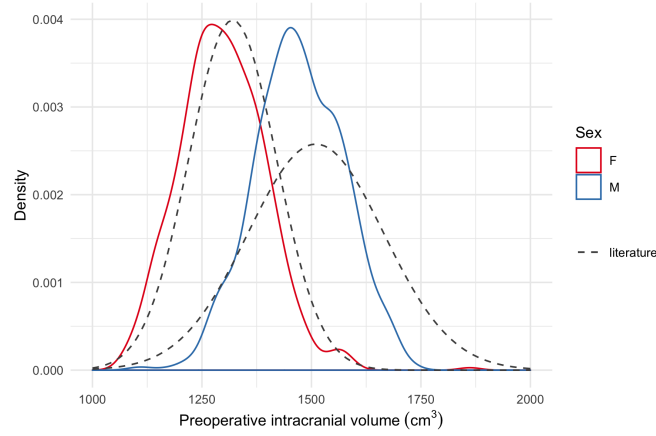


FIGURE 3.10: Distributions of preoperative intracranial volume for historical cases. The dashed lines are normal densities from [1], reproduced in Table 3.14.

TIV in 52 control patients to be $1512.5 \pm 128.2 \text{ cm}^3$ for males ($n=24$) and 1316.8 ± 97.6 for females ($n=28$). The volumes for Alzheimers patients was nearly identical. Lastly, Blatter et al listed TIV for 192 subjects and stated “A significant difference in the TICV was observed between male ($1558 \text{ cm}^3 \pm 97 \text{ cm}^3$) and female ($1352 \text{ cm}^3 \pm 115 \text{ cm}^3$) ... subjects” [119]. This reference also split the TIV by decade of life. 105 subjects were female and 89 were male. A summary of these references is given in Table 3.14.

Reference	subjects	Male TIV in cm^3 (n)	Female TIV cm^3 (n)
Lüders 2002 [1]	young, healthy, volunteer students	$1510 \pm 400^\dagger$ (50)	1320 ± 100 (50)
Jenkins 2000 [118]	52 healthy controls age 52 ± 12	1512.5 ± 128.2 (24)	1316.8 ± 97.6 (28)
Blatter 1995 [119]	health subjects age 16-65	1558 ± 97 (89)	1352 ± 115 (105)
This work (2021)	1181 untreated adult glioma patients	1473 ± 98 (702)	1298 ± 95 (479)

TABLE 3.14: Total intracranial volume (TIV) measured using 3D imaging for several different populations. † This value was reported in Table 1 of [1] but it is inconsistent with the reported mean and range of values. The true value is likely around 155 based on the other values in that table.

Using the data from Lüdders 2002 reproduced in Table 3.14 our data matches fairly well, see Figure 3.10. Excessive smoothing might have biased the measurements to be slightly small compared to population data. For our data we measured TIV to be 1298 ± 95 for females ($n = 479$) and 1473 ± 98 for males ($n = 702$). These numbers only include brain masks of acceptable data quality.

3.3.5.2 Tumor Volume

Preoperative and postoperative tumor volumes for each patient were measured by neurosurgery researchers for prior studies [14, 15]: specifically, T1 contrast enhancing volume, T2 FLAIR lesion volume, and T1 hypointense volume. Enhancing volume measurements also include necrotic tissue within the enhancing volume when present. These reference measurements provide excellent benchmark values for automatically segmented tumor volumes.

To make a proper comparison to the reference values, we combined the appropriate labels from the CLARA three-class segmentation (enhancing tumor, non-enhancing and necrosis, and edema, Figure 3.7). The specific combination depends on the enhancement status of the tumor. For purely non-enhancing and enhancing tumors, the reference T1 enhancing volume was compared to the non-enhancing plus enhancing regions. This accounts for necrosis present in the reference T1 enhancing measurements. For enhancement status “both”, the non-enhancing tumor core is not included in the reference enhancement measurement due to a large amount of non-enhancing tumor in addition to enhancing tissue. An example of this pattern is shown in the right-hand side of Figure 3.7. For this case, only the enhancing label is counted as enhancing volume. Reference T2 FLAIR volume was equivalent to the sum of the three sub-regions from CLARA for all cases. The part of the tumor measured for the reference T1 hypointensity is not a direct combination of the labels present in the CLARA segmentation which means a direct comparison was impossible. Summing the non-enhancing and enhancing sub-regions provided the closest agreement but appears to consistently underestimate the T1 hypo intense volume, Figure 3.11. This is likely because some of the T1 hypointensity is sufficiently FLAIR hyperintense that it is labeled as edema instead. The relations between the CLARA subregions and reference measurements are given explicitly in Table 3.15.

Enhancement Status	T1 enhancing volume	T2 FLAIR volume	T1 hypointensity
Enhancing	neh + enhanc	neh + enhanc + edema	neh + enhanc
Non-enhancing	neh + enhanc	neh + enhanc + edema	neh + enhanc
Both	enhanc	neh + enhanc + edema	neh + enhanc

TABLE 3.15: How CLARA segmented volumes are added to compare to reference T1 enhancing and T2 FLAIR volumes. Enhanc is enhancing tumor (blue in Figure 3.7), neh is nonenhancing tumor core and necrosis (green in Figure 3.7).

Preoperative volumes agreed quite well with reference measurements. The enhancing volume measurements mostly agreed well but had more variability that warrants further investigation. Comparisons of these values are in Figure 3.11. Preoperative values agrees quite well for both measurements. However, there are still several cases where CLARA identified some enhancing volume where reference measurements measure no enhancing volume. This is usually due to small enhancing vessels mistaken for enhancing tumor. We could have corrected this by excluding enhancement preoperatively for clinically non-enhancing tumors. But for 75% of cases, it would have adjusted the total volume by less than 0.2 cm^3 so we elected not to.

Postoperative measurements of total tumor volume showed systematic over-estimation, although the overall correlation was still strong at $R = 0.83$. Enhancing volume and T1 hypointensity are not well estimated by automatic measurements. Several cases had zero reference volume but some small non-zero amount segmented, likely due to treatment effect. The agreement of postoperative enhancing tumor volumes could be improved if the patient's clinical designation of gross-total or subtotal resection were used to exclude the enhancement. But, this potentially biases the results so a correction was not made.

Excluding the bad data had a positive effect on the correlation with reference values. Figure 3.12 shows the agreement between segmented tumor volume and reference tumor volume for the three data quality levels. As the data quality increases, the correlation also increases and the points generally fall closer to the agreement line. The root-mean-square-error also decreases with better data quality, Table 3.16. There were a few exceptions where there is disagreement between the volumes but the data is otherwise ok. There are cases that are within the limits of human judgement, such as where to draw a boundary on faint FLAIR hyper-intensity. In some cases, there was considerable volume of both non-enhancing tumor and necrosis. Since the reference values only count necrosis, this inflated the enhancing volume and is not actually an error. It is a consequence of using three classes in CLARA not four.

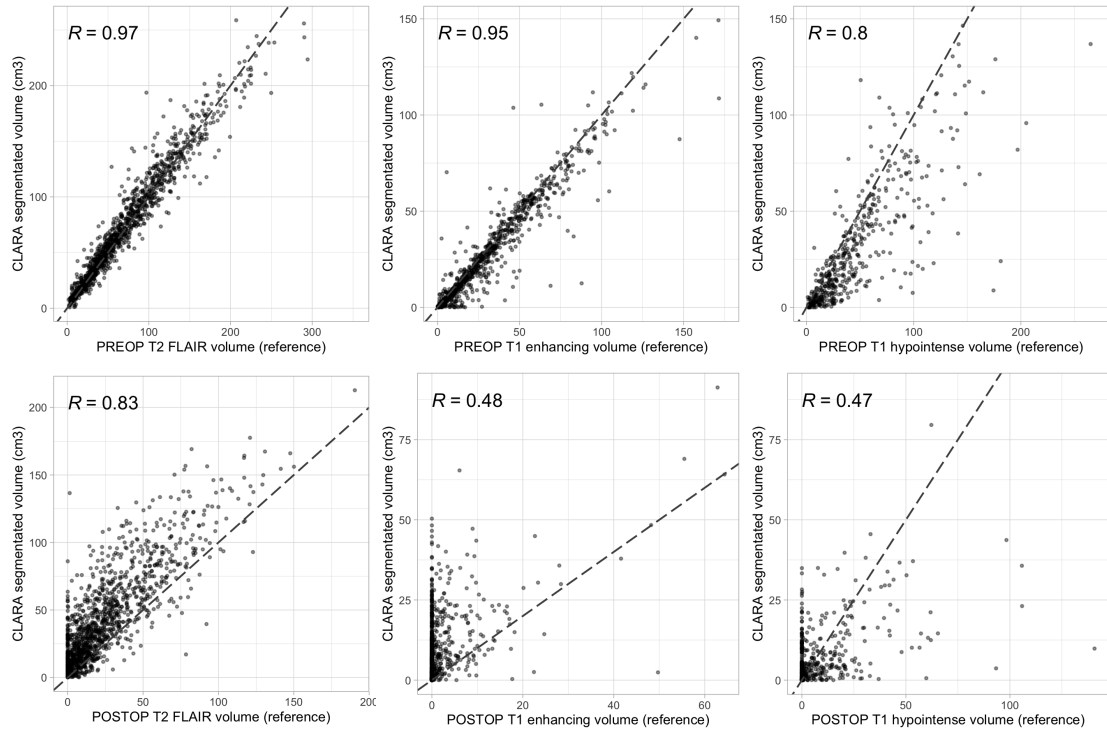


FIGURE 3.11: Comparison of tumor volumes segmented by CLARA and volumes measured previously by neurosurgery. T2 FLAIR volume included the entire FLAIR lesion or total visible tumor volume. Enhancing volume includes both enhancing tissue and tumor necrosis. T1 hypointense volume includes tumor core outside of enhancing volume. The dashed lines indicate agreement.

	data quality	T2-FLAIR RMSE	T1-EV RMSE
1	exclude	24.22	19.36
2	acceptable	20.46	9.35
3	ok	12.22	9.31

TABLE 3.16: Root mean square error in cm^3 for preoperative segmented tumor volume (T2-FLAIR and T1 enhancing) compared to reference tumor volume. Separate values are given for each level of data quality.

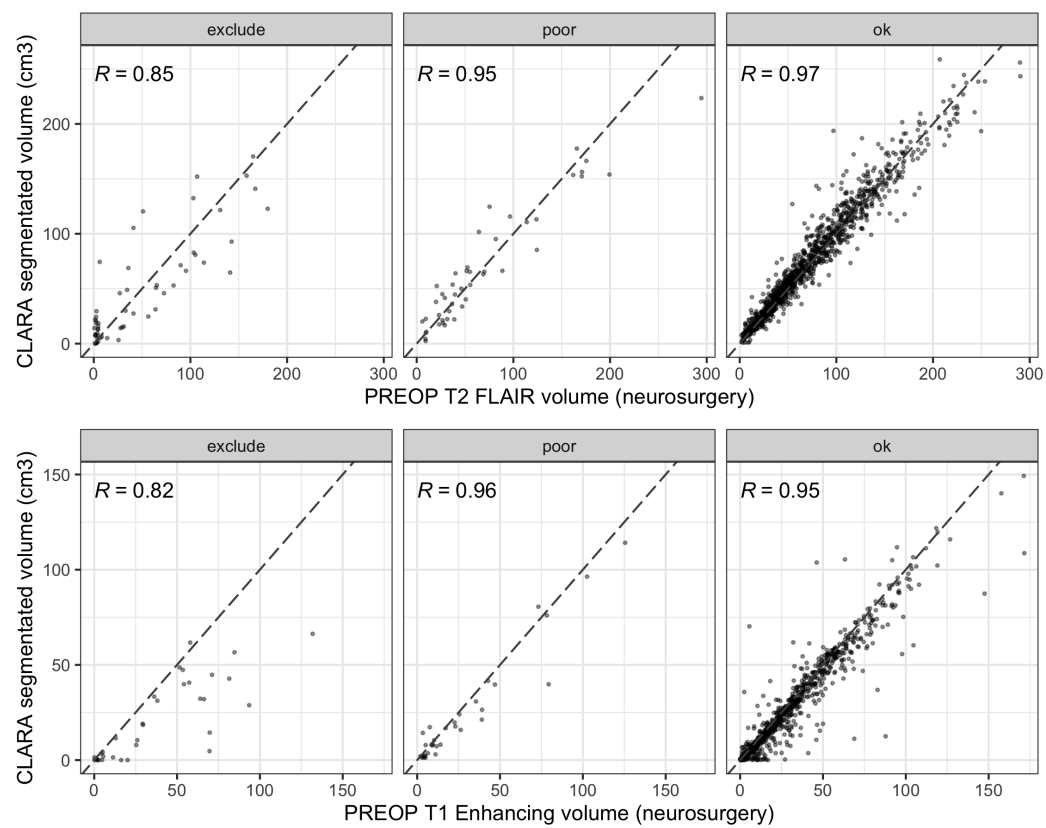


FIGURE 3.12: Preoperative T2 (top) and enhancing tumor volumes (bottom) for historical cases compared to reference values. Each plot shows values for a different level of segmentation quality and the dashed line indicates agreement. “Poor” data is acceptable with minor errors.

3.4 Discussion

Using a combination of automated image processing and manual data review the primary results achieved in this chapter were:

1. Classification and organization of over 3,500 MR imaging studies containing over 46,000 image series from a single institution using a custom library of regular expressions applied to study and series descriptions.
2. Effective automated processing of image studies including registration, segmentation, and normalization which produced acceptable results for 86% of preoperative cases and 72% of postoperative cases.
3. Verification of data quality enabled by a custom R shiny web app.

Identifying the four primary image contrasts (T1w, T1w with contrast, T2w, and FLAIR) needed for image analysis from an arbitrary study is a challenging task due to lack of a standardized series naming scheme between vendors or institutions. Rule-based solutions using DICOM metadata or artificial intelligence have been proposed to tackle this [120]. The goal in this work was to classify a finite total number of series from a single institution. This allowed a simpler approach based on regular expressions for the study and series descriptions alone. Because the regular expression libraries in Tables 3.2 and 3.3 were built in conjunction with the finite number of series descriptions contained in the retrospective data, they may not be flexible enough for prospective classification of new series.

The curation of this large-scale historical image database with annotations not only enables the survival analysis discussed in later chapters, but also serves as a resource for further investigation. The sheer volume necessitated use of a computing cluster in this work. However, processing a single study, as would be the case in clinical practice, can be performed in just over a few hours. The most time-consuming step by far is skull stripping during registration pre-processing since each of four images must have the brain outlined individually. We chose to use an atlas based multi-contrast skull strip algorithm [114]. While the results were fairly reliable, this algorithm performs 4 - 9 fully deformable image registrations per image stripped as well as post-processing with patch matching. This means the total time could be reduced considerably by substituting a

faster skull stripping algorithm [121]. However, finding one that works effectively on many image contrasts (T1w, T2w, and FLAIR) and on postoperative data is difficult.

Preoperative total intracranial volume (TIV) served as a surrogate for the accuracy of the skull-stripping algorithm. While the individual measurements came from annotations labeled as “brain mask” the TIV is a more appropriate measure since it includes the CSF which is part of the mask. On postoperative data, the mask also included any resection cavity which was not part of the brain but was part of the intracranial contents. Although, the postoperative data is not included in Figure 3.10 or Table 3.14. Unlike tumor volumes, we do not have patient specific reference measurements for TIV so we are instead limited to comparing the overall distribution to population values [1, 118, 119]. We found an underestimation of TIV compared to the reference distributions. This might be explained by masks with small amounts of inferior cerebellum excluded were still designated as acceptable quality and hence included in the final data.

We compared segmented tumor volume with a reference tumor volume as an approximate measure of segmentation quality. Overall, Figure 3.11 showed good agreement with reference values with a strong correlation of $R = 0.97$ for total tumor volume. Some deviation from human-generated reference values is expected even for a perfect algorithm given that tumor segmentation is inexact. In a small study where ten radiation oncologists were asked to contour four glioblastoma gross tumor volumes, the standard deviations in the T2-FLAIR volumes ranged from 22.6 cm³ to 62.6 cm³ [122]. This is comparable or even larger than the RMSE of 12.22 cm³ to 20.46 cm³ for the acceptable segmentations from CLARA. In other words, the distance from true values (RMSE) is comparable to the variability in the truth itself.

While the preoperative tumor volumes agreed quite well, Figure 3.11 illustrated a systematic overestimation of postoperative FLAIR hyperintense volume relative to the reference measurements. The regression coefficient for postoperative whole tumor volume versus reference was 1.089. These data suggest that the residual disease volume is overestimated by about 9%. This phenomenon is well established in neurosurgical literature and is generally attributed to edema or ischemia [123–125]. Although, these studies demonstrate significant differences between early and late postoperative T2w and FLAIR hyperintensity as groups, there is little data on the expected patient-level effect. A study in non-enhancing gliomas by Belhawi et al. found a 20% overestimation

on early vs late postoperative FLAIR volume [126] using linear regression. This is larger than our observed 9% but may be explained by the difference in tumor grades present. Intraoperative MRI or ultra-early (immediate after wound closure) is less susceptible to these treatment effects [124]. But, the retrospective analysis in this work limits us to the available early preoperative imaging. One solution to mitigate the overestimation of residual tumor is with diffusion-weighted imaging [125]. A large number of our cases do have postoperative diffusion weighted imaging that could be incorporated to better estimate residual disease.

Comparison with reference tumor volume gives an incomplete assessment of the overall segmentation performance. While agreement in general suggest a good segmentation, it is not entirely sensitive or specific for segmentation quality since it does not capture the spatial distribution of the segmentation. To illustrate: we observed some counter-examples. Figure 3.12 shows some cases where failed (exclude) cases had good agreement with reference volume. This is due to a segmentation with the correct approximate size placed in the wrong location in the brain. We also saw cases with a much larger or smaller total tumor volume than the reference but was acceptable. These were due to faint T2-FLAIR hyperintensity which creates a large uncertainty in tumor extent. We expect human readers would have similarly discrepant. Lastly, segmenting the enhancing or necrotic compartment of a tumor is somewhat arbitrary as well. Many of the cases with discrepant enhancing volumes in Figures 3.11 and 3.12 have overall total volume agreement even though the individual labels, i.e. enhancement versus edema, might disagree.

Ideally, we would have compared the automatic segmentations to a complete manual segmentation on the historical data set using metrics like Dice similarity or Hausdorff distance. Unfortunately, such human segmentations were not available. CLARA was trained on 285 cases from the brain tumor segmentation challenge [34]. On the validation set it achieved average Dice scores of 0.851 for tumor core, 0.773 for enhancing tumor, and 0.903 for whole tumor³. Since the historical data was pre-processed to match the Brain Tumor Segmentation Challenge data, it would be reasonable to assume the Dice scores with human-generated ground truth would be similar to those values.

³ngc.nvidia.com/catalog/models/nvidia:med:clara_mri_seg_brain_tumors_br16_full_no_amp

The previous counter examples where good tumor volume agreement was achieved with bad segmentations and vice versa illustrated the need for careful data review. For complex image processing basic data quality, annotation accuracy, intensity statistics, and other properties all need to be assessed for accuracy. Almost every step in Figure 3.1 introduces a failure mode with some frequency. To check all aspects at once we used a custom data review dashboard and found that the QA procedure, while laborious, conferred great confidence in the data quality moving forward. Using the dashboard, we even found instances of bad data quality in the Brain Tumor Segmentation Challenge data. In fact, we marked about 6% of the 285 cases as having image quality bad enough to exclude from further analysis. It is unclear how extreme cases like the ones shown Figure 3.9 have impacted the training and evaluation of the best segmentation algorithms. Indeed, these cases were used to train the CLARA segmentation model we applied. In short, this simply underscores the need to carefully review data, even data coming from popular and well-documented sources.

There are existing tools assess raw image quality [127, 128]. however good image quality alone does not guarantee the success of downstream processes. Many of these other processes have some ways to identify uncertainty like test-time dropout for segmentation [129] but ultimately these still require either human intervention for the final decision. Lastly, when exploring a new data set it is very difficult to know a-priori which failure modes will be encountered and which metrics can accurately identify them. This makes automated data review a “chicken and the egg” problem where such methods cannot be assessed without manually reviewing the data first to get ground truth. For work such as this one where the goal is to use the curated data for further research, discovering the best metrics for QA retrospectively provides little benefit for the project at hand. Although, doing so would be a rich area of future work.

3.5 Summary of Curation Effort

Step	Curation achieved	Resources	# Patients	# Studies	# Series	Computer time (core hrs)	Labor (FTE)	Data size
Collect clinical data from neuro-surgery	quantified clinical outcomes	Clinical research team	2,584	5,168		-	years ³	-
Download DICOM image data from PACS	images accessible for processing	DICOM database	1,935	4,326	64,346	Weeks	2 weeks	12 TB DICOM files
Classify studies, remove non-MR non-brain	data temporally organized	desktop & custom regex library	1904	3,728	46,329	0.5	1 week	-
Classify series, remove unused	Raw files ready	desktop & custom regex library	1,899	3,667	27,662	0.5	3 weeks	-
Process best series for each case image data	full results and feature data available	computing cluster	1,851	3,591	23,888	167,000	30 weeks ²	181 GB Raw NIfTIs 3.9 TB after processing
Review image data, exclude unusable	validated data quality	MRDQED app ¹	1,851	1,807 ok, 673 acceptable, 741 bad	23,888	27,000	8 weeks	-
TOTAL	publishable results		1,851	3,591	23,888	194,000	12 months	16 TB

TABLE 3.17: Summary of curation effort. Three steps that take an appreciable amount of time: A) Downloading raw DICOM data from PACS. This takes took a few weeks since we throttle and only transferred during nights and weekends. B) Processing the image data from raw DICOMs to a final data matrix can be done in about 24 hours on a moderate sized computing cluster. C) Data review takes about 80 human hours to review all studies. This includes breaks and distractions. Additional time to re-processed and review bad data included. ¹[R shiny app](#). ²Includes to develop and test processing pipeline. ³Not included in total labor amount.

Chapter 4

Preoperative Imaging-Pathologic Estimators of Survival

4.1 Introduction

Prognosis for glioma patients varies greatly depending on the overall grade of the tumor, clinical factors, and treatment. Clinically, patients that live longer also tend to be younger and have high mental performance status [14, 15], these two factors are recognized as independently important and are commonly controlled for as covariates in survival models. Mental status is measured by the Karnofsky performance status scale which scores a patient’s ability to independently function in daily life from 0 to 100 in increments of 10. A score of 80 or above means they are still able to carry out normal activities without special assistance [2, 130]. The list of possible scores with description is reproduced from Oken et al. [2] in Table 4.1.

The single most powerful survival factor currently known is the tumor grade described by the World Health Organization (WHO) scale [20]. The WHO scale grades tumors from I to IV with a higher grade being more malignant and carrying worse prognosis. In this work we consider adult patients with tumors of WHO grade II or higher (grade I tumors are usually found in pediatric populations). Historically, the WHO grade has been determined based on tissue histology, using such pathological features as the presence of mitoses, microvascular proliferation, or necrosis in order to characterize the tumor. Some imaging features like contrast enhancement are known to be associated with higher

Description	Karnofsky Status
Normal, no complaints	100
Able to carry on normal activities. Minor signs or symptoms of disease	90
Normal activity with effort	80
Care for self. Unable to carry on normal activity or to do active work	70
Requires occasional assistance, but able to care for most of his needs	60
Requires considerable assistance and frequent medical care	50
Disabled. Requires special care and assistance	40
Severely disabled. Hospitalization indicated though death non-imminent	30
Very sick. Hospitalization necessary. Active supportive treatment necessary	20
Moribund	10
Dead	0

TABLE 4.1: The Karnofsky performance scale, reproduced from: *Oken, M.M., Creech, R.H., Tormey, D.C., Horton, J., Davis, T.E., McFadden, E.T., Carbone, P.P.: Toxicity And Response Criteria Of The Eastern Cooperative Oncology Group. Am J Clin Oncol 5:649-655, 1982. [2]*

WHO grade [33]. As of 2016, the discovery of important mutations or molecular markers like IDH1/2 mutation [86] have shifted the WHO grading scale from histologically based to histologic and molecular. In fact, certain histologic subtypes can now be completely defined based on key genetic markers as is the case with oligodendroglioma with chromosomal 1p/19q co-deletion [14, 15].

While the prognostic difference between glioma grades is large, the grading system still depends on having tissue specimens available. Obtaining tissue data is difficult, expensive, and incurs some risk like all surgery. Furthermore, tissue samples are often obtained concurrently with tumor debulking so the tumor grade is obtained after treatment decisions (e.g. to operate) have already been made. In contrast with tissue data, MRI is relatively cheap, safe, and easy to obtain. This is the primary motivation behind the predictive modeling and imaging-pathology correlations developed in Chapter 2 and has also spurred the field of “radiomics” searching for strong image-based biomarkers to inform prognosis and treatment [35]. For gliomas, radiomic analysis is confounded by well-established intra-tumoral heterogeneity [9]. In short, feature measurements over whole tumor volumes or even large radiographically distinct sub-regions can fail to account for differences in underlying histology. This leads to “averaging out” possible

survival signal. Thus, there is a need to quantify the true biological heterogeneity in order to discover meaningful biomarkers.

4.1.1 Summary of Analysis

In this chapter, we examine the full histologic heterogeneity using the predictive models developed in Chapter 2 and applied to curated data from Chapter 3. We address the following primary hypothesis: *Image features provide additional prognostic information for overall survival relative to age, KPS, and tumor grade.* If the hazard ratio in multivariate proportional hazard models for an image feature is significantly different from 1 then the imaging provides additional prognostic information not captured in grade or clinical measures. If the primary hypothesis is not supported, a secondary hypothesis is: *The prognostic information contained in image features (imaging phenotype) is non-inferior to the information provided by histologic tumor grade when controlled for performance status and age.*

The results in this chapter are presented roughly in this order: First, we correlated the baseline clinical characteristics like age, mental status, clinical histologic grade, and tumor volume with survival and reproduced the known relationships established in the literature. For tumor volume, we analyzed the automatically segmented tumor volumes which were used for feature extraction.

Next, we applied survival modeling techniques to raw image feature data. The primary results are hazard ratios from Cox proportional hazards models for overall survival in all patients. Inclusion of age, mental status (KPS), WHO grade and image features as inputs demonstrate the independence of prognostic features. The prognostic value of each feature is computed individually and the best feature derived from each image type (T1, T1C, T2, FLAIR) is presented as a candidate biomarker. We repeated this analysis to include features derived from estimated pathology maps (Ki67, CD, ERG, local grade). Just like the raw image features, we present mainly the best performing feature from each map type and list additional results in the appendix. For completeness, we performed repeat analysis on the subset of patients with each WHO grade (II, III, IV). However, for clarity we focus our results on the analysis of the combined cohort consisting of multiple WHO grades. Subset analysis is discussed when it provides noteworthy results beyond the data presented in the combined analysis.

In addition to the prognostic value of imaging, we are particularly interested if the prognostic value of predicted biological heterogeneity is better than features from raw images in terms of survival stratification. In general, these predictive models synthesize the existing imaging and encode learned correlations with pathology. So, it is reasonable to expect they will produce strong prognostic biomarkers. However, it is also possible that the reduction from a four-dimensional space (T1, T1C, T2, FLAIR) to a one-dimensional space (e.g. Ki67) will lose some important information. There is also still some error associated with the model predictions and this noise may interfere with prognostic stratification. Regardless of the differences between raw image features and predicted pathology features, using the nonlinear random forest to transform the image data adds significant complexity. The cost of this increased complexity should be weighed against potential improvement in risk stratification.

4.1.2 Image Feature Complexity

The field of radiomics is concerned with finding imaging measurements that correlate with a particular diagnosis or outcome [131]. In general, measurements are function of three parameters: the image being measured, a region being measured over, and the feature being computed. For example, the mean T2w image intensity over a visible lesion. Mean intensity is a simple and interpretable feature. But, sometimes information may be contained in so-called higher order features that measure subtle characteristics of the image gray levels. For instance, the "short run low gray level emphasis" measures how prominent short dark structures are in a region. While these complex features may have the ability to identify subtle differences between tumor types, the trade-off is that they are much more difficult to interpret and tend to be highly sensitive to image acquisition parameters and preprocessing [36, 132, 133].

The complexity of a measurement also depends on the underlying image. Raw diagnostic images are generally the simplest. But, it is possible to also measure features over derived or functional images like diffusion anisotropy or pharmacokinetic parameter maps like K^{trans} from DCE. Non-linear combinations of several images, such as the estimated pathology maps in Chapter 2, are yet even more complex. As complexity increases, features lose interpretability. Although, regression to a biologically relevant variable

hedges this complexity by adding interpretability. The ultimate goal is to discover biomarkers that meet several criteria. Namely we want features that are:

- Reproducible: describing some real quantity that can be accurately described and reproduced.
- Interpretable: able to be logically defined and visibly connected to some real quantity describing the disease. Hand-crafted features used in this work carry at least some interpretability by the nature of being human designed. This is in contrast to deep learning features which are generated entirely based on data.
- Actionable: the value of prognostic stratification is that it can be used to guide individual patient treatment. As an example, genetic biomarkers may suggest the use of certain chemotherapeutic agents. For imaging, we prefer biomarkers that can identify effective targets for spatially focused therapies like surgery and radiation.
- Testable: the value of quantitative biomarkers should be measurable and consistent on an individual patient basis so that they can be evaluated prospectively in clinical trials.

4.2 Methods

4.2.1 Clinical Data Sources and Description

Patient data for the retrospective historical cohort was collected under a HIPAA-compliant and institutional review board approved retrospective protocol at MD Anderson Cancer Center (PA12-0753 Chart/Imaging Review of the Patterns of growth in Gliomas, PI Schellingerhout). The Interdisciplinary Brain and Spine Center database was queried for patient records between June 1, 1993 and May 31, 2018. The search was limited to patients who were ultimately diagnosed with glioma and underwent first surgical resection at MD Anderson. The returned clinical data for each patient included several types of information.

- Basic information: medical record number, surgery date, date of birth, and sex.

- Outcome information: follow-up time, vital status at follow-up, mental status, and post-operative complications if applicable.
- Diagnosis and treatment information: histologic diagnosis, WHO grade, preoperative and postoperative imaging dates, surgery date, type of surgery, previous surgery if applicable.
- Reference tumor measurements: enhancement pattern, T2-FLAIR volume, T1 contrast-enhancing volume, T1 hypointense volume, and extent of resection based on T1 enhancing or T2-FLAIR volume. Additionally, tumor laterality and eloquent involvement.

From this information, overall survival was calculated using the surgery date, latest follow-up time, and vital status. The patient's age at time of surgery was calculated using date of birth and surgery date. Unfortunately, information on other treatments like radiation or chemotherapy before or after surgery was not available. However, treatment courses for glioma patients has been mostly standardized since the introduction of the Stupp protocol [4] meaning most patients receive the same chemoradiation protocols. Using single-institution data also reduced variability in therapeutic choice which minimizes confounding.

4.2.1.1 IDH Mutation Data

To obtain genetic information like IDH1 mutations status, we joined the clinical data with molecular testing results collected through a separate MD Anderson protocol (Proactive program, PI: Dr. John De Groot). This data was originally gathered using a combination of molecular diagnostics lab sequencing results and manual data abstraction of pathology and clinical documentation. IDH mutant status was confirmed by either codon R132H immunohistochemical staining or IDH1/2 sequencing results. Wild-type status was confirmed by sequencing which also checks for non-canonical IDH1 mutation and IDH2 mutations. Four WHO grade IV patients had negative R132H IHC staining and were assumed to be IDH wild-type since only about 10% of gliomas have non-canonical IDH1 mutations or IDH2 mutations [86]. In total, IDH mutation status was present for about 25% of the 2,588 patient records queried.

Several studies have confirmed that IDH1 mutation rates in glioma vary according to grade [134–136] and age [19, 137, 138]. In the subset of data with known IDH status we observed similar rates of mutation by subgroup. The proportions are listed in Table 4.2 and agree well with published values. In a meta-analysis by Sun et al. of 937 low-grade glioma patients, 744 (79.4 %) had IDH1 mutations [134]. This falls between the two rates we observed for young and old patients. Studies by Hartmann et al. found that 6.2% (15/237) of GBM patients ≤ 60 years old were IDH1 mutant compared to just 1% (2/237) of GBM patients over 60 years old. We found a much higher mutation rate among young GBM patients but had a lower age threshold of 55 years. There is also evidence that IDH1 mutation rates varies with histologic subtype [135, 136] but this would cause too fine of a distinction to be meaningful given that the differences between subtypes are smaller than with age and Grade.

Since IDH mutation status is was available for only a fraction of the historical cases, it posed a substantial missing data problem for survival analysis. One solution would be to randomly imputed the IDH1 mutation status using the observed mutation rates among cases with the same WHO grade and age category (above or below 55 years old). Imputing molecular data in this way has been performed in similar studies [139]. But, we deemed the proportion of missing data is too large to meaningfully impute. Instead, we chose to exclude IDH mutation status from the overall analysis and instead report repeated analysis on the subset with known IDH status as supplementary results in appendix Section A.5. This decision was also supported by evidence that MR imaging is strongly predictive of IDH mutation status [109]. Therefore, image information should also be able to capture the relevant prognostic information contained in IDH mutation status. For future work, a model like the one presented in [109] could be used to appropriately impute the missing IDH mutation status.

WHO Grade	Age group	Observed mutation rate	N
II	<55	0.878	311
II	>55	0.737	46
III	<55	0.806	405
III	>55	0.300	84
IV	<55	0.195	545
IV	>55	0.038	923

TABLE 4.2: IDH mutation rates stratified by patient age (above or below 55 years) and WHO grade. The number of cases per group is given as N. IDH mutation status was confirmed by IDH1 R132H immunohistochemical staining or sequencing.

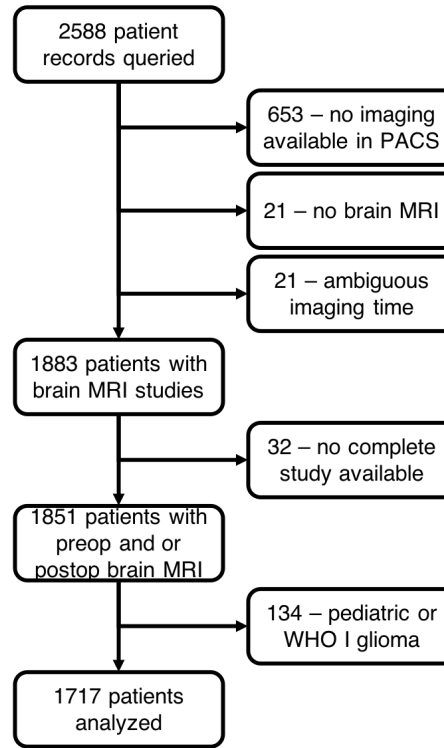


FIGURE 4.1: Flowchart for patient selection in the historical data. Ambiguous imaging time means the imaging and surgery were on the same day. A complete study includes at least one of each: T1-weighted pre-contrast, T1-weighted post-contrast, T2 weighted, and FLAIR images.

4.2.1.2 Imaging Data

MR Imaging data was downloaded from the hospital picture archiving and communication system (PACS) using each patient's preoperative and postoperative imaging dates. For many older studies, imaging was unavailable since it was not digitally included in the patient's medical record. Additional information on image data curation is in Section 3.2. The final cohort of patients for historical analysis met the following criteria:

1. Adult, age 18 years or older
2. Histologic diagnosis of WHO II, III, or IV glioma
3. At least one full imaging study (preoperative or postoperative) with T1, T2, FLAIR, and contrast-enhanced images.

4.2.2 Feature Extraction and Mathematical Description of Image Features

We use Pyradiomics (v3.0) to extract image features [36]. The features are described in the [Pyradiomics documentation](#). To contextualize the features, it is easiest to consider each image I with K voxels as a function that maps each voxel to a real value. $I : \{1, 2, \dots, K\} \rightarrow \mathbb{R}^K$ We can define a region to measure features over as a subset of the image's domain $L \subseteq \{1, 2, \dots, K\}$ Let X be the set of image values for the N_p voxels in region L . $X_L = \{I(i) : i \in L\}, |X| = N_p$. We use the notation $X(i)$ to denote the individual values in the set X . All features we consider in this work other than 3D diameter are a functions of the set of voxel intensities X , which means they are functions that map the I and L to real values. Shape based features like volume do not use intensity information and instead just use the region indices:

$$\text{Voxel Volume: } V_L = \sum_{k=1}^{N_p} V_k \quad (4.1)$$

$$\text{Max 3D Diameter: } D_{\max} = \max(\|\vec{v}_i - \vec{v}_j\|_2) \quad (4.2)$$

Where V_L is the voxel volume of region (label) L and V_k is the volume of each voxel and is a constant for each image. The vector \vec{v}_k is the vector in 3D space at the location of voxel i in L . The 3D diameter is the maximum Euclidean distance between pairs of points. Practically, this is computed using vertices on the convex hull [35]. The first-order intensity-based features use the image values as computed by pyradiomics:

$$\text{Mean: } \bar{X}_L = \frac{1}{N_p} \sum_{i=1}^{N_p} X(i) \quad (4.3)$$

$$\text{Maximum: } \max_L(X) = \max(\{X(i), \dots, X(N_p)\}) \quad (4.4)$$

$$\text{Minimum: } \min_L(X) = \min(\{X(i), \dots, X(N_p)\}) \quad (4.5)$$

$$\text{Percentile: } P_k(X) = \min\{c : |\{X(i) : X(i) \leq c\}|/N_p \leq k/100\} \quad (4.6)$$

$$\text{Median: } \text{med}_L(X) = P_{50}(X) \quad (4.7)$$

$$\text{Total Sum: } \Sigma X_L = V_k \sum_{i=1}^{N_p} X(i) = \bar{X} N_p V_k \quad (4.8)$$

$$\text{Total Energy: } V_k \Sigma X_L^2 = V_k \sum_{i=1}^{N_p} X(i)^2 \quad (4.9)$$

$$\text{Root Mean Square: } \text{RMS}(X) = \sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} X(i)^2} \quad (4.10)$$

$$\text{Intensity at Volume: } I_{V'} = P_q(X) : q = 100 \left(1 - \frac{V' \text{cm}^3}{N_p V_k}\right) \quad (4.11)$$

V_k is the volume of each voxel and is a constant for each image and $P_k(X)$ is the k -th percentile of the values of X . The total sum feature was added manually to pyradiomics. Intensity volume histogram features were also manually computed. The intensity at volume features are quantile measurements that use volume instead of fraction of total data. For example, the intensity at Volume $V' = 5 \text{ cm}^3$ for a region with total volume $|L| = N_p V_k = 250 \text{ cm}^3$ is equivalent to the $1 - 5/250 = 0.98$ or 98th percentile. Note, this is technically a different definition than the one given by the [IBSI](#) but difference is negligible. An illustration is shown in Figure 4.2. The purpose of these quantile measurements is analogous to Windsorizing [140] and is to provide robust alternatives to maximum and minimum features which rely on the (often unstable) value of a single voxel. For simplicity, we refer to the IntensityAtVolume1000 feature as the “1 cc intensity” and so on for other volumes.

Recall, each image measurement is specified by three attributes: An image being measured, a region over which the measurement is being made, and the feature being computed. For example, Mean_T2_wholetumor is the mean T2-weighted intensity over the whole visible tumor. We considered 22 intensity features for each of the 7 different

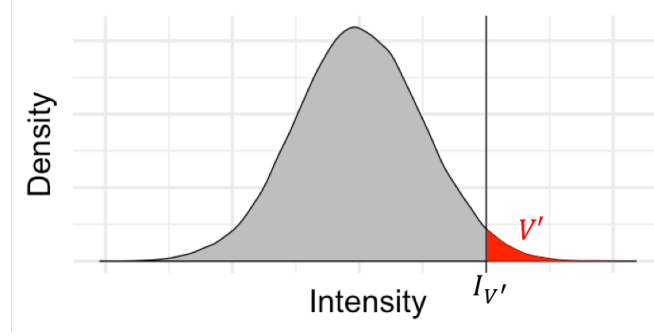


FIGURE 4.2: Illustration of Intensity At Volume features. On the density plot of voxel values the total area of the density plot is equal to the region volume being measured over. The Intensity At Volume feature for a volume V' is the intensity such that the area in the upper tail (red) has area corresponding to the volume V' .

images and 5 different regions, plus two shape features over each of the 5 regions as well as the voxel volume and 3D diameter of the predicted lower-grade and higher-grade tumor regions. This gives a total of 784 different features measurements. The possible combinations are summarized in Table 4.3.

Image	Region	Feature
T2	Whole tumor	Maximum
T1	Enhancing volume	Median
T1C	Non-enhancing tumor	Mean
FLAIR	edema	Minimum
Ki67	whole brain without CSF	Total Sum
CD	lower grade	Root Mean Square
ERG	higher grade	Total Energy
		IntensityAtVolume (0.01,0.1,1,10 cc)
		Percentile (10, 90-99)
		Voxel Volume
		Max 3D Diameter

TABLE 4.3: Images, regions (Figure 3.7), and features used to compute image measurements. Shape features Voxel Volume and Max 3D Diameter only depend on the region being measured over not the underlying image.

The three tumor region measurements mentioned in Table 4.3 correspond to the CLARA segmented regions. Namely, the non-enhancing tumor, enhancing tumor, and edema correspond to the green, blue, and yellow labels in Figure 3.7. The whole tumor region is the union of all three labels. The other three regions are the normal, lower-, and higher-grade regions from the predicted grade maps. For these, we computed the volumes and diameters only. In total, we computed 784 features per patient.

4.2.3 Survival Analysis

We employed standard survival analysis methods including the Kaplan-Meier method and Cox proportional Hazard modeling [141]. These methods are well established in the literature including previous analysis of extend the value of extent resection for glioma patients [14, 15].

4.2.3.1 Concordance Index

We calculated the concordance index or C-index between each feature and overall survival¹. The C-index captures association between the feature and outcome in the presence of right censoring by estimating the probability that a larger feature value corresponds to longer survival. To calculate the C-index, consider all possible pairs of subjects. Each pair is either concordant, discordant, or tied depending on the values of the feature x and outcome y . Pairs where the survival times cannot be unambiguously ranked (e.g. two censored times or tied times) are excluded as incomparable.

	$y_i < y_j$	$y_i = y_j$	$y_i > y_j$
$x_i < x_j$	concordant	Tied Y	discordant
$x_i = x_j$	Tied X	Tied XY	Tied X
$x_i > x_j$	discordant	Tied Y	concordant

TABLE 4.4: Definition of concordant and discordant pairs in terms of feature values x and survival times y .

Using the number of pairs in each category, the concordance index C is estimated as

$$C = P(x_i > x_j | y_i > y_j) = \frac{\text{Concordant} + \text{Tied X}/2}{\text{Concordant} + \text{Discordant} + \text{Tied X}}$$

4.2.3.2 Proportional Hazard Model

In addition to computing the C index we use proportional hazards modeling to parametrically examine the relationship between time to an event and an exposure variable [142]. It uses the hazard function $h(t)$ which models the instantaneous failure rate at a time t for a population. One definition is $h(t) = -S'(t)/S(t)$ where $S(t)$ is the survival function giving the fraction of the population alive at time t . The proportional hazard model

¹See cran.r-project.org/web/packages/survival/vignettes/concordance.pdf

assumes that the hazard function for a member of the population with covariate vector \mathbf{Z} is of the form

$$h(t) = h_0(t) \exp(\beta^T \mathbf{Z})$$

Where $h_0(t)$ is the baseline hazard and β is a vector of parameters. The hazard ratio $h(t)/h_0(t)$ represents the increased chance of death relative to the baseline case $\mathbf{Z} = \mathbf{0}$. A benefit of the proportional hazards model is that by fitting using partial likelihood the baseline hazard does not need to be known or estimated. For details, see appendix Section A.6.1.

4.2.3.3 Cross-Validated Binary Thresholds

The proportional hazards model assumes a continuous relationship between feature value and hazard. However, this may not be the case across the entire range of a given feature's values. Instead, there may be a threshold where patients above and below have very different risk characteristics. For example, a volume of contrast enhancement equal to 0 (i.e. non-enhancing) versus greater than zero (i.e. enhancing). Performing a binary categorization using a threshold also makes the features clinically useful since they allow simple categorization of patients into high- and low-risk.

To look for such thresholds, we searched for an optimal stratifying threshold for each feature. For a specific feature X , the value for a particular patient j is X_j and we find the optimal stratification $HR_{max}(X)$ among the training set.

$$HR_{max}(X) = \operatorname{argmax}_i \max \{ \exp(\beta), \exp(-\beta) \} : \sum_j \delta(X_j \leq i) > n_{\min} \ \& \ \sum_j \delta(X_j > i) > n_{\min} \quad (4.12)$$

Where β is the coefficient from a Cox proportional hazard model fit using the two groups of patients with $X_j \leq i$ and $X_j > i$. In other words, β maximizes the log partial likelihood in Equation A.3. Using the maximum of $\exp(\beta)$ and $\exp(-\beta)$ allows a larger feature value to be associated with either increased or decreased survival. The

parameter n_{\min} is a minimum group size which we set as 20% of the number of training cases unless otherwise specified and δ is an indicator function which counts the number of cases above and below the cutoff i . $HR_{\max}(X)$ can be found with a line search over the unique values of X . The search is restricted to just the values of X that provide a significant (log-rank) difference in survival between the two groups when possible. When no threshold provides a significant survival difference, the largest non-significant hazard ratio is reported. Since we test many features, we correct the univariate significance level to account for the number of features tested per base image type using the Benjamini-Hochberg method which controls the false discovery rate [143]. We report the corrected p -values for univariate tests. When those features and their high risk/low risk predictions are used in multivariate analysis against clinical factors, the p -values are not corrected since there is no longer a multiple comparisons problem.

Since optimizing survival differences carries risk of over-fitting, we wrapped this analysis in a 10-fold cross validation [144]. The data was split into 10 folds with roughly equal distributions of survival times and censoring. Then, the cases in each fold were assigned to the high-risk or low-risk group using a threshold i calibrated on the remaining 9 folds. We pooled the test set predictions so that every case had one prediction associated with it. We then measured the hazard ratio between the predicted high risk and low risk groups. Unless explicitly stated, all survival results reported are based on cross-validated survival estimates. Statistical differences in between groups is assessed using a log-rank test. Technically, the fold structure introduces a small amount of dependency in the predictions which violates an assumption of the log-rank test. An ideal solution would be bootstrap estimation of the distribution of the log-rank test statistic to establish a significance level [144]. But, this incurs significant computational cost and is the subject of future work.

We also repeated this procedure using multivariate Cox models that also include known prognostic clinical factors. First, we included known important clinical factors: low Karnofsky performance status ($KPS < 80$) [130], and age > 55 . This allows the prognostic capability of each feature to be contextualized. The importance of these other clinical factors have been previously established [14, 15, 86]. Next, we added tumor grade (low grade WHO II or high grade III and IV) to age and KPS as third covariate in the Cox model. This allowed us to evaluate the independent prognostic power of the image features relative to the WHO grade of the tumor. For the subset analysis of

patients with known IDH mutation status, IDH mutation status (mutant or wild-type) is included in the multivariate analysis as well.

Finally, we attempted to validate the best prognostic image features on an independent data set from the 2018 Brain Tumor Segmentation Challenge (BraTS). For each of the cases with known survival and acceptable image quality we applied the optimal hazard ratio from the WHO grade IV historical data to the BraTS cases and calculated the hazard ratio between groups above and below the cutoff. This external data set for validation provided a strong challenge to the prognostic power of each feature. The results of this experiment are listed in the appendix Section A.5.1.

4.3 Results

4.3.1 Key Results

- Advanced age, low KPS, and high WHO grade are all strongly associated with worse prognosis. These clinical factors are well established by previous neurosurgical literature [14, 15].
- Larger contrast enhancing volume and total visible tumor volume are associated with poor survival. This result holds for both human drawn (known from literature) and automatically segmented (this work) tumor volumes. As expected, larger tumor volumes confer a worse prognosis.
- Maximum contrast enhancing brightness was the single best imaging predictive feature, with a brighter enhancement associated with worse survival. Other features that correlated strongly with the volume of enhancing tumor were highly prognostic as well.
- When maximum contrast enhancement was added to known clinical prognostic factors (age, KPS, and grade) we found additional predictive information to still be significant and additive.
- Features from synthetic pathology maps did not perform as well as maximum contrast enhancement although heightened cell density, proliferation, and vascularity were still independently associated with worse prognosis.

4.3.2 Patient Cohorts and Clinical Data Summary

Of the 6,563 total cases from 5,319 unique patients, there were 2,588 patients whose first resection was at MD Anderson Cancer Center. Among those, 1851 had imaging suitable for further analysis (Figure 4.1). Among those, 449 patients had previous surgery listed including “Biopsy-burr hole” (n=2), “Biopsy-stereotactic” (n=160), “Biopsy-open” (n=57), “Biopsy-NOS” (n=210), “shunt” (n=4), “Resection” (n=3), “Other” (n=13). Since a majority of patients with previous surgery had just biopsy, we elected to include these patients in the analysis. Any cases with substantial postoperative changes to the brain from prior surgery were excluded during the data review procedure. For other exclusions: 4 cases were listed as having no imaging and 18 listed as CT imaging only. No additional filtering is applied based on histology to exclude diagnoses like pilocytic astrocytoma, gliosarcoma, etc. although these histologies were rare. The actual histologic diagnoses of the patients are tallied in Table 4.6.

Among the remaining candidate cases, patients were included or excluded based on the availability of MR imaging. Additional cases with pediatric patients, WHO grade 0 (i.e. no grade), or WHO grade I tumors were included in image processing but excluded from this survival analysis. A summary of exclusions is given in Figure 4.1. Clinical characteristics of the resulting 1181 patients with preoperative imaging that also passed data review and were subsequently included in survival analysis are summarized in Table 4.5. As expected, patients with higher grade tumors tend to be older, have a decreased rate of IDH1 mutation, and have larger preoperative tumor volume.

Grade	N	Age (mean ± sd)	Sex M/F	median KPS	IDH1 MUT/WT (confirmed)	median tumor vol (cc)	median OS (weeks)
II	207	40 ± 12	121/86	90	91/12	37.16	NA
III	246	43 ± 14	135/111	90	59/24	47.25	498
IV	728	59 ± 13	446/282	90	16/201	72.64	71

TABLE 4.5: Clinical data summary for all 1181 preoperative cases. IDH1 mutation status is listed for cases where IDH1 mutation status was explicitly mentioned in the clinical record. Median overall survival (OS) was not reached for the WHO II group.

	II	III	IV
Anaplastic Astrocytoma	86	163	0
Oligodendroglioma	91	61	0
Mixed Oligoastrocytoma	14	19	0
Glioblastoma	0	0	712
NOS	5	0	0
Other	11	3	16

TABLE 4.6: Histologic diagnoses by WHO grade for all patients. The histologic grades are mostly consistent with the 2012 WHO grading scale. NOS=Not otherwise specified.

4.3.3 Clinical Factors, Tumor Volume, and Survival

Several clinical factors are known to be strongly associated with overall survival including age, mental status (KPS), and high clinical tumor grade (III or IV). Our data showed this relationship as well and the hazard ratios are given in Table 4.7 for all cases combined. The results for each subset of WHO grade are listed in Table A.2, Table A.3, and Table A.4. These factors serve as the benchmark for the prognostic value of imaging biomarkers and compare well with published values. Landmark publications by Sawaya and colleagues analyzed a very similar population (though only Grade IV disease), from our same institution [14, 15].

	HR	CI	p	
KPS < 70	2.142	[1.824, 2.515]	1.41e-20	***
Age > 55	2.664	[2.264, 3.134]	3.43e-32	***
Grade WHO III+	4.251	[3.057, 5.912]	7.70e-18	***

TABLE 4.7: Multivariate cox proportional hazards model hazard ratios and significance for all cases using only clinical factors that strongly influence survival. Confidence intervals are 95%

One simple image-based survival correlate from literature is tumor volume. Optimal stratifications are shown in Figure 4.3 for automatic measurements of T1 enhancing and T2 FLAIR lesion volumes. In the combined patient cohort with all grades, all tumor volume measurements show significant survival differences, see Figure 3.11. Note, these automatic volume measurements agree very well with reference measurements as described in Section 3.3.5. The univariate and multivariate survival stratification based on tumor volumes are listed in Table 4.8 and the same analysis for each subset of WHO grades are given in Table A.5, Table A.6, and Table A.7. Note, for WHO II and III cases there are not enough tumors with and without enhancement to perform a meaningful survival analysis.

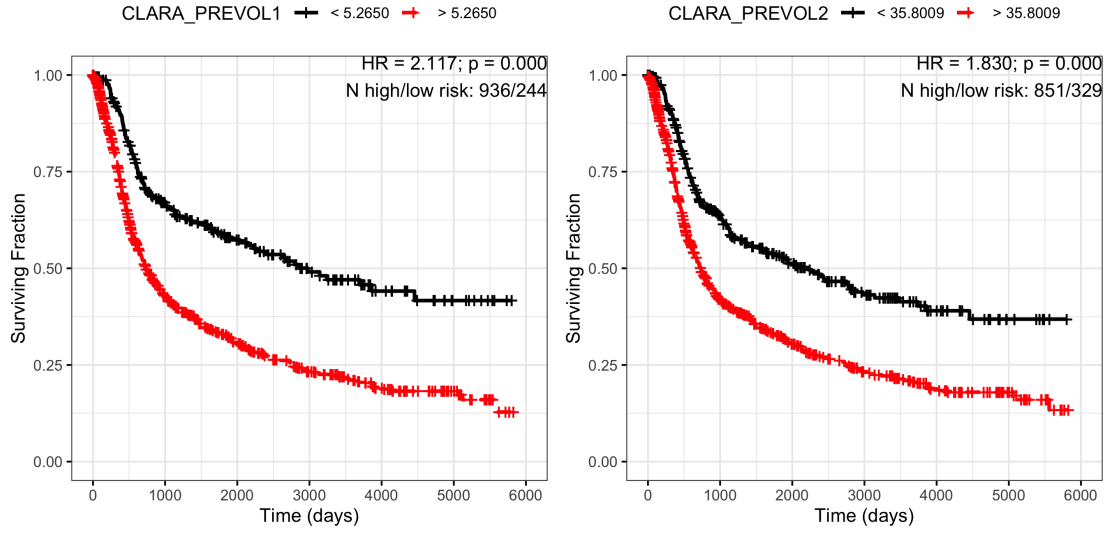


FIGURE 4.3: Survival curves with an optimized cutoff based on preoperative tumor volume. Left: survival stratified by T1 enhancing volume Right: Survival stratified by total T2 visible tumor volume. p-values from log-rank test.

		Univariate			Multivariate (age+KPS)			Multivariate (age+KPS+grade)				
CLARA EV > 5.27	2.117	[1.721, 2.603]	1.21e-12	***	1.701	[1.378, 2.100]	7.57e-07	***	1.582	[1.281, 1.955]	2.11e-05	***
CLARA WT > 35.80	1.830	[1.530, 2.188]	3.50e-11	***	1.574	[1.312, 1.888]	9.98e-07	***	1.389	[1.158, 1.668]	4.11e-04	***

TABLE 4.8: Cox proportional hazards model hazard ratios and significance for tumor volumes, multivariate ratios include controls for age, KPS and grade. EV = enhancing volume, WT = whole tumor

4.3.4 Survival Based on Preoperative Image Features

To investigate relationships between image features, clinical data, and survival we use the C-index, Cox proportional hazard models, and the Kaplan-Meier method [141]. First, we removed features with Spearman rank correlation greater than 0.8 with either total tumor volume or enhancing volume. Without this step, most of the best features are simply reproducing the prognostic effect of tumor volume. Mutual correlations between some selected features are shown in Figure 4.4. For the remaining features in Table 4.3 we identified the ones with the greatest univariate cross-validated hazard ratio for each image type in Table 4.9. In addition to the C-index, we identified the optimal threshold value for each feature using predicted high- and low-risk patients from the 10-fold cross-validation. To find the optimal threshold we searched over the possible values to find the greatest (or smallest) hazard ratio with a (log-rank) significant survival difference between the groups. A minimum group size of 20% is enforced to avoid spurious thresholds.

The following paragraphs describe in detail the single best feature (by hazard ratio) for each image and synthetic pathology map type. Shape features are considered as their own “image” type. The results reported in this section are over the combined cohort of all WHO grades (II, III, and IV). The corresponding results for individual grade subsets are listed in the appendix, Section A.4.2

Panel A: Univariate analysis										
image	region	feature	IS		Univariate					
			C	Cut	HR	95% CI	p			
TC	wholetumor	IntensityAtVolume100	0.681	1.61	5.097	[3.93, 6.61]	1e-33	***		
CD	wholetumor	IntensityAtVolume10	0.662	7.68e+03	4.209	[3.33, 5.33]	7e-32	***		
-	enhanc	VoxelVolume	0.642	0.786	3.479	[2.75, 4.40]	3e-24	***		
ERG	enhanc	IntensityAtVolume10	0.593	3.82	2.936	[2.36, 3.65]	3e-21	***		
Ki67	enhanc	IntensityAtVolume10	0.635	13.7	2.645	[2.27, 3.09]	4e-34	***		
T2	enhanc	IntensityAtVolume10	0.615	0.623	2.385	[1.93, 2.95]	7e-15	***		
T1	wholetumor	Median	0.573	0.74	2.027	[1.64, 2.50]	2e-10	***		
FL	enhanc	Maximum	0.596	3.55	1.835	[1.56, 2.16]	1e-12	***		
GR	higher	Maximum3DDiameter	0.591	43.1	1.424	[1.22, 1.66]	1e-05	***		
Panel B: Multivariate analysis										
image	region	feature	Multivariate (age+KPS)				Multivariate (age+KPS+grade)			
			HR	95% CI	p		HR	95% CI	p	
TC	wholetumor	IntensityAtVolume100	3.549	[2.72, 4.63]	1e-20	***	2.777	[2.12, 3.64]	1e-13	***
CD	wholetumor	IntensityAtVolume10	2.909	[2.28, 3.71]	7e-18	***	2.274	[1.78, 2.91]	7e-11	***
-	enhanc	VoxelVolume	0.642	[1.92, 3.12]	4e-13	***	1.923	[1.50, 2.46]	2e-07	***
ERG	enhanc	IntensityAtVolume10	2.151	[1.72, 2.69]	2e-11	***	1.741	[1.39, 2.18]	2e-06	***
Ki67	enhanc	IntensityAtVolume10	1.650	[1.40, 1.95]	4e-09	***	1.339	[1.13, 1.58]	6e-04	***
T2	enhanc	IntensityAtVolume10	1.807	[1.45, 2.25]	1e-07	***	1.559	[1.25, 1.94]	7e-05	***
T1	wholetumor	Median	1.643	[1.33, 2.03]	5e-06	***	1.598	[1.29, 1.98]	2e-05	***
FL	enhanc	Maximum	1.453	[1.23, 1.72]	1e-05	***	1.333	[1.13, 1.57]	7e-04	***
GR	higher	Maximum3DDiameter	1.219	[1.04, 1.42]	1e-02	*	1.106	[0.95, 1.29]	2e-01	

TABLE 4.9: Best image features from each image type among patients with all WHO grades (II III IV) in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate (Panel A) and multivariate (Panel B) models. p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons.

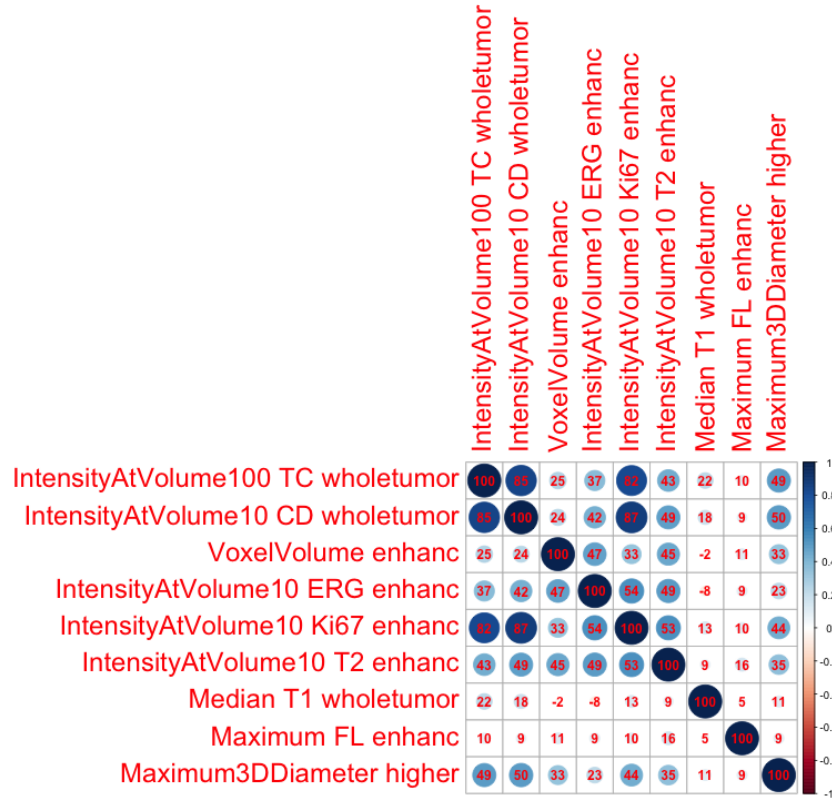


FIGURE 4.4: Visualization of the correlation matrix between the features with largest hazard ratios, Table 4.9. Individual Pearson correlations are listed as percentages (i.e. 85 instead of 0.85). The mutual correlations are reduced since we only select features which are correlates less than 0.8 with enhancing or total tumor volume.

TC: Maximum T1 Enhanced Intensity A stabilized maximum intensity of >1.624 on T1 contrast enhanced images using a whole tumor volume of interest is a poor prognostic indicator with a univariate hazard ratio of 5.305, Figure 4.5. This was the largest univariate hazard ratio observed among all image features. The maximum contrast enhanced intensity (with 0.1 cc volume constraint) also had the strongest overall correlation with survival ($C = 0.681$). Furthermore, we found good stability of the optimal cutoff value with a standard deviation of just 0.034 (2%). The in-sample cutoff of 1.61 falls less than one standard deviation from the mean cross-validated cutoff of 1.624, showing good generalization. Figure 4.5 visually shows this good agreement between the in-sample survival cutoff for high risk/low risk cases and the 10 individual cutoffs from cross-validation since the pooled across folds predictions line up almost exactly with the in-sample predictions.

The survival difference remained significant in multivariate analysis with age and KPS

taken into account, although the hazard ratio was reduced to 3.55. This is a large survival difference, exceeding in magnitude that of age or KPS, but still smaller than the comparable hazard ratio for high WHO grade of 4.25 (Table 4.7). Multivariate analysis accounting for tumor grade and contrast enhancement at the same time, showed the hazard ratio for high maximum TC to decrease somewhat to 2.78. However, the relation is still significant indicating that the contrast enhancement adds prognostic value even after tumor grade information is taken into account.

The strong performance on maximum contrast enhancement is not surprising because enhancement is a hallmark characteristic of high-grade gliomas and these high-grade gliomas have much worse prognosis. However, the max TC feature captures more than just the presence of contrast enhancement, it provides a threshold for the maximum brightness of the enhancement at about 1.6 times the brightness white matter relative to CSF.

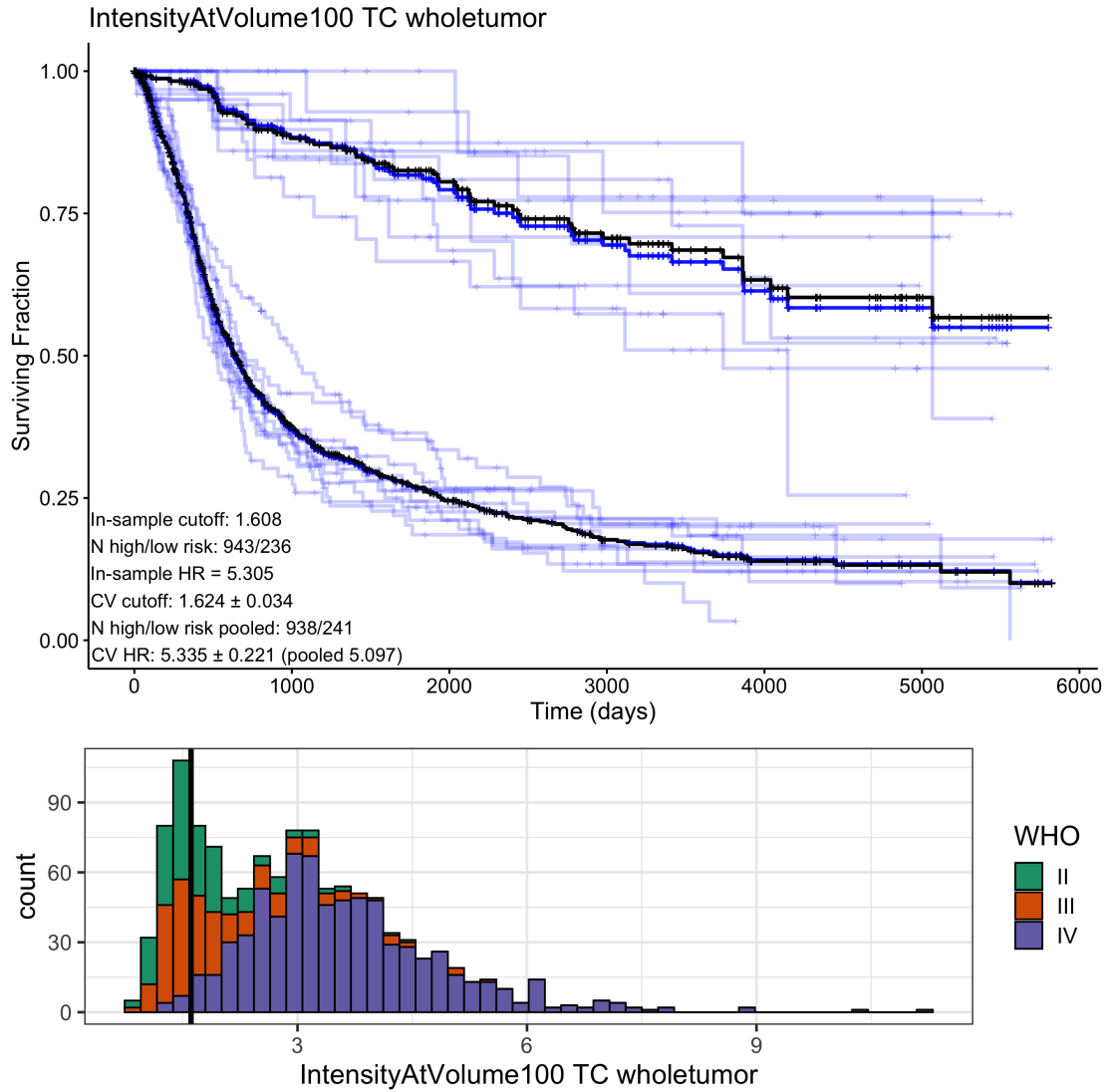


FIGURE 4.5: Best feature for stratifying survival for WHO grades II, III, and IV cases for TC image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

Cell Density: Maximum Cell Density Maximum cell density >7680 nuclei/mm² over the visible tumor volume was a strong indicator of poor prognosis with a univariate hazard ratio of 4.209, Figure 4.6. The maximum feature was stabilized by enforcing a 0.01 cm³ volume constraint. Maximum cell density also showed a high concordance with survival ($C = 0.662$), which is comparable to the best overall feature from contrast enhancement, Table 4.9. The optimal threshold was incredibly stable in cross-validation with only one patient switching between high and low risk groups between cross-validation and in-sample results.

The large hazard ratio of 4.2 was moderated by the inclusion of clinical factors age and KPS which reduced the independent hazard ratio to 2.91 while maintaining statistical significance. This multivariate hazard ratio is larger than both age and KPS (2.66, 2.14) which suggests maximum cellularity may be a stronger prognostic factor than these clinical covariates. It is, however, still smaller than the effect of WHO grade with a hazard ratio of 4.25. When max CD and high WHO grade are considered together, CD still maintains an independent hazard ratio of 2.27, comparable to age and KPS. So, maximum cellularity clearly represents independent prognostic information.

Heightened cellular density (CD) is caused by tumor growth constrained by brain anatomy as well as infiltration into surrounding normal brain. This feature essentially measures the maximum cellularity inside the visible tumor lesion. As previously mentioned, this heightened cellularity is a hallmark of highly malignant tumor hence the large hazard ratio between groups. Interestingly however, the effect was not quite as strong as is seen with contrast enhancement (Table 4.9) despite the two being 0.85 rank correlated.

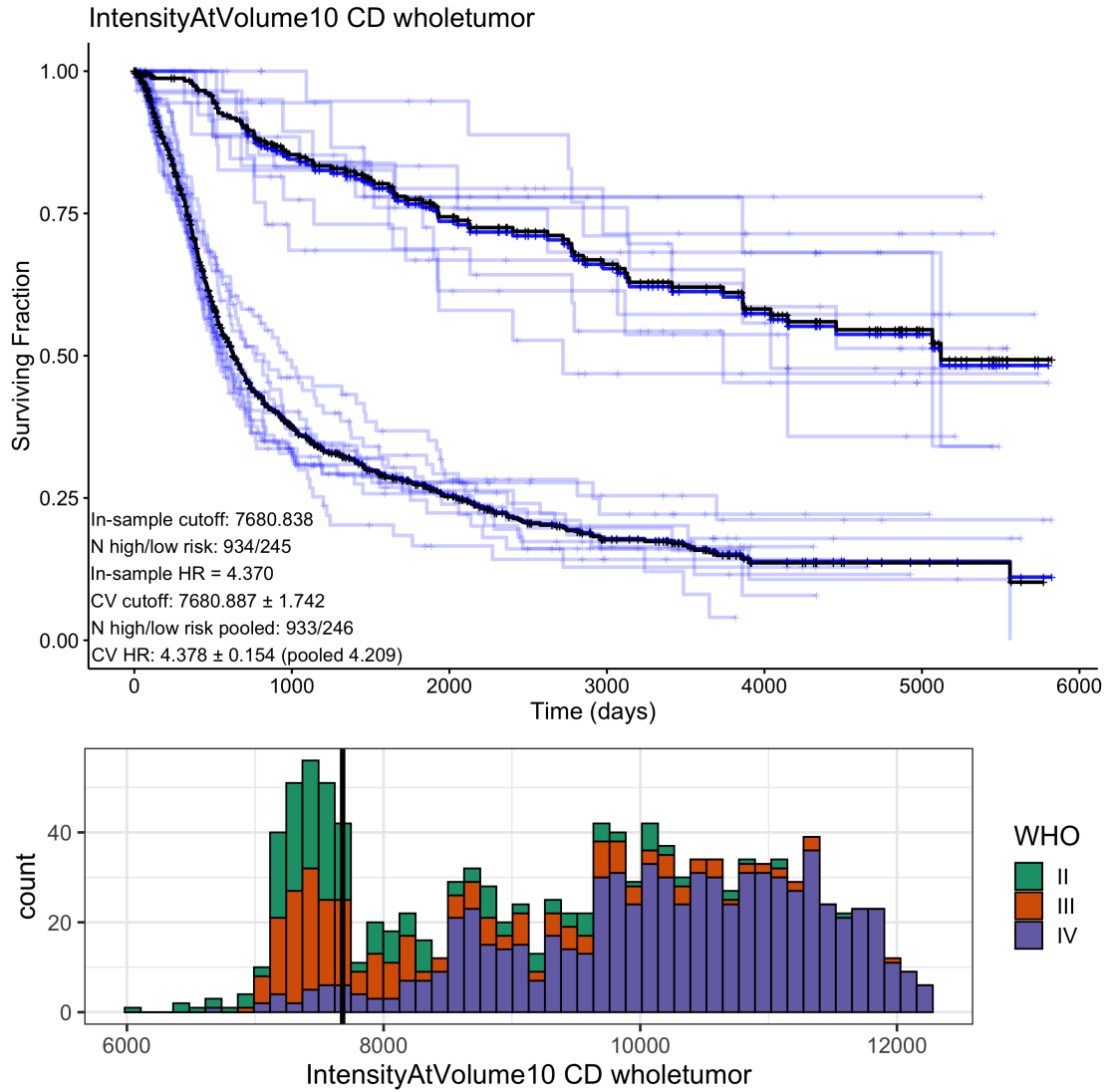


FIGURE 4.6: Best feature for stratifying survival for WHO grades II, III, and IV cases using the estimated cell density map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

Shape: Enhancing Tumor Volume The best shape feature was the volume of the enhancing tumor region $>0.786 \text{ cm}^3$ which indicated a worse prognosis with univariate hazard ratio of 3.48, Figure 4.7. The small cutoff near zero means this feature is essentially detecting the presence or absence of contrast enhancement. We saw a high concordance index of $C = 0.642$ and a very stable optimal cutoff value. The in-sample value of 0.786 was nearly identical to the mean of the cross-validated thresholds at 0.788 (standard deviation 0.007). In Figure 4.7 the in-sample and cross-validated survival

curves are indistinguishable meaning there was effectively no overfitting.

In binarized cross-validation, hazard ratio was 3.48 between groups. This generally agrees with the tumor volume analysis (Table 4.8). But, the cross-validated hazard ratio of 3.48 is much larger than the in-sample hazard ratio of 2.11 in Table 4.8. This is likely because of the difference in volumes being measured. In order to compare to reference values, the CLARA segmented enhancing volume in Table 4.8 combined enhancing and non-enhancing labels as described in Table 3.15. However, the enhancing voxel volume image feature in Table 4.9 counts only the enhancing label shown in blue in Figure 3.7. Note that the prognostic information in the enhancing tumor volume is more predictive of survival than the total tumor volume.

The univariate hazard ratio for enhancing volume was reduced by the multivariate inclusion of age and KPS, falling to 2.45. This put in on approximately equal footing with age and KPS (hazard ratios 2.7 and 2.1 respectively) but still much smaller than the hazard ratio of 4.3 for WHO grade. In multivariate analysis accounting for grade and presence of contrast enhancement, we still found an independent hazard ratio of 1.92, which supports enhancing volume as an independent predictor. Similarly to the previously mentioned max TC feature, these results make sense due to the known importance of contrast enhancement. While Max TC is a measure of enhancing brightness, this feature captures its presence or absence which can indicate fundamental differences in the tumor vasculature. The presence of a biologic “switch” for angiogenesis is well known for gliomas, and is the biologic feature that corresponds most closely higher grade and poor prognosis [145].

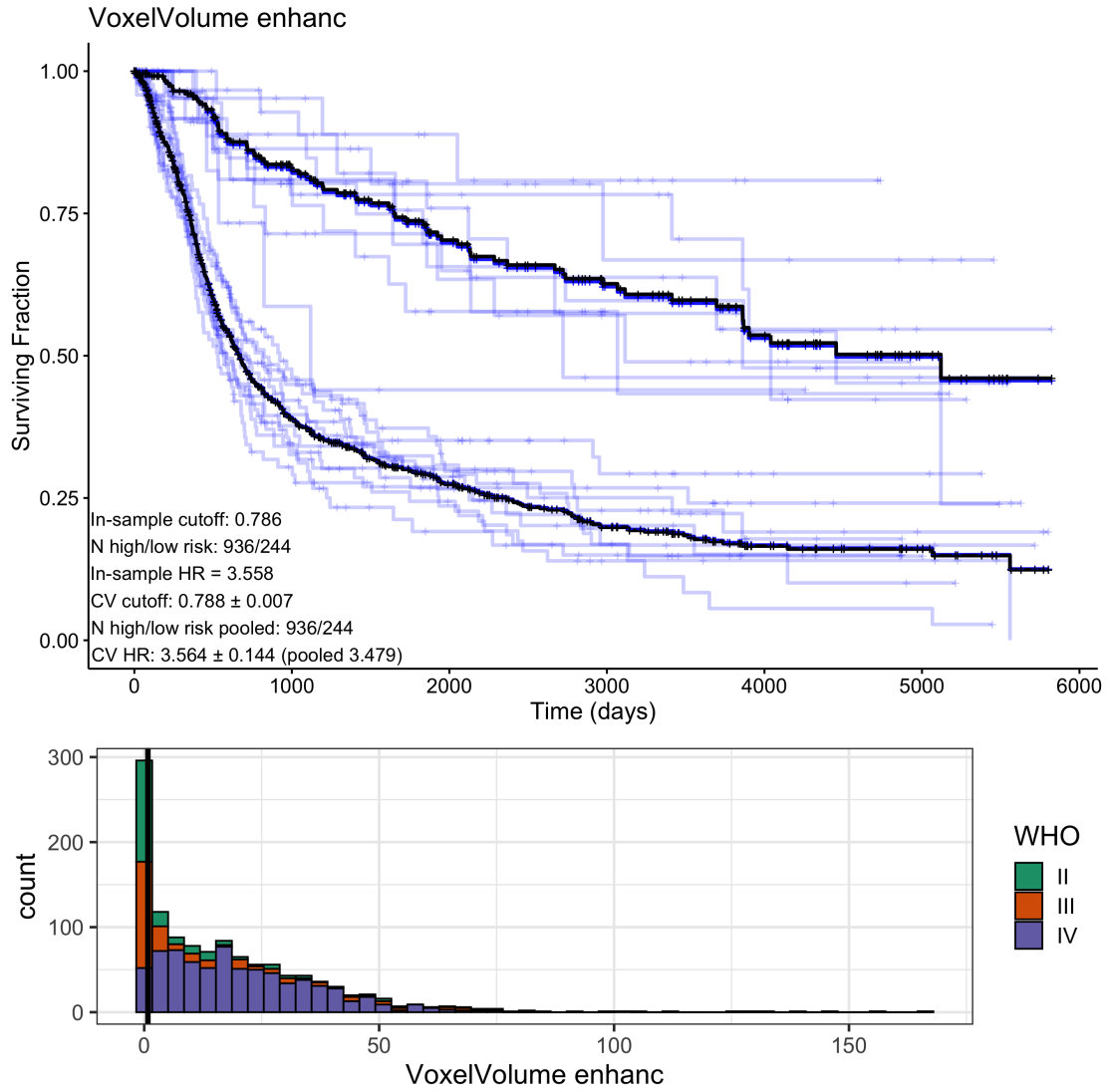


FIGURE 4.7: Best shape feature for stratifying survival for WHO grades II, III, and IV cases. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

ERG: Maximum ERG Expression Similar to the cell density and Ki67, we found a maximum ERG expression ($> 3.82\%$ positive area) in the enhancing tumor sub-region to be a negative prognostic factor with a univariate hazard ratio of 2.96. This is lower than the best features based on contrast enhancement and cell density with hazard ratios > 4 but still very considerable, Figure 4.8.

The maximum ERG (with 0.01 cm^3) constraint had a C-index of 0.593 which was comparable to many of the other features in Table 4.9. The optimal threshold of 3.82 was also very stable under cross-validation as evidenced by almost perfectly superimposed

survival curves in Figure 4.8. The univariate hazard ratio was reduced in multivariate modeling with age and KPS to 2.15, although statistical significance was retained. This means a fair portion of the information is redundant with these clinical factors. The hazard was reduced further by the inclusion of age, KPS, and high WHO grade to 1.74. Again, there is some overlap between maximum ERG and WHO grade, but nonetheless it retains significance as an independent prognostic factor.

It makes intuitive sense that ERG expression, related to vascularity, would correlate with prognosis. High grade gliomas tend to have increased vascular recruitment and angiogenesis as part of their aggressive profiles. Interestingly however, the univariate hazard ratio for maximum ERG is not as large as for contrast enhancement which is based on tissue vascular permeability. The overall correlation is also fairly small at just 0.37 (Figure 4.4). It may be the case that the presence of vascular proliferation measured by ERG is not generally sufficient to create the leaky vasculature necessary for contrast extravasation. Overall though, the 0.1 cc max ERG intensity showed significant prognostic value relative to age, KPS, and grade with a multivariate hazard ratio of 1.74.

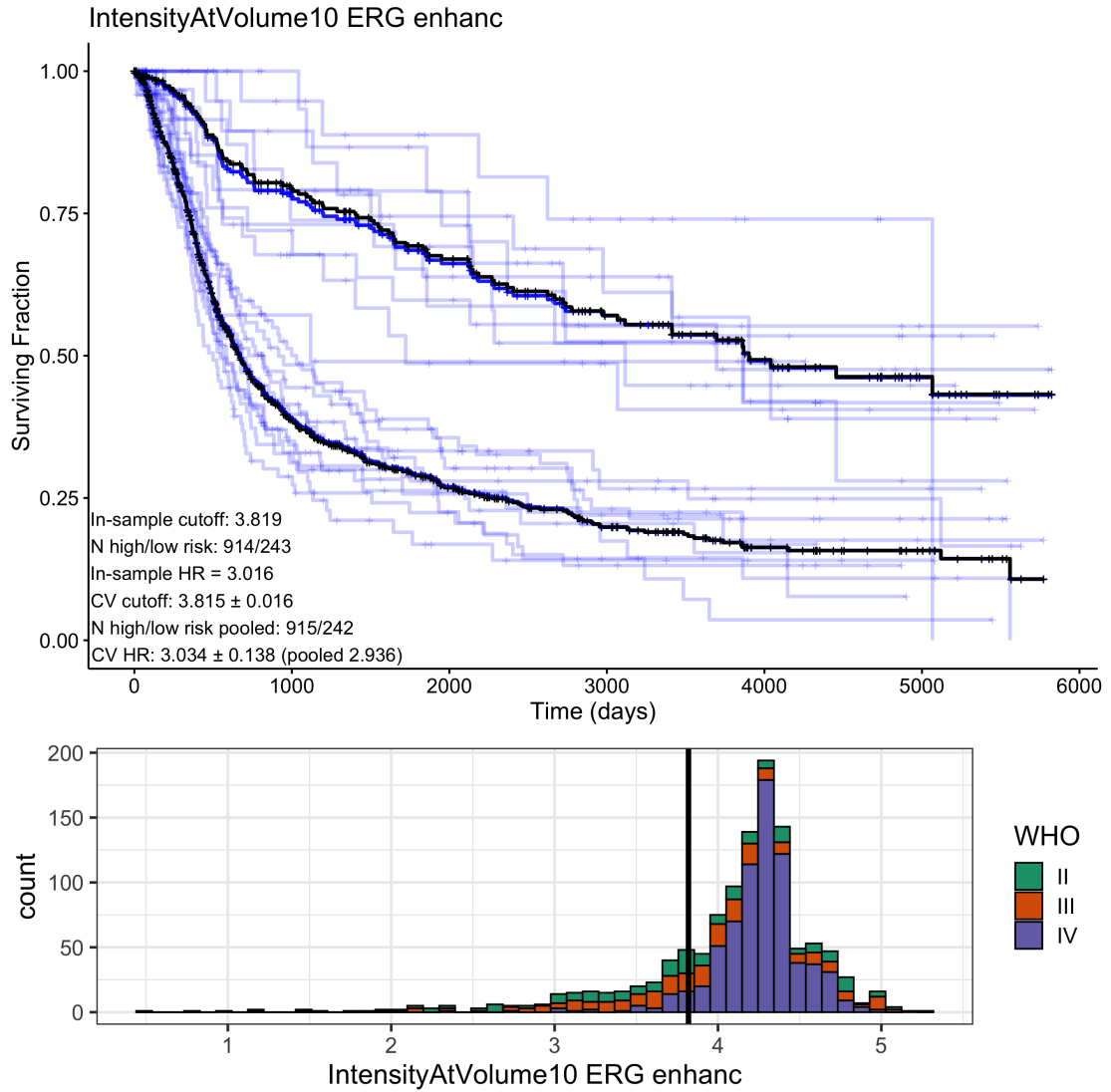


FIGURE 4.8: Best feature for stratifying survival for WHO grades II, III, and IV cases for the estimated ERG map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

Ki67: Maximum Ki67 Expression We found estimated Ki67 $>13.7\%$ in the enhancing tumor sub-region was associated with poor prognosis and a univariate hazard ratio of 2.65. The C-index for max Ki67 (0.01 cm^3 volume) was fairly high at 0.635 and the optimal threshold was stable under cross-validation having a standard deviation of just 0.6 percentage points. The optimal threshold also falls nicely around the center of the data set with 45% of the data in the high-risk category and 55% in the low-risk category, Figure 4.9.

With the inclusion of age and KPS, the hazard ratio between groups was reduced considerably to 1.65 although it retained statistical significance. This suggests that the highly proliferative tumors may preferentially occur in older patients or correlate with deteriorating mental status. Including WHO grade only reduced the multivariate hazard ratio a small amount to 1.34, which was still significant. Interestingly, the multivariate hazard ratio for the Ki67 feature was smaller than the multivariate ratios for T2 and T1 features which had smaller univariate hazard ratios.

The correlation with clinical factors makes sense based on the fact that older patients tend to present with IDH wild-type high-grade tumors which are more aggressive (proliferative) than the IDH mutant counterparts. So, clinical factors have more overlap with the effect of proliferation. Overall though, this focal measure of proliferation indeed represented an independent prognostic factor. However, note that even small pockets of about 0.01 cc of highly proliferative tissue could indicate poorer prognosis, which is expected given that heightened proliferation is a hallmark of aggressive disease [13, 30].

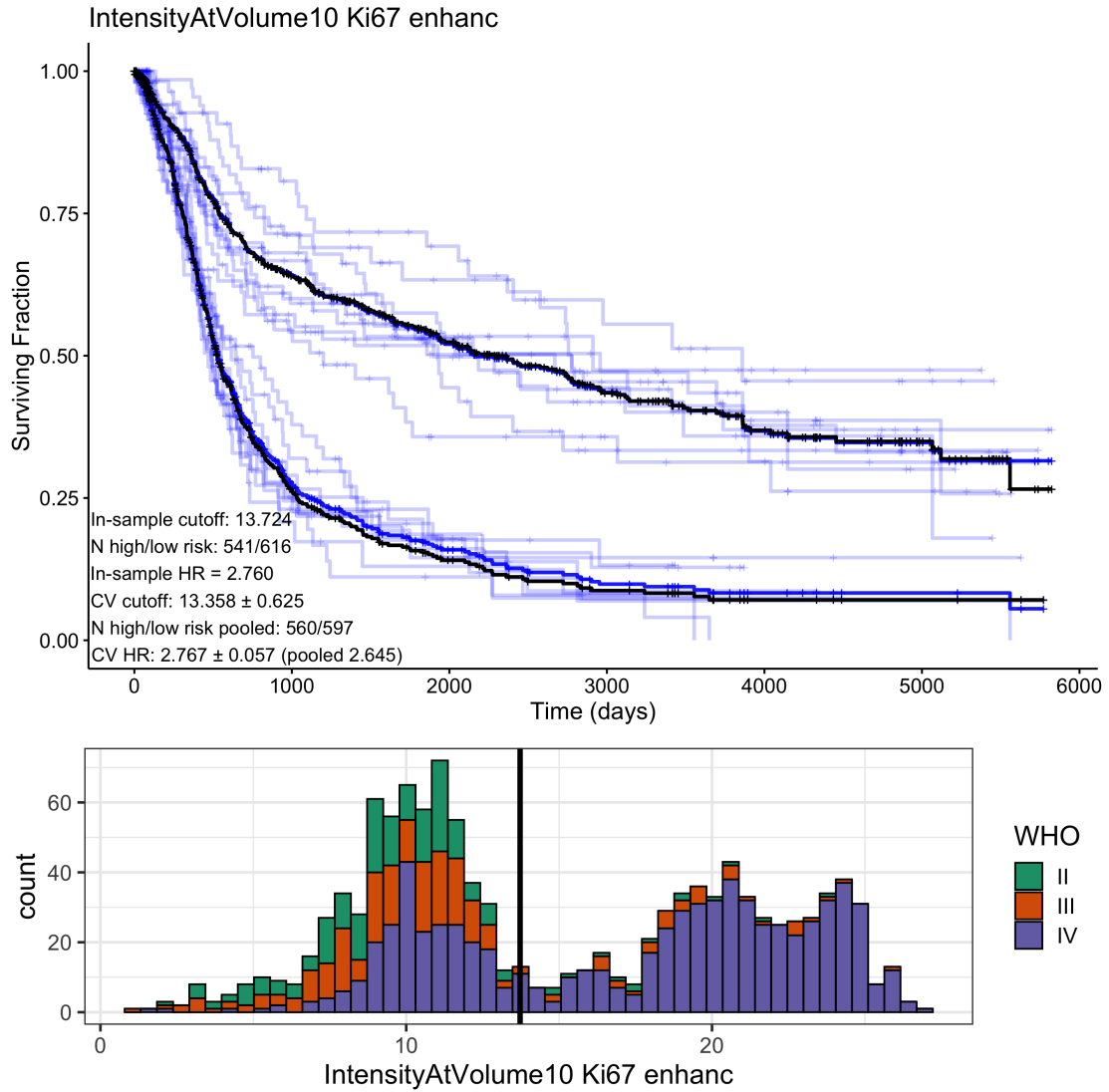


FIGURE 4.9: Best feature for stratifying survival for WHO grades II, III, and IV cases for the estimated Ki67 map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

T2: Maximum T2-Weighted Intensity The best T2-weighted image feature was the maximum intensity over the enhancing tumor sub-region. A value of 0.623 or greater was associated with a worse prognosis with univariate hazard ratio 2.39, Figure 4.10. This was a smaller hazard ratio than the best features from Ki67, CD, and ERG which suggests there may be less prognostic information than is contained in the synthetic pathology maps. Nonetheless, we observed a good concordance index of 0.615 for maximum (0.01 cm³ constraint) as well as excellent generalization. The cross-validated cutoff

values were 0.624 ± 0.003 and the in-sample optimal cutoff was 0.623. Again, the survival curves in Figure 4.10 show indistinguishable in-sample and cross-validated survival curves.

The difference in survival remained significant with age and KPS accounted for in multivariate modeling. The hazard ratio for high max T2 decreased to 1.81, making it less prognostic than age and KPS which have hazard ratios > 2 . With tumor grade also accounted for, the hazard ratio only slightly decreased to 1.56 and retained significance which again suggests that the T2w image adds independent prognostic information to tumor grade.

This feature is similar to the best overall feature of 0.1 cc TC intensity as it represents maximum intensity within a sub-region of the tumor. There is a moderately high correlation between these two features of 0.43 as illustrated in Figure 4.4 which suggests some redundancy of information. Although, the 0.1 cc TC intensity picks up strongly on the differences between enhancing and non-enhancing tumors whereas the 0.01 cc T2 intensity is more general since all gliomas generally appear T2 bright. The in-sample optimal cutoff value around 0.62 (on a scale from modal brain intensity to CSF intensity) suggests a maximum T2 brightness threshold for identifying highly aggressive pathology with worse prognosis.

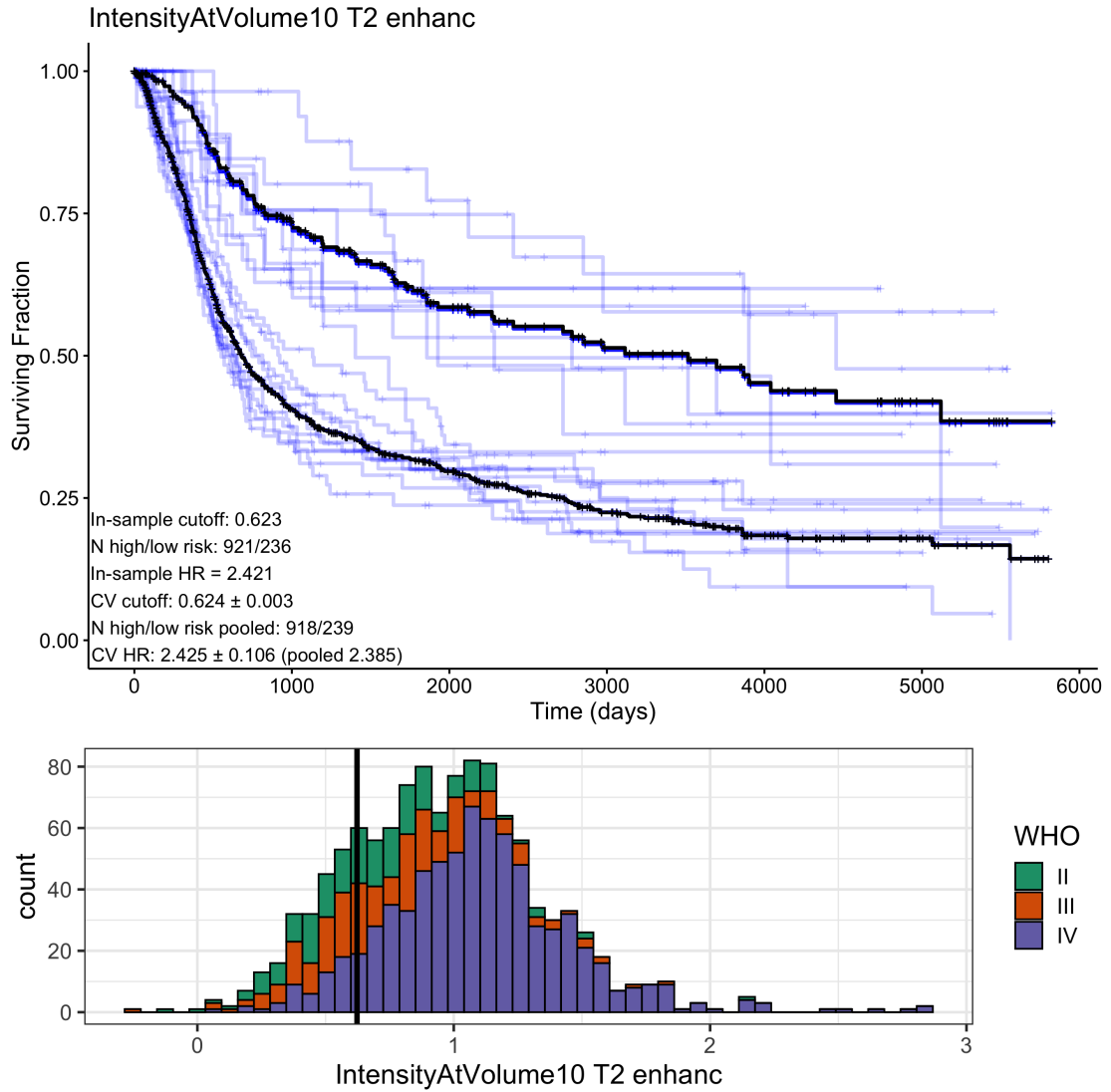


FIGURE 4.10: Best feature for stratifying survival for WHO grades II, III, and IV cases for T2 image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

FLAIR: Pointwise Maximum FLAIR Intensity The best feature based on the FLAIR image a voxel-wise brightness inside the segmented enhancing volume > 3.55 which was associated with a worse prognosis with univariate hazard ratio of 1.84. The intensity scale is CSF mode to brain mode so this number corresponds to a maximum tumor intensity which is about 3.5 times as bright as the brain overall. The point-wise maximum had a good C index of 0.596 and good agreement between the in-sample threshold (3.548) and the mean cross-validated cutoff of 3.554. Figure 4.11 shows a very slight separation of the in-sample and cross validated curves in the high-risk group but

this is negligible.

Comparing point-wise maximum FLAIR alongside age and KPS moderated the hazard ratio down to about 1.45. This was still significant but smaller than the effects of KPS and age (hazard ratios 2.7 and 2.1). Further inclusion of tumor grade in the multivariate model alongside age, KPS, and max FLAIR pushed the hazard ratio down just a little more to 1.33. This effect was still significant and so the max FLAIR does indeed provide independent information to tumor grade.

This feature can be interpreted much in the same way as the T2 intensity: representing a focal T2-weighted hyperintensity. Interestingly though, the rank correlation between the maximum FLAIR intensity and 0.01 cc T2 intensity is very small at 0.16 (Figure 4.4). This might be due to instability in the point-wise maximum feature.

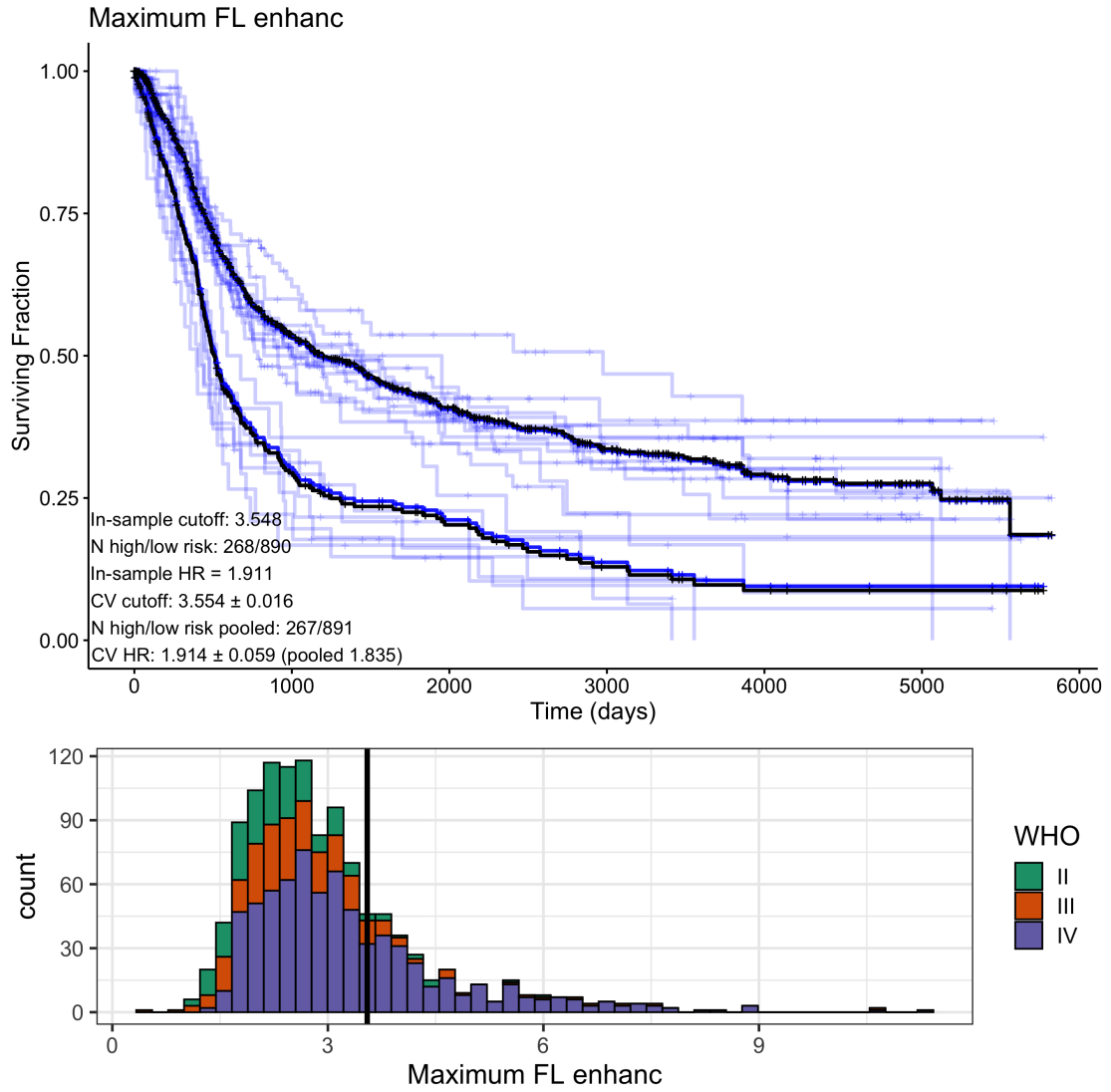


FIGURE 4.11: Best feature for stratifying survival for WHO grades II, III, and IV cases for FLAIR image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

T1: Median T1-Weighted Pre-Contrast Intensity The median T1 pre-contrast brightness over the whole tumor volume of interest was the best performing feature from the T1w image. A brighter median >0.740 was associated with a worse prognosis with univariate hazard ratio 2.037, Figure 4.12. This was slightly better than the best FLAIR feature (HR 1.83) but lower than all the other raw image features. Concordance between median T1 intensity and survival was more modest at 0.573, the smallest among all features in Table 4.9. However, there was still very good generalization shown with the cross-validated thresholds (mean 0.739) and the in-sample threshold of 0.740.

The univariate hazard ratio of 2.04 was reduced slightly in multivariate analysis compared with age and KPS to 1.64, which is smaller than the hazard ratios for KPS and age but still considerable and significantly prognostic. Further, when grade was also included as a covariate, the hazard ratio only decreased to 1.60 and remained significant. This means that tumor grade does not account for much of the effect of high median T1 intensity that is not already accounted for by age and KPS.

The hazard ratios greater than 1 suggest brighter T1w intensity corresponds to worse prognosis. This is contrary to the usual appearance of gliomas which is T1 hypo-intense. More investigation is needed to discover what is being detected, possibly the presence of methemoglobin or other blood products. Petechial hemorrhages are frequently apparent in high grade gliomas on pathology. The difference in survival may be due to differences in the various grades above and below the 0.74 threshold value. About two-thirds of the WHO II and WHO III cases are above the threshold but nearly 87% of WHO IV cases are above the cutoff. So, the survival difference may be due in part to separation of the WHO IV cases. This may not have been detected by our multivariate analysis that groups the WHO grades III and IV into a single high-grade class.

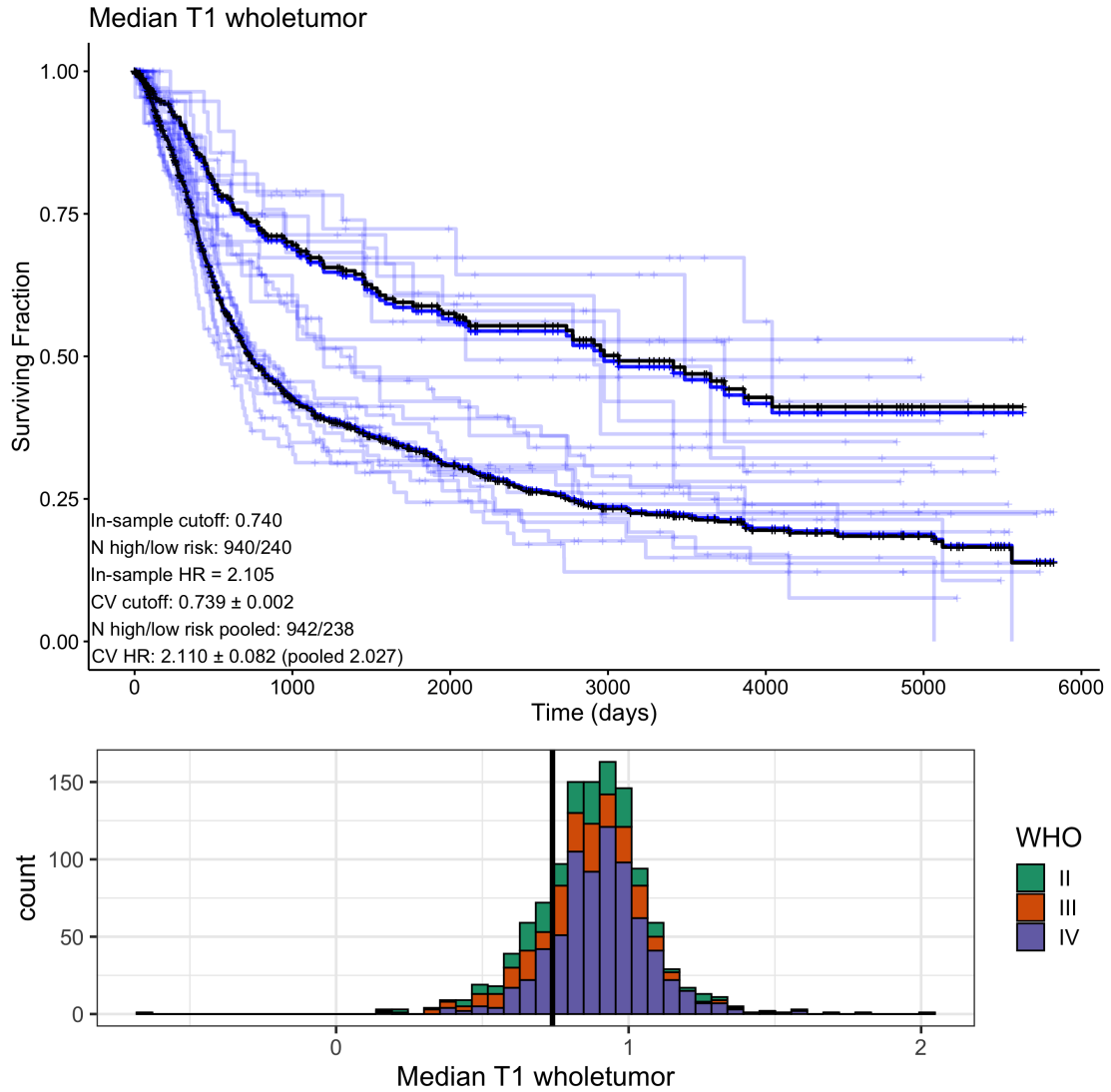


FIGURE 4.12: Best feature for stratifying survival for WHO grades II, III, and IV cases for T1 image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

Local Grade: Diameter of Higher-Grade Region When using a binary threshold, the maximum diameter of predicted high grade disease was significantly prognostic with a hazard ratio of 1.42 in cross-validation. This corresponded to a threshold of 43.1 mm, Figure 4.13, and a larger diameter meant worse prognosis. Although this was the smallest univariate hazard ratio in Table 4.9, high-grade diameter had a C-index of 0.591 which was only the second smallest and comparable to features with larger hazard ratios like maximum ERG. There was some instability observed in the 43.1 mm threshold though. The mean cross-validated threshold of 43.72 mm agreed with the

in-sample of 43.12, however the standard deviation was 9.99 mm. We can also see some overfitting in Figure 4.13 since the in-sample black curves are clearly outside the pooled cross-validated predictions (dark blue). So, the in-sample results should be used with caution.

Grade diameter was still prognostic in multivariate modeling with age and KPS, although with a comparable small hazard ratio of 1.22. Interestingly but expected, higher grade diameter was the only image feature in Table 4.9 which was not independently prognostic when compared against high tumor grade. Being redundant with tumor grade fairly expected since the WHO grading scale is meant to separate groups of tumors based on whether they contain high grade disease. Overall, diameter of predicted high grade disease was a univariate significant predictor although it did show some overfitting. It seems to be capturing much of the same information as high WHO grade which is expected based on its design.

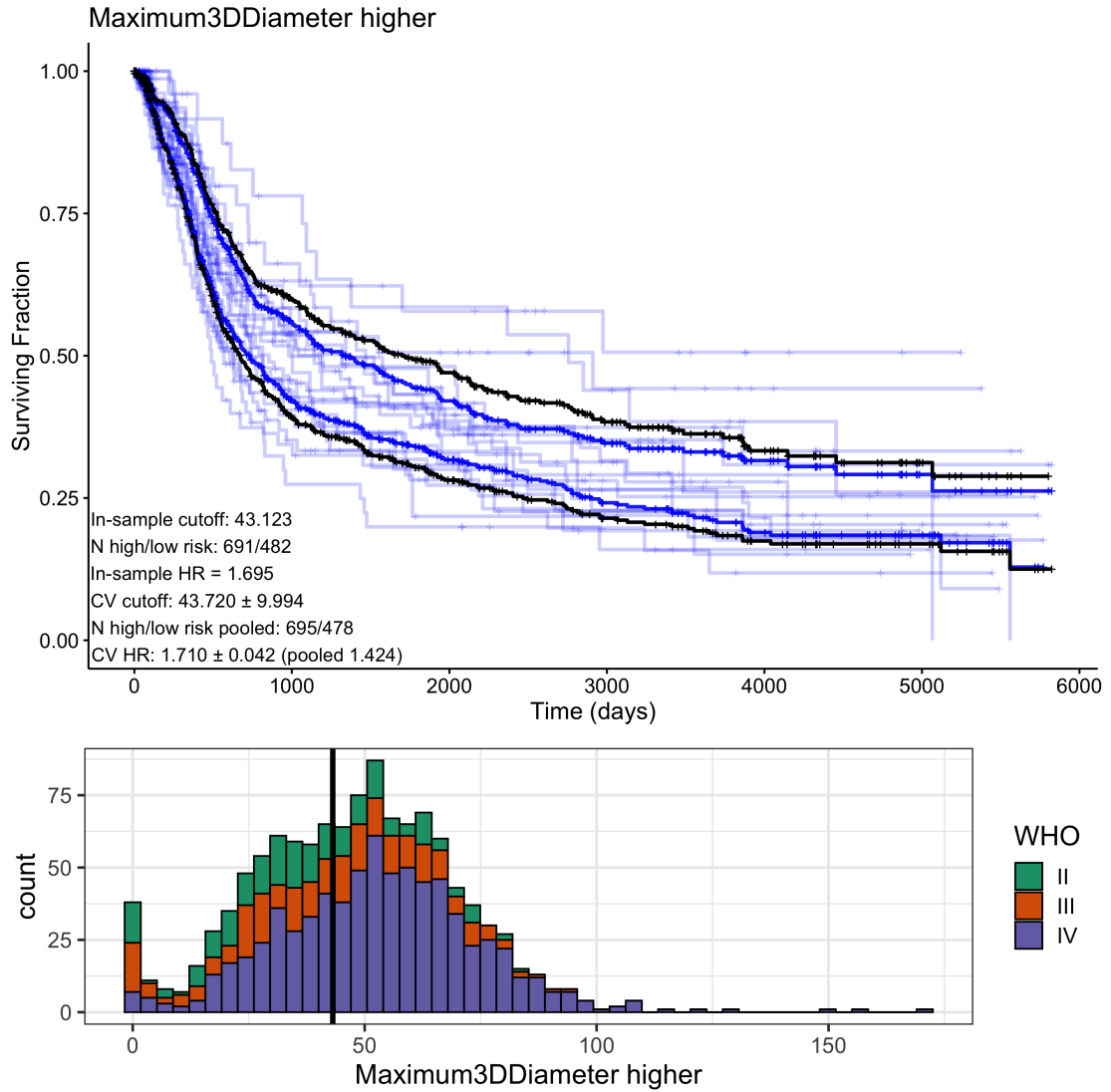


FIGURE 4.13: Best feature for stratifying survival for WHO grades II, III, and IV cases for the estimated grade map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

Subset Analysis Overall the T1 contrast enhanced features outperform the estimated pathology features. However, one case where the estimated pathology maps performed substantially better than the raw image features was for WHO II cases. The 10th percentile of Ki67 and CD over the whole tumor provided very strong stratification with hazard ratios of 6.28 and 10.24 respectively. See appendix Table A.8 and survival curves in Figure A.8. The univariate p values associated with these features does not cross the significant threshold corrected for multiple comparisons among features. However this is likely due to the small number of deaths in the low-risk group. Further

investigation is needed to validate or confirm these interesting results.

Nonetheless this is consistent with clinical intuition that low grade tumors that have overall higher proliferative activity of cellularity tend to be more aggressive and hence have worse prognosis. A higher value in a low percentile means there is an overall increase in cellularity or proliferative activity.

4.3.5 Summary

In summary, the results using predicted pathology maps are similar to the results using raw image features. However, the best features are the ones that correlate with contrast enhancing volume. This confirms that the biology of enhancement (angiogenesis as measured by enhancement) outweighs the biology of proliferation (CD and Ki67) in terms of prognostic significance, and is consistent with the clinical evolution of low to higher grade disease observed in some patients. Contrast enhanced features were highly significant across the whole population. Features derived from cell density were highly prognostic across as well and showed some promise among WHO II cases. This is likely because the pathology estimate maps synthesize the complementary information in the raw images themselves.

4.4 Discussion

We retrospectively examined 1181 untreated glioma patients from MD Anderson Cancer Center and evaluated the prognostic stratification of imaging biomarkers. We found several simple image metrics that provided strong survival stratification, even when accounting for differing preoperative demographics. Specifically, we found the best prognostic stratification among all patients had a hazard ratio of 5.10 (multivariate 2.77) using the 0.1 cc TC intensity over the whole tumor. This is likely due to the enhancing characteristics between high versus low grade tumors which explains the drop in hazard ratio in the multivariate analysis. For WHO grade II cases, the 10th percentile of whole tumor estimated cell density showed an impressive hazard ratio of 10.2 (multivariate 10.8) which far exceeds the best hazard ratio of 4.592 using maximum T2 brightness. Unfortunately, statistical significance was not reached.

We attempted to validate the survival biomarker predictions using cases from the MICCAI Brain Tumor Segmentation Challenge (Section A.5.1). This data set is widely used to benchmark tumor segmentation algorithms and imaging biomarkers. Although, historically the survival prediction task has been very difficult [103]. Unfortunately, we found that the optimal thresholds on prognostic image features trained on the historical data generally failed to validate on the BraTS data. This could be due to a variety of reasons: First, while the BraTS data was processed similarly to the historical data, it was provided already co-registered and skull stripped which might have been performed slightly differently. Also, we extracted features from the provided “ground-truth” segmentations instead of the automatically segmented ROIs from CLARA. While all segmentations were acceptable, the manual versus automatic segmentation procedure might have systematically affected the image features.

Overall, our work confirms the established relationships between clinical factors and survival. In general, patients who are older, have more symptoms at presentation, and have IDH wild-type tumors have a poorer prognosis [14, 15]. An exception is for WHO II cases, where the effect of age is less prominent. However, this may simply be due to the fact that patients with low grade tumors tend to be younger (Table 4.5).

Some previous studies in the neurosurgical literature have examined enhancement brightness and survival. Lacroix et al. measured “enhancement grade” in glioblastoma as the relative brightness of contrast enhancement on T1w images [14]. The grades are low signal intensity, high signal intensity, and signal intensity equal to fat. Only the highest grade was significantly associated with survival (univariate) in untreated patients. This is analogous to the maximum T1 enhancing intensity or the 0.1 cc intensity feature, which we also found to be a strong significant predictor of overall survival in the combined patient cohort.

There is a large body of work examining correlation between image features in glioma, especially glioblastoma, and survival. In general, the “radiomics” field follows the same methodology we employ here [35, 36, 146]. Namely, image processing, feature extraction, and a search over all features for candidate biomarkers. More recently, these features have been incorporated into survival predicting models. Bae et al. used a random survival forest trained on several intensity, shape, and texture features to predict progression free and overall survival in 217 glioblastoma patients [147]. However, the complexity of

the features needed and incremental results shows the need for continued research [148]. Another addition to radiomics is the use of “deep” features based on neural network features. Lao et al. computed nearly 100,000 such features and found a substantial stratification in survival among 112 total patients [149]. The authors employed a LASSO, which uses L1 and L2 regularization on the cox partial likelihood to perform joint variable selection and model training. In this work, we focused on the prognostic ability of individual features, choosing to improve the quality of our building blocks, rather than adding to the complexity of nonlinear feature combinations. Estimated pathology maps already combine multiple sources of information in the same way models like LASSO can use features from multiple MR contrasts.

Perhaps the most similar work to our synthetic pathology estimates is one study by Li et al. who used radiomic features to predict high and low Ki67 expression ($\pm 10\%$) in WHO II and III tumors [150]. They found that the strongest radiographic correlate with high Ki67 was spherical disproportion and a multivariate hazard ratio of 2.37 for high Ki67. Note: they used the actual Ki67 not the Ki67 as predicted by a model.

4.4.1 Limitations

Our preoperative survival analysis is subject to a few common limitations of such studies. First, we had no control over the specific imaging sequences used in the historical data. This is in contrast to the tightly controlled research protocol used to generate data in Chapter 2. While intensity normalization handles much of the variability in image contrast due to different scanners and acquisition parameters, the question of the true accuracy of models estimating proliferation, grade, CD, and ERG on the retrospective data is not entirely answered.

Next, we were only able to acquire ground truth IDH mutation status data for around 25% of the cases analyzed. This is primarily because IDH1 status was not routinely collected before 2015. A similar problem precluded gathering other key molecular factors like 1p19q co-deletion, MGMT methylation, P53, or ATRX mutations [13, 87, 135]. Molecular factors including IDH1 have become extremely important to the diagnosis and prognosis of glioma patients [20] and not having detailed information is a limitation.

4.4.2 Future work

Future work primarily includes expanding the scope of biomarkers to include higher-order texture or deep filter features which may contain additional information. Furthermore, imaging like diffusion-weighted imaging and T2* weighted images were curated as part of Chapter 3 but the image features were not analyzed against survival. In addition, it may be beneficial to continue exploring the interaction between features from different MR contrasts, for instance mean T2w intensity and max T1C intensity together. Although, this adds considerable complexity to the analysis.

Additionally, it may be useful to further subset the patients based on histologic diagnosis or other clinical features which define significant subpopulations. This includes removing patients with NOS or "other" histologic diagnoses in Table 4.6. For now, the analysis is carried out stratifying on WHO grade only with the various subgroups like astrocytoma and oligodendroglioma merged together.

Chapter 5

Biologically Based Extent of Resection

5.1 Introduction

Biomarkers from Chapter 4 show that preoperative quantitative image measurements can identify subsets of patients with differences in baseline risk. This is consistent with the underlying radiomics hypothesis that imaging captures tumor characteristics and heterogeneity that reflect the underlying biology. We explicitly modeled this relationship by predicting the underlying tumor biology via synthetic pathology maps. This allowed us to focus in low-complexity intensity and shape features rather than rely on obscure texture or deep filter features to find meaningful prognostic stratification. Furthermore, using simplified features like mean or maximum intensity makes the resulting biomarkers clearly actionable. For example, maximum proliferation being unfavorable suggests that removing the highly proliferative disease with surgery would be beneficial. Beyond peroperative biomarkers, postoperative and extent of resection (EOR) features present a paradigm to evaluate and potentially guide therapy.

The goal of this chapter is to define and evaluate postoperative image features and extent of resection based on the difference, or “delta” between postoperative and preoperative features. By doing so, we reinforce the prognostic value of image features by showing that they reflect prognosis at more than just the time of diagnosis. Additionally, this improves our understanding of how surgical resection interacts with the heterogeneous

nature of gliomas. As of now, maximal resection is recommended for all glioma patients but is often not possible due to the risk of neurological deficits [41]. It is unclear how the characteristics of residual disease factor into the likelihood of recurrence or death and this complicates decision making during treatment. Furthermore, extent of resection using radiographic findings is poorly defined and inconsistent in its evaluation [151, 152]. Using quantitative image features will help disambiguate this evaluation.

Established literature on extent of resection focuses on removal of contrast enhancement. The benefit of gross total resection (GTR) has been overwhelmingly confirmed by observational and retrospective studies, especially for Glioblastoma [14, 15, 41, 151, 152]. However, there is still no consensus on what constitutes GTR, even for quantitative studies measuring percentage of enhancing tumor removed [152]. Recent evidence has pointed to no difference in benefit from GTR within molecular subgroups of glioma [139]. This suggests that the variance in GTR definitions is not due to differences in effectiveness on different populations, but instead points to a need for a better, clearer, definition of extent of resection.

5.1.1 Summary of Analysis

This chapter examines the prognostic effect of radiographic and histologic heterogeneity in the context of surgical resection. Given the results in Chapter 4 that indeed show MR image features hold prognostic information, we now approach a new hypothesis: *Changes in image features measured before and after surgery inform survival differences for glioma patients.* These “delta features” provide new potential definitions for extent of resection. Furthermore, we expect to find that features quantifying changes in disease burden as estimated from graphical pathology maps will provide better extent of resection measures than raw image features alone.

In this chapter, we collected and processed postoperative data for a large portion of the patients in Chapter 4. The methods for data curation are mostly the same with the addition of a deep learning segmentation model for excluding postoperative cavities. Using the matched preoperative and postoperative data, we first evaluated the postoperative features in the same way as preoperative features. Then, we evaluated the delta features to examine the impact of surgery. We evaluated these postoperative imaging

features and EOR features in the same way as the preoperative features: using the C-index and proportional hazards model. This includes both univariate and multivariate Cox modeling with age KPS, and grade as covariates. For clarity, we focus on the most prognostic feature for each image type (T1, T2, T1C, FLAIR) and estimated pathology map (Ki67, CD, ERG, local grade) over the combined cohort. Alongside the feature analysis we also recreate the classical extent of resection (EOR) measurements based on contrast enhancement in order to compare with known values from literature. This is analogous to tumor volume comparisons from Chapter 4 in the sense that it illustrates how the large-scale feature analysis includes these established results as a subset.

For preoperative features, a larger feature value is generally associated with worse prognosis so hazard ratios were reported using the below-cutoff group as a reference. When measuring extent of resection, the opposite is the case: A larger value (greater delta between pre- and post-resection) corresponds to a better prognosis. Thus, the hazard ratios reported in this section are made > 1 by using the above-cutoff group as the reference.

5.2 Methods

5.2.1 Patient Cohorts and Data

Among the 1380 with preoperative imaging suitable for analysis, 1271 (92%) also had immediate postoperative imaging (within 72 hours) available which allowed measurement of extent of resection. The proportions of WHO grades for these patients were roughly the same as the preoperative data with 241 WHO II gliomas, 256 WHO III, and 774 WHO IV cases. Overall, the postoperative imaging data was of slightly lower quality and fewer high-resolution sequences were available. This is primarily because the purpose of immediate postoperative imaging is to identify any residual enhancing tumor volume and screen for complications like intracerebral hemorrhage or cerebrospinal fluid leaks [153]. Having recently received craniotomy, these patients are also less tolerant to the long scan times needed for high-resolution imaging and often cannot be aligned perfectly in the bore of the scanner. These issues led to increased rates of motion artifacts and more difficult image registrations.

5.2.2 Definitions and Mathematical Description of Extent of Resection

The individual features in Section 4.2.2 define disease burdens in terms of the underlying image (e.g. Volume of T2-FLAIR or Maximum Ki67). An illustration of what EOR captures is illustrated in Figure 5.1. Comparing preoperative (PRE) and postoperative (POST) features for the same patient gives rise to two ways to quantify extent of resection (EOR) as a difference in features. For each feature in Section 4.2.2, we independently calculated the values on the preoperative and postoperative images using their own set of labeled regions. Then we computed the difference of feature values either as a difference or a fraction of the preoperative value.

$$\text{Reduction: } X_L^{\text{PRE}} - X_L^{\text{POST}} \quad (5.1)$$

$$\text{Fractional Reduction: } \frac{X_L^{\text{PRE}} - X_L^{\text{POST}}}{|X_L^{\text{PRE}}|} \quad (5.2)$$

Where X_L^{PRE} and X_L^{POST} are real numbers corresponding a feature listed in Table 4.3, for some fixed image type and region type, on preoperative and postoperative images respectively. The absolute value in the denominator of Equation 5.2 ensures that a greater resection corresponds to a reduction in image feature regardless of sign. In total, 1256 such delta features were extracted. These EOR features are labeled with either **EORreduc-** to notate the signed reduction in feature or **EORfrac-** to denote fractional reduction. For example: **EORreduc-VoxelVolume_enhanc** is the difference of the preoperative contrast enhanced volume and postoperative contrast enhanced volume.

These definitions serve as a logical extension of clinical EOR measurements. Any EOR feature that provided significant risk stratification naturally gives rise to a definition of gross total resection (GTR) and subtotal resection (STR). Clinically, GTR is defined using 98% - 100% of contrast enhancing volume removed [14, 15]. This corresponds to a fractional reduction (Equation 5.2) of the voxel volume feature measured over the enhancing tumor region.

Note, because the preoperative and postoperative images occupied different physical spaces, the voxel values of the images themselves were not directly compared. Each study (preop and postop) also had its own set of labels for brain, tumor, enhancement,

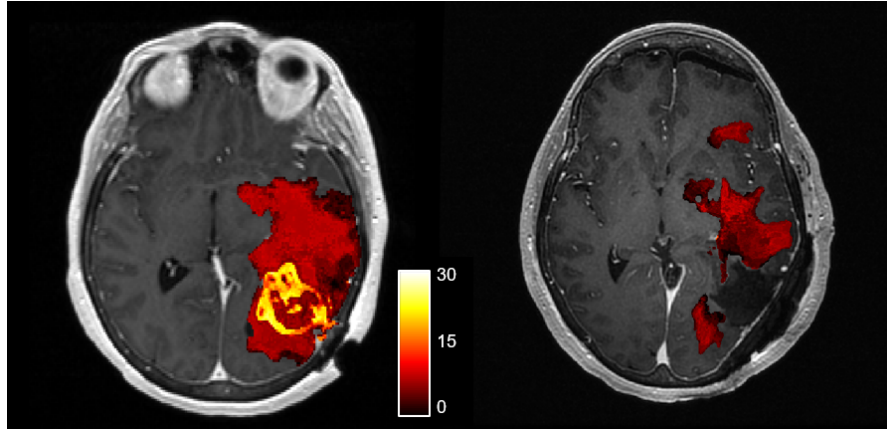


FIGURE 5.1: Estimated Ki67 maps on preoperative and postoperative imaging overlaid on a T1w contrast enhanced image. The high Ki67 areas on the preoperative image have been removed by surgery and a resection cavity is seen. We hypothesized that surgical removal of high-proliferative activity areas would lead to improved survival. The images shown are of the same patient at roughly the same axial location but appear differently due to slice orientation.

etc. Furthermore, these definitions of EOR features do not take into account locality of the image characteristics. For example, a reduction in maximum intensity may be using voxel values that correspond to different anatomical locations in the brain on preoperative and postoperative imaging. Therefore a lot of these features, especially fractional reductions, often took on negative values. Since we are searching specifically for meaningful and interpretable EOR measures, in survival analysis we excluded features that have optimal EOR thresholds less than 0 or show increased overall survival with decreasing EOR feature values. These cases were likely caused by instability in the image features and are not biologically meaningful.

5.2.3 Curation of Postoperative Data

Postoperative images were processed using the same data processing pipeline in Section 3.2.3. The tumor segmentation step was modified for postoperative data to exclude the resection cavity from the residual disease segmentation. The NVIDIA CLARA model was trained to identify gliomas on preoperative using the Brain Tumor Segmentation Challenge data. So, the training data did not contain postoperative changes. In practice, we saw that the model almost always segmented postoperative changes, including the resection cavity, as non-enhancing tumor and gave unacceptable residual disease segmentations. As a solution, we trained a small U-net based deep learning model to roughly segment out the postoperative cavity. We then subtracted the cavity mask from the

tumor segmentation. An example is shown in Figure 5.2. Normal tissue segmentation is performed over brain mask minus any voxels identified as either tumor or resection cavity.

5.2.3.1 Resection Cavity Segmentation Model

The resection cavity segmentation model utilized a modified "pocket" version of the widely used DenseNet architecture [154, 155]. DenseNet uses convolution blocks arranged in down sampling and an up sampling paths. Each path has either four max-pooling or four transposed convolution layers. A channel-wise concatenation operation links each layer in the down sampling and up sampling paths. The blocks consist of two densely connected convolutions followed by point-wise convolution. Each convolution is followed by rectified linear unit (ReLU) activation. However, instead of using a standard DenseNet, where the number of feature maps doubles at each downsampling layer, we used a modified DenseNet where the number of feature maps was kept constant throughout the network. This modification reduced the computational footprint of the model while also retaining segmentation performance.

The dataset used for training the model contained 64 patients, where each patient has a set of T1, T2, T1C, and FLAIR images. Segmentation masks were manually drawn by a neuroimaging researcher (1 year experience) using ITK-SNAP [156]. Of the 64 patients, 15 came from publicly available datasets (BraTS 2013 [34], Ivy GAP, TCGA-GBM, and TCGA-LGG [59]), and the remaining 49 were curated from internal a sample of the historical patients analyzed in this study. For pre-processing, each image was resampled to an isotropic voxel resolution of $1 \times 1 \times 1 \text{ mm}^3$, zero-padded to size $240 \times 240 \times 160$, and normalized using z-score intensity normalization. We trained on 80% of the images and use the excluded 20% as a validation set. After training the model to convergence, the mean validation Dice score was roughly 0.80. The model was implemented in Python using the Keras toolkit (version 2.1.6-tf) and trained on an NVIDIA Quadro RTX 6000 GPU.

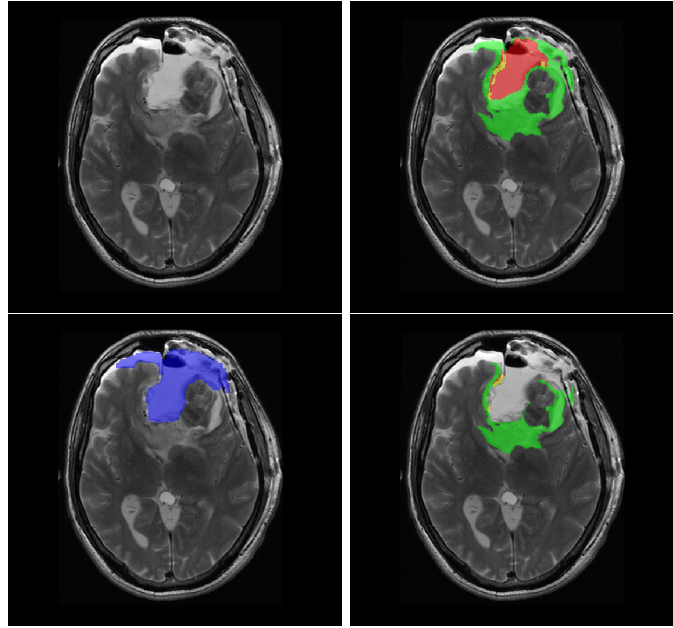


FIGURE 5.2: T2-weighted postoperative image and illustration of residual disease segmentation. CLARA (top right) provides a segmentation of residual tumor that also falsely segments the resection cavity as disease. A custom-made U-net identifies the postoperative cavity (bottom left) and the cavity is subtracted from the tumor segmentation to more accurately label the residual volume (bottom right). Red: non-enhancing tumor, yellow: enhancing tumor, green: edema, blue: resection cavity.

5.2.4 Survival Analysis

Survival analysis of postoperative and extent of resection (EOR) features followed the same methodology as the preoperative features, Section 4.2.3. In short, we evaluated the C-index and optimal binary threshold for each feature in both univariate and multivariate (age, KPS, WHO grade) analyses. As before, we applied some basic filtering on the features as well. First, any features highly correlated (Spearman > 0.8) with postoperative enhancing volume, total residual volume, or reduction in tumor volume were removed, except the enhancing and total volume features themselves. We also removed features that were undefined or missing for more than 25% of the cases in a given patient cohort. These were almost exclusively 1 cc and 10 cc Intensity Volume features that are undefined if the given sub-region is less than 1 or 10 cc respectively. Lastly, we also filtered out features that showed unrealistic correlations with survival due to the unstable nature of some EOR features. These features had EOR thresholds less than zero or had greater extent of resection conferring a worse prognosis. This allowed our search for the best features to identify the most useful and independently prognostic candidates.

5.3 Results

5.3.1 Key Results

- Reduction in enhancing volume, measured manually or automatically, led to improvement in survival. For patients with total resection of enhancement, reduction in automatically segmented T2-FLAIR volume also showed survival benefit in high-grade (WHO IV) cases with hazard ratio 1.42. This is consistent with literature [14, 15].
- Postoperative image features from raw images were significantly associated with survival independent of clinical factors whereas features from synthetic pathology measurements were generally not.
- EOR features based on estimated biological heterogeneity of cell density was more prognostically significant than raw-image based features. Reduction in overall cellularity gave a univariate hazard ratio of 1.9, which is larger than the conventional measure based on contrast enhancement at a hazard ratio 1.7.
- Further investigation is necessary to validate these findings due to the large amount of treatment effect on postoperative data.

5.3.2 Patient Data Summary

Among 1271 patients with paired preoperative and postoperative imaging, 811 had processed imaging that passed both preoperative and postoperative review. A greater proportion of the postoperative data failed the review procedure due to image quality issues. Specifics are given in Section 3.3: Table 3.10 and Table 3.11. Among the cases in the final cohort, 564 had a high-resolution 3D T1 post-contrast or FLAIR image used as a fixed image. The remainder used lower-resolution T2 weighted or FLAIR images. Summaries of the clinical data and tumor histologies included in the final analysis are given in Table 5.1 and Table 5.2.

Grade	N	Age (mean ± sd)	Sex M/F	median KPS	IDH MUT/WT (confirmed)	median postop tumor vol (cc)
II	144	40 ± 13	86/58	90.00	63/8	8.96
III	164	43 ± 13	89/75	90.00	44/17	12.70
IV	502	59 ± 12	307/195	90.00	9/148	22.89

TABLE 5.1: Clinical data summary for all cases with preoperative and postoperative imaging. IDH1 mutation status is listed for cases where IDH1 mutation status was explicitly mentioned in the clinical record.

	II	III	IV
Anaplastic Astrocytoma	62	106	0
Oligodendroglioma	63	45	0
Mixed Oligoastrocytoma	8	11	0
Glioblastoma	0	0	491
NOS	3	0	0
Other	8	2	12

TABLE 5.2: Histologic diagnoses by WHO grade for all patients with preoperative and postoperative imaging. NOS=Not otherwise specified.

5.3.3 Survival Based on Postoperative Image Features

After removing features that were strongly correlated with postoperative tumor volume or postoperative enhancing volume (Spearman $r > 0.8$), several image features were significantly associated with survival. Table 5.3 shows the features from each image type with the highest hazard ratio between binarized groups. These methods are described fully in Section 4.2.3. On the postoperative imaging several features from local pathology estimates were significantly associated with survival although the results were not as strong as the raw image features. The results reported in this section are over the combined cohort of all WHO grades (II III IV). The corresponding results for individual grade subsets are listed in the appendix, Section A.4.3.

Panel A: Univariate analysis										
image	region	feature	IS		Univariate					
			C	Cut	HR	95% CI	p			
T1	enhanc	10Percentile	0.590	0.343	2.120	[1.63, 2.76]	3e-07	***		
FL	edema	10Percentile	0.578	1.28	1.781	[1.48, 2.14]	2e-08	***		
-	wholetumor	VoxelVolume	0.613	66.3	1.753	[1.45, 2.12]	1e-07			
TC	wholetumor	IntensityAtVolume1000	0.584	1.43	1.657	[1.36, 2.02]	5e-06	***		
T2	wholetumor	IntensityAtVolume1000	0.554	0.545	1.614	[1.25, 2.09]	1e-03	**		
ERG	enhanc	Maximum	0.522	4	1.523	[1.20, 1.93]	2e-03	**		
CD	wholetumor	IntensityAtVolume1000	0.569	6.34e+03	1.506	[1.19, 1.91]	2e-03	**		
Ki67	nonenh	TotalSum	0.557	1.74e+04	1.403	[1.17, 1.68]	1e-03	**		
GR	higher	Maximum3DDiameter	0.575	61.5	1.290	[1.05, 1.58]	3e-02	*		
Panel B: Multivariate analysis										
image	region	feature	Multivariate (age+KPS)				Multivariate (age+KPS+grade)			
			HR	95% CI	p		HR	95% CI	p	
T1	enhanc	10Percentile	1.838	[1.41, 2.40]	7e-06	***	1.728	[1.33, 2.25]	5e-05	***
FL	edema	10Percentile	1.482	[1.23, 1.78]	3e-05	***	1.473	[1.23, 1.77]	4e-05	***
-	wholetumor	VoxelVolume	1.439	[1.19, 1.75]	2e-04	***	1.289	[1.06, 1.56]	1e-02	*
TC	wholetumor	IntensityAtVolume1000	1.359	[1.11, 1.66]	3e-03	**	1.240	[1.01, 1.52]	4e-02	*
T2	wholetumor	IntensityAtVolume1000	1.377	[1.06, 1.78]	2e-02	*	1.352	[1.04, 1.75]	2e-02	*
ERG	enhanc	Maximum	1.170	[0.92, 1.49]	2e-01		1.112	[0.87, 1.42]	4e-01	
CD	wholetumor	IntensityAtVolume1000	1.255	[0.99, 1.59]	6e-02	.	1.136	[0.89, 1.44]	3e-01	
Ki67	nonenh	TotalSum	1.130	[0.94, 1.36]	2e-01		1.108	[0.92, 1.33]	3e-01	
GR	higher	Maximum3DDiameter	1.084	[0.88, 1.33]	4e-01		1.031	[0.84, 1.27]	8e-01	

TABLE 5.3: Best postoperative image features from each image type among patients with all WHO grades (II III IV) in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate (Panel A) and multivariate (Panel B) models. p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample.

T1: Overall T1-Weighted Pre-Contrast Brightness A 10th percentile intensity >0.343 over the enhancing sub-region on a T1-weighted pre contrast image was associated with a worse prognosis and univariate hazard ratio of 2.120. Having a higher 10th percentile means that the region is overall brighter and almost entirely above the 0.343 cutoff. This feature had a reasonable concordance with survival at 0.590 and had a very stable optimal cutoff value (Figure 5.3) with the in-sample and mean cross-validated cutoffs differing by just 0.001.

When preoperative age and KPS are taken into account alongside bright T1, the hazard ratio was reduce somewhat to 1.838. This is slightly smaller than the HR for age and KPS (> 2) and much smaller than the preoperative HR for WHO Grade (>3) but still remained significant. Furthermore, when high WHO grade (III/IV) was also used as a covariate the hazard ratio only decreased slightly to 1.728. This is larger than the hazard ratio of 1.55 for tumor volume EOR (Table 5.6). So, not only is there is quite a bit of independent prognostic information present in the postoperative imaging relative

to tumor grade, the postoperative T1 brightness may be more prognostic than just reduction in tumor volume. Interestingly, the relation between the threshold >0.343 and worse prognosis suggests that brighter T1 pre-contrast values correspond to worse prognosis which is opposite clinical intuition. It is possible that the feature is detecting bright treatment effect or blood products that happen to correlate with prognosis as well. Although, the optimal cutoff (Figure 5.3) is about 0.35 on the scale from CSF to brain mode so it is still in the hypo-intense range.

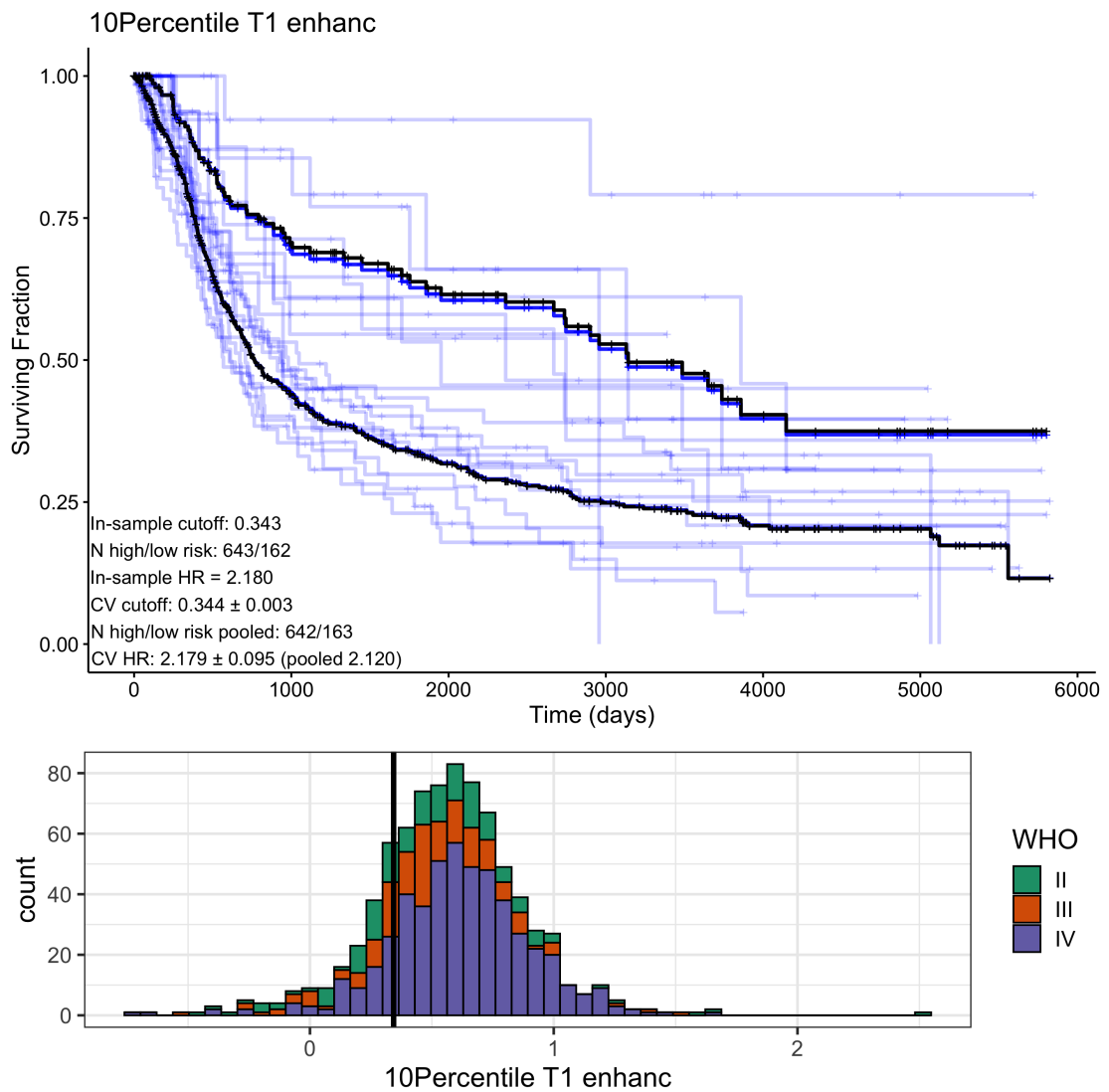


FIGURE 5.3: Best postoperative feature for stratifying survival for WHO grades II, III, and IV cases for T1 image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

FLAIR: Overall Edema Brightness The 10th percentile FLAIR intensity over the edema region being greater than 1.28 was associated with worse prognosis with univariate hazard ratio 1.781, Figure 5.4. Like the T1w feature, this feature effectively captures the overall brightness and the threshold indicates that postoperative peritumoral edema more than about 28% brighter than the rest of the brain confers a poor prognosis. This feature had a moderate C-index of 0.578 which is comparable to the rest of the rest of the postoperative features. Figure 5.4 shows possible a small amount of overfitting by in-sample results since the black in-sample curves are clearly outside the blue cross-validated curves. However, there is still good agreement in the cutoff values and hazard ratios which are on the order of 1% different.

Like other features, the hazard ratio was reduced some to 1.482 in multivariate analysis taking into account age and preoperative KPS. This means the effect is smaller than age and KPS but still remains significant. The multivariate hazard ratio is also quite similar to the multivariate hazard ratio for reduction in tumor volume (1.55, Table 5.6). This suggests that the brightness of edema on FLAIR imaging may be similarly prognostic.

This result makes sense because brighter FLAIR intensity is suggestive of residual tumor. The same patients with extensive resections in terms of volume are also likely to have less FLAIR hyperintensity. Also note, the FLAIR image performed better than the base T2w image possibly due to better suppressing some treatment effects that could confound the feature measurements.

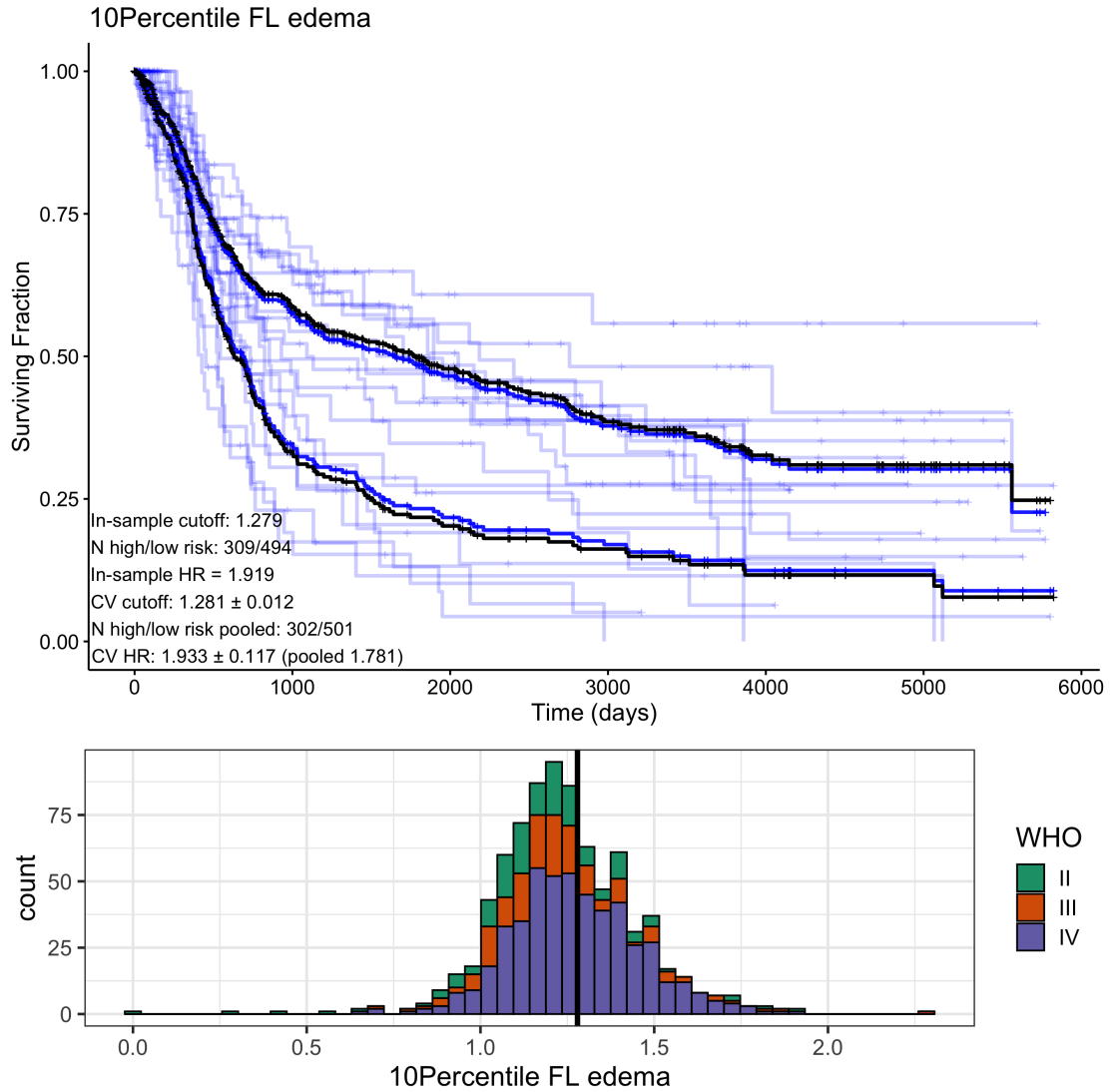


FIGURE 5.4: Best postoperative feature for stratifying survival for WHO grades II, III, and IV cases for FLAIR image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

Shape: Total Residual Volume The best postoperative shape feature was the volume of the residual visible tumor edema. A volume $>66.3 \text{ cm}^3$ was associated with a univariate hazard ratio of 1.753, Figure 5.5. Among all the postoperative features, this one had the highest C-index of 0.613. However, we did observe some overfitting, shown in Figure 5.5 by the separation of black and blue survival curves. This just means the in-sample hazard ratio of 2.03 may be overly optimistic, and the true optimal cutoff may be closer to the average cross-validated value of 60.3 cm^3 .

However, the postoperative residual volume remained significant in relation to age and

KPS with a slightly reduced hazard ratio of 1.44 (smaller than either age or KPS themselves). The hazard ratio was only reduced slightly with further comparison to tumor grade, having a final significant multivariate hazard ratio of 1.29. This is smaller than the comparable hazard ratio based on reduction in tumor volume (1.55) which suggests that the information contained in the postoperative volume alone is not enough to encompass the full prognosis. Nonetheless, it did provide significant and independent information to age, KPS, and grade. Overall this feature makes sense as a prognostic indicator based on the extensive literature studying extent of resection in terms of radiographically visible tumor [14, 15].

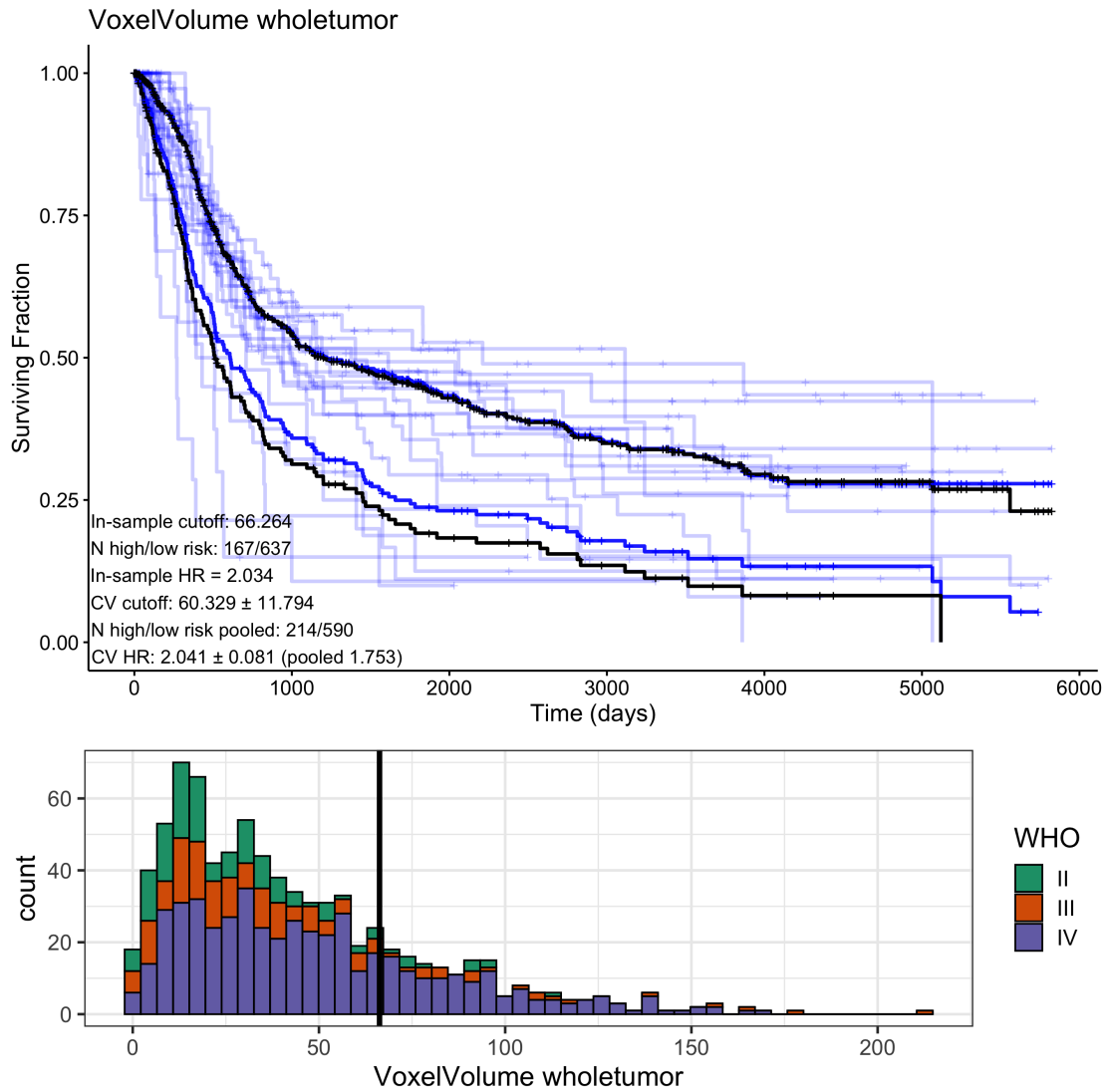


FIGURE 5.5: Best postoperative shape feature for stratifying survival for WHO grades II, III, and IV. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

TC: Maximum Contrast Enhanced Intensity Similarly to the preoperative results, we found the most prognostic contrast enhanced postoperative image feature to be the standardized (1 cc) maximum intensity over a whole tumor VOI. A value greater than 1.4 was associated with a worse prognosis and hazard ratio of 1.657, Figure 5.6. The concordance (C-index) between max postoperative TC and survival was still good at 0.584 and the optimal cutoff showed good stability in cross-validation analysis with a standard deviation of just 0.006.

The hazard ratio remained significant in multivariate analysis against age and KPS and

reduced slightly to 1.36 which was similar to other postoperative features. When tumor grade was included as a covariate the maximum postoperative enhancing intensity still remained significant with a hazard ratio of just 1.24. This is much smaller than the comparable hazard ratios for any of the clinical factors including reduction in tumor volume.

It is worth noting that this feature is only defined for cases with at least 1 cc of residual enhancing volume as a consequence of the feature definition (and to de-noise the more complex postoperative imaging, where small areas of T1 hyperintensity and enhancement are frequently seen, and difficult to interpret). This means patients without any residual enhancement were excluded from the calculation of hazard ratios and C-index, so the interpretation of the 1 cc T1C should be taken with caution. However, the result is consistent with the preoperative results in Section 4.3.4 where the best survival stratification between thresholded groups was achieved using the 0.1 cc T1C intensity. On the preoperative data the hazard ratio was much larger (5.1, Table 4.9) which may be due to the better preoperative image quality or the presence of treatment effect on postoperative imaging. Further investigation to exclude treatment effect and quantify true enhancing behavior should identify similar results to the preoperative case.

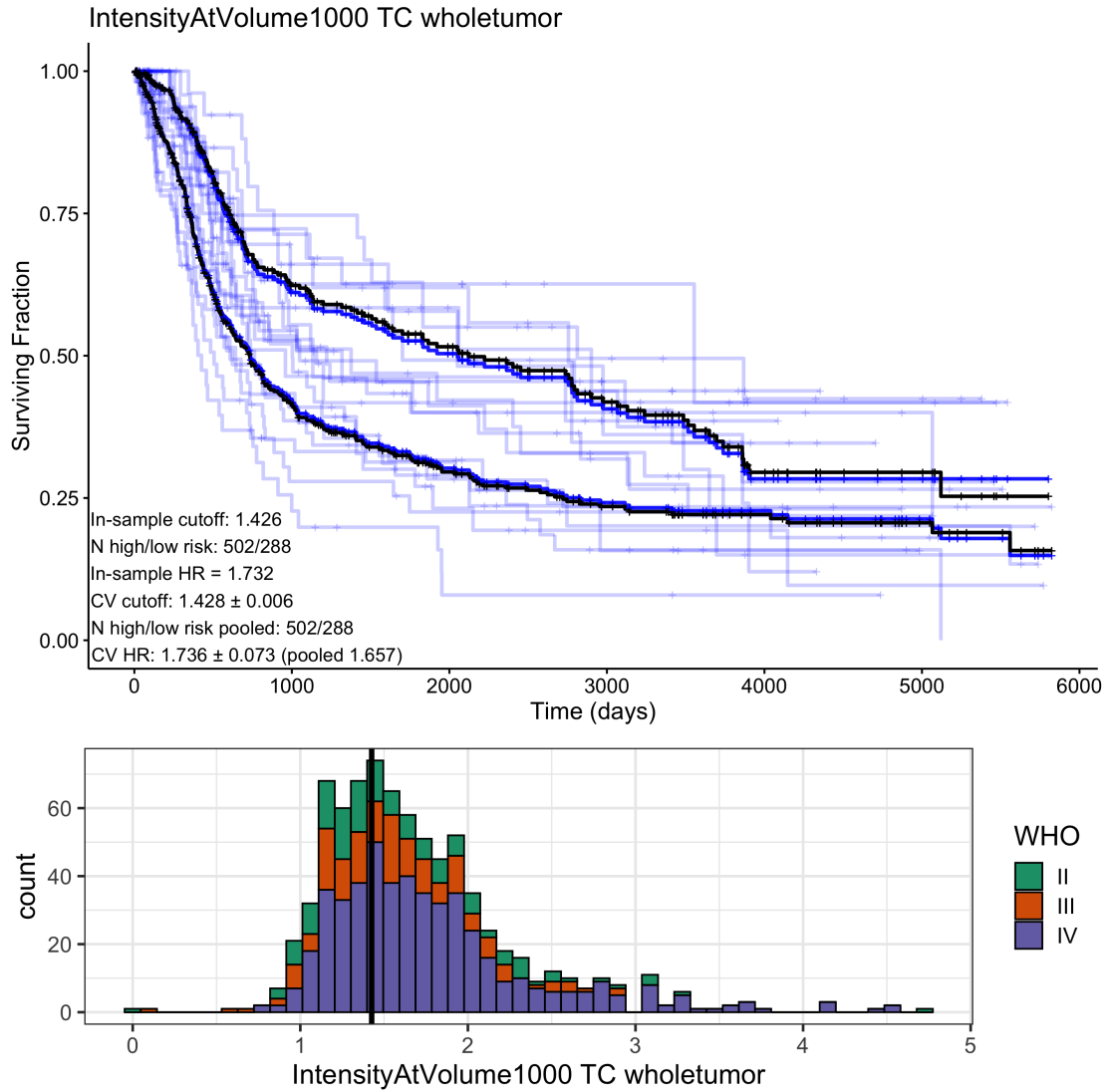


FIGURE 5.6: Best postoperative feature for stratifying survival for WHO grades II, III, and IV cases for TC image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

T2: Maximum T2-Weighted intensity The best postoperative T2-weighted image feature was the same feature for the postoperative contrast enhanced image: the maximum intensity in the whole tumor VOI. Patients with a maximum intensity >0.545 had a worse prognosis with hazard ratio 1.614, Figure 5.7, which was the smallest postoperative hazard ratio among the raw image types. This threshold corresponds to a brightness about half-way between normal brain and CSF. The 1 cc T2w intensity over the residual tumor had a moderate C-index of 0.554 but we observed excellent stability of the optimal threshold value, recovering 0.545 both in-sample and in cross-validation.

When compared against age and KPS the relationship between maximum postoperative T2w intensity and survival remained significant with a hazard ratio of 1.377 and was nearly unchanged with the inclusion of grade (HR 1.35). This may be due to the fact that gliomas of all grades are T2 hyperintense so there is little redundant information provided by grade. This supports the use of max T2 intensity as an independent prognostic factor with a slightly smaller hazard ratio than whole-tumor extent of resection. However, the same caveat as the TC results still holds, the 1 cc volume constraint means the feature is only defined for patients with 1 cc or more of residual tumor. But, this is almost always the case.

Large T2-weighted brightness makes clear sense as a prognostic factor since T2 brightness is a standard radiographic finding. Brighter T2 intensity likely signals malignant or active tumor compared to more moderate signal from edema or treatment related effects.

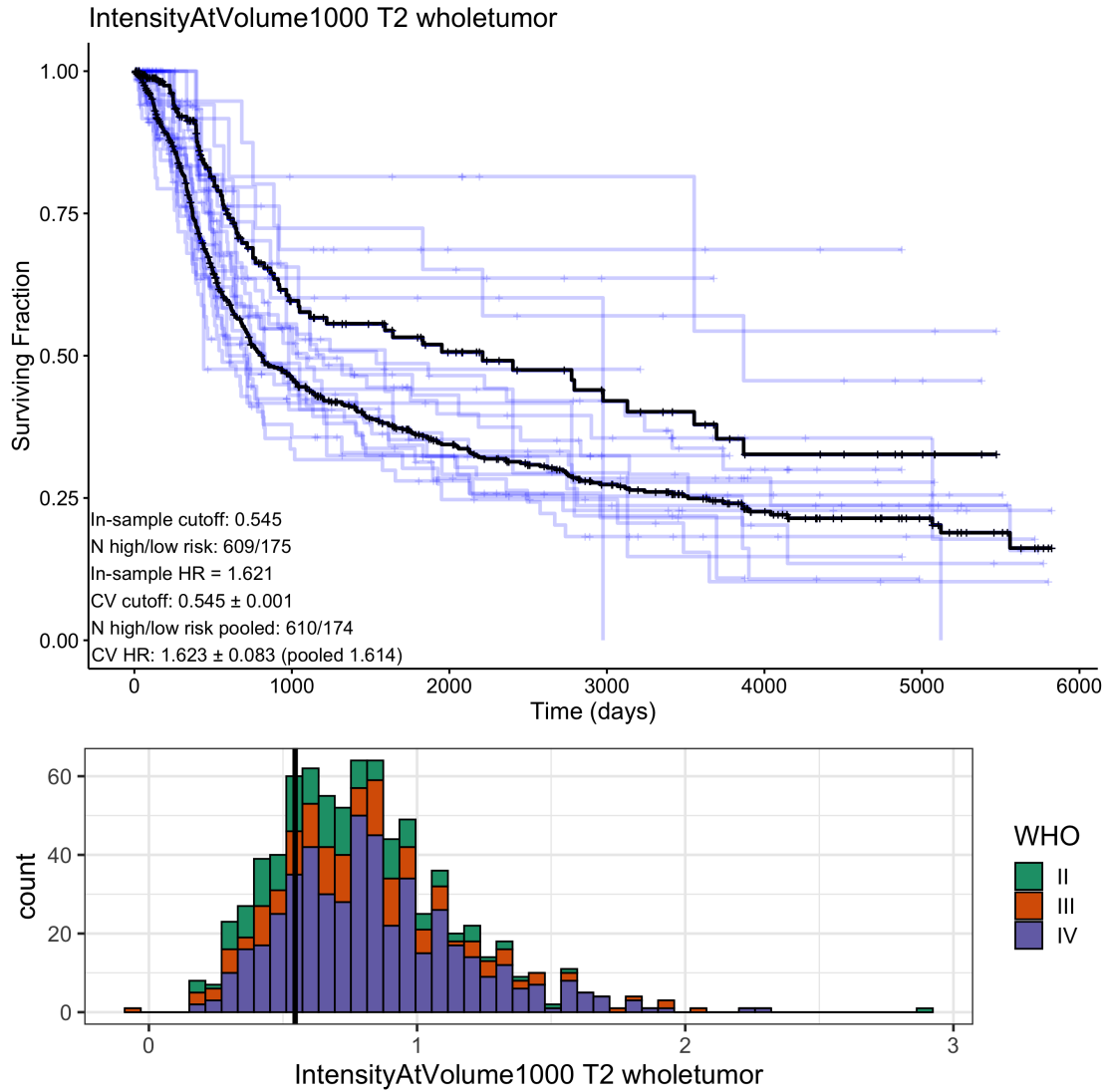


FIGURE 5.7: Best postoperative feature for stratifying survival for WHO grades II, III, and IV cases for T2 image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

ERG: Maximum ERG Expression Patients with maximum postoperative ERG $>4.00\%$ in the residual enhancing sub-region had worse prognosis with a univariate hazard ratio of 1.523, Figure 5.8, which was the highest among the synthetic pathology features. Despite good generalization of the optimal threshold and univariate hazard ratio, the C-index with survival was only moderate at 0.554.

Unlike the raw image features, the maximum ERG was not significant compared against age and KPS (or grade). This may be due to the already smaller hazard ratio in univariate analysis that is redundant with other clinical factors. We expected maximum ERG

to be prognostic because ERG is a vascularity marker and angiogenic tumor with compromised blood-brain-barrier tends to be contrast enhancing. However, we know that enhancing subregion measurements on the postoperative data are potentially unreliable due to treatment effect. This may have weakened the ability to detect pockets of vascular tissue. Although, this postoperative result is still consistent with the preoperative results where the best performing ERG feature (maximum within enhancing subregion) was associated with a univariate hazard ratio of 2.94.

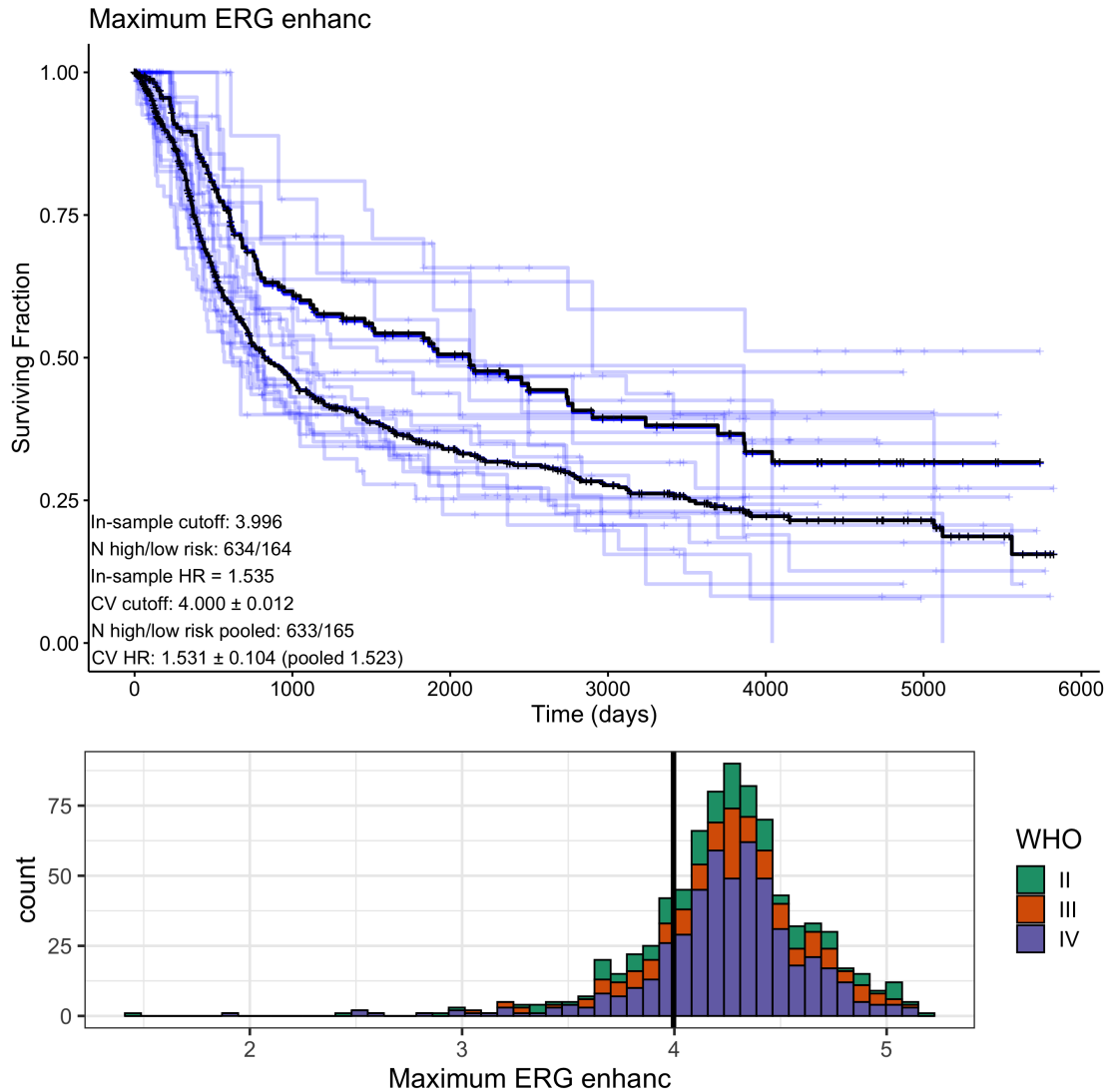


FIGURE 5.8: Best postoperative feature for stratifying survival for WHO grades II, III, and IV cases for ERG map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

Cell Density: Maximum Cell Density The best postoperative feature for cellularity was the maximum (with 1 cm³ constraint) over the whole tumor VOI at a threshold of 6343 nuclei/mm². This was a smaller threshold than the preoperative data of 7680 nuclei/mm² which makes sense because the overall cellularity of these tumor has been reduced by surgery. A maximum cell density >6343 nuclei/mm² was associated with a worse prognosis and univariate hazard ratio of 1.506 and C-index of 0.569. The threshold and hazard ratio was stable between in-sample and cross-validation as shown in Figure 5.9.

Like the other simulated pathology estimates, the maximum cell density was not independently prognostic of age and KPS (or grade). This is surprising since the preoperative maximum cell density had some of the best performance among all features. We can attribute the lack of multivariate significance to either confounding of the cell density estimates by postoperative treatment effect or to redundancy with the clinical factors.

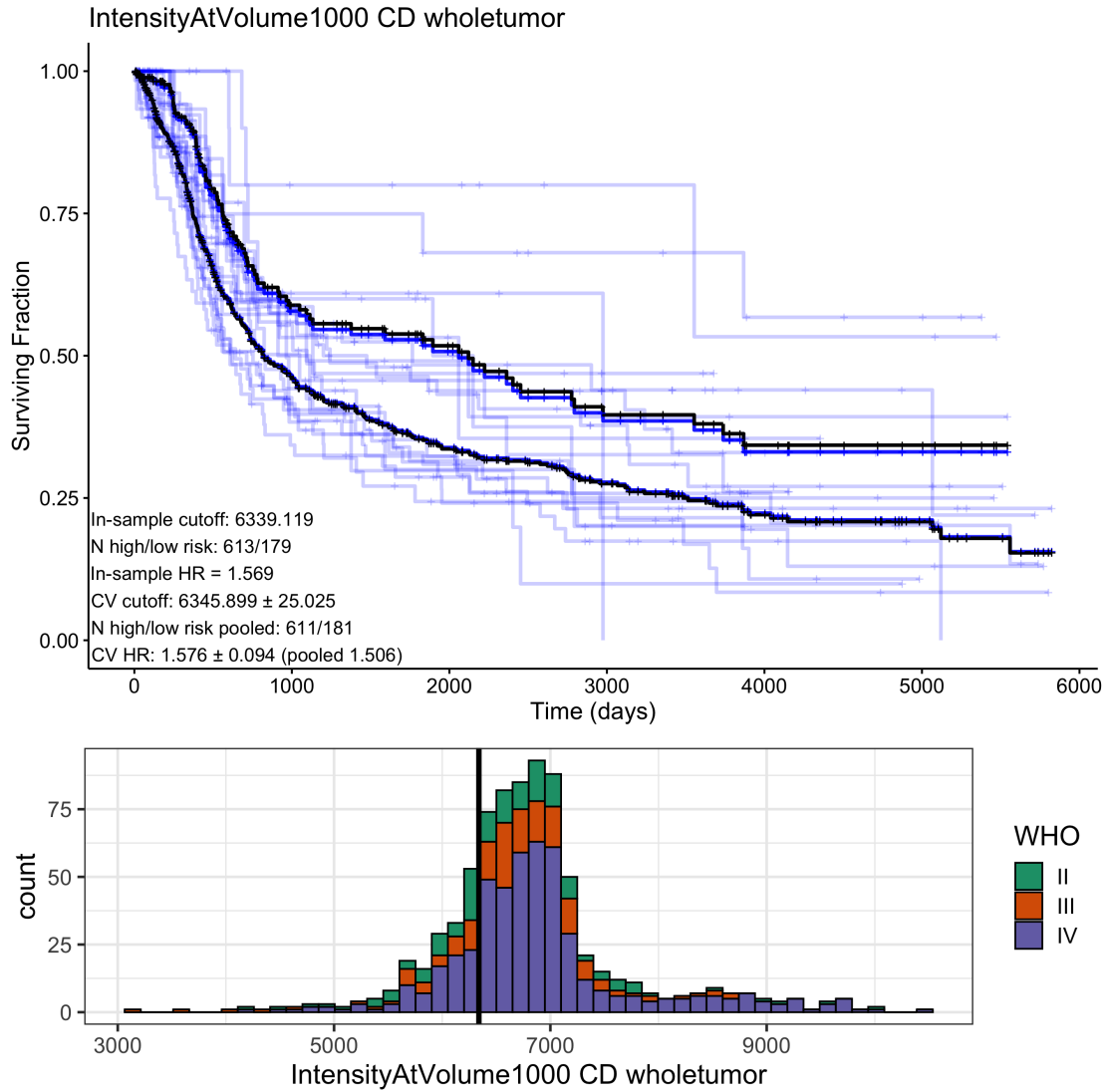


FIGURE 5.9: Best postoperative feature for stratifying survival for WHO grades II, III, and IV cases for CD map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

Ki67: Proliferation-Weighted Non-Enhancing Volume For estimated proliferation (Ki67), the best feature by hazard ratio was the total sum of Ki67 over residual non-enhancing tumor core, which can be interpreted as weighting the voxels in the residual non-enhancing tumor by their Ki67 expression. A value > 1740 was associated with a worse survival and univariate hazard ratio of 1.403, Figure 5.10. The concordance with survival was modest at 0.557 however the stability in cross-validation was very good as shown by overlap of black and blue curves in Figure 5.10.

Again, we observed that the sum of Ki67 was not prognostic when compared to clinical

factors age, KPS, and grade. We expected to find that residual proliferating tumor (or its summation) would be a crucial prognostic factor. After all, dividing tumor cells are what lead to recurrence after treatment. However, we found only univariate significant results based on the postoperative measurements alone. Like cell density measurements, one explanation may be confounding by postoperative treatment effect.

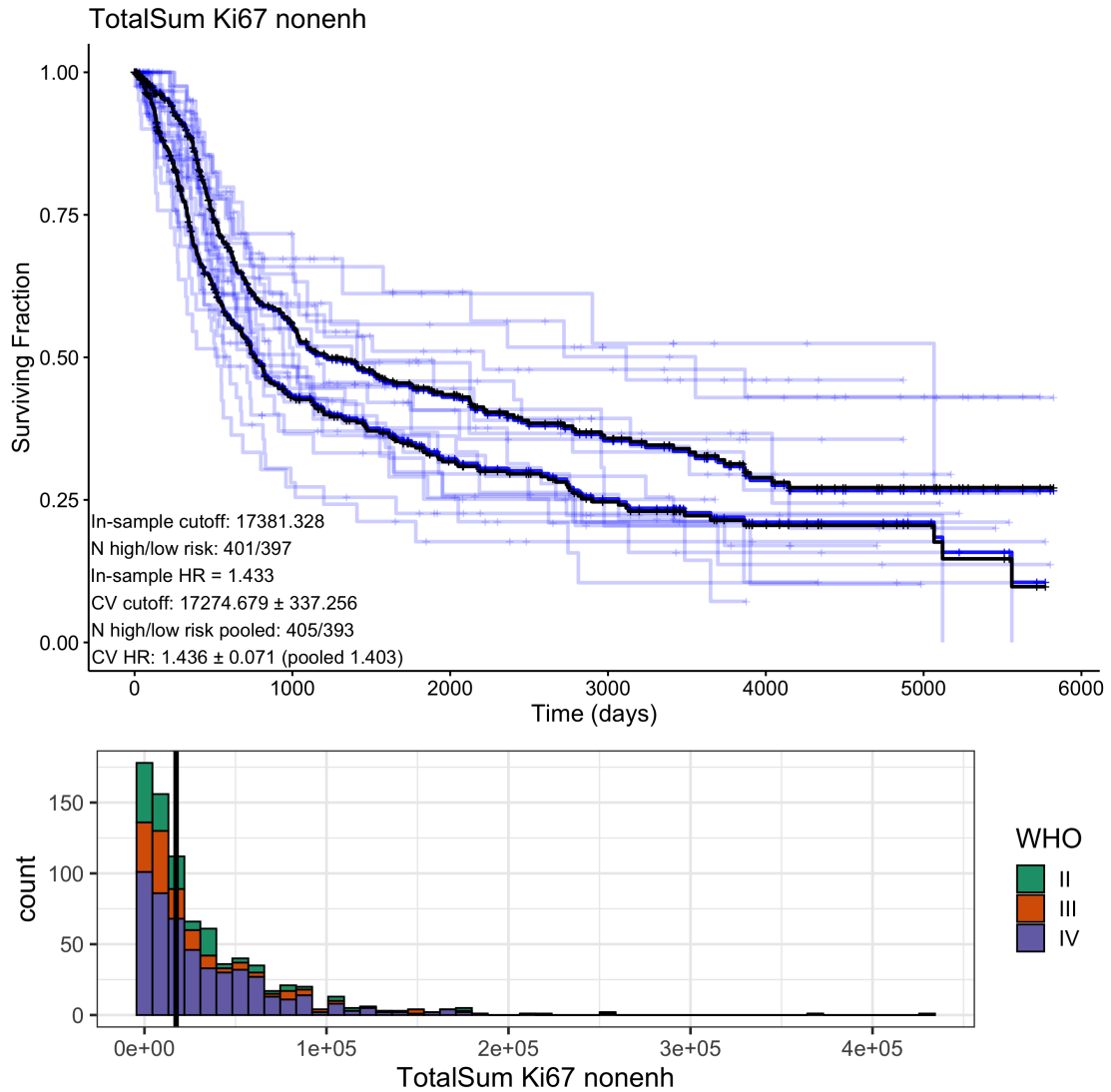


FIGURE 5.10: Best postoperative feature for stratifying survival for WHO grades II, III, and IV cases for Ki67 map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

Local Grade: Diameter of Higher-Grade Region The largest hazard ratio among local grade features was the maximum diameter (i.e. size) of estimated higher

grade disease. However, the relation was non-significant even in univariate analysis. One reason may be due to instability in the 3D diameter feature which does not enforce the distance be between contiguous points. For example, a complete resection with just two small pieces of residual high grade can have a very large diameter if they are separated spatially. This is less likely to happen in the preoperative case where the tumor is one contiguous unit. Indeed we found the diameter of high grade disease was strongly univariate significant (HR 1.42) on the preoperative data. For completeness, the cross-validated survival curves are shown in Figure 5.11.

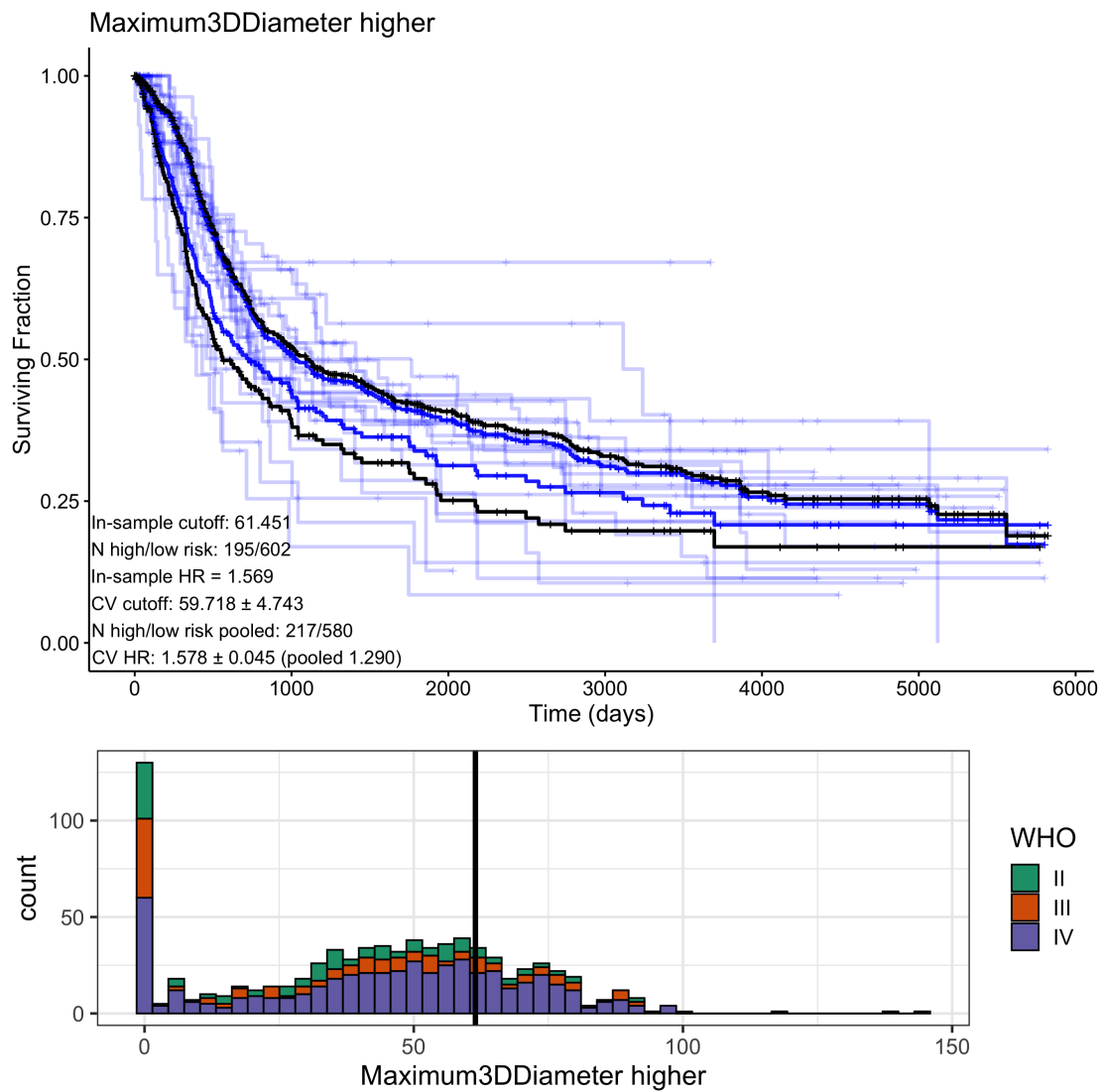


FIGURE 5.11: Best postoperative feature for stratifying survival for WHO grades II, III, and IV cases for grade map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Note: the survival difference from this feature was non-significant after multiple comparison correction. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

Subset Analysis Like the preoperative results, we found that estimated pathology performed markedly better on the subset of WHO II cases (appendix Table A.11). However, the median CD values over the edema sub-region had so few events in the low-risk category that the hazard ratio could not be meaningfully estimated. This shows the potential to use Median CD as a powerful biomarker for detecting long term survivors after surgery in the low-grade subset based on the survival curves in Figure A.14 even though statistical significance was not reached. Similarly, the maximum ERG in the edema subregion had an impressive point-estimate of hazard ratio of 7.2 but did not achieve statistical significance. Survival curves are shown in the appendix Figure A.14.

5.3.4 Clinical Extent of Resection

Several studies have examined extent of tumor resection for glioma patients using conventional radiographic findings. In particular, fractional reduction of contrast enhancement and T2-FLAIR abnormality. We replicated this analysis using the segmented tumor volumes and measured the volume EOR (Equation 5.2). For comparison, we also analyzed the reference EOR measurements provided alongside clinical data. These results are similar given the agreement between segmented and reference tumor volumes in Figure 3.11.

Lacroix et al [14] found that extent of resection of T1 enhancing volume was still a significant prognostic factor in a multivariate cox model after controlling for advanced age (greater than 55 years) and low preoperative performance status (KPS). Specifically, for grade IV tumors they found a hazard ratio of 1.4 associated with removal of 98% or more of enhancing volume. We duplicated this analysis, subset on our Grade IV cases, and computed the EOR measurements done by the software algorithm (Table 5.4). We found the optimal cut-off threshold using the segmented result to be 93% with a univariate hazard ratio of 1.66. Figure 5.12 shows survival curves for cases above and below the 93% EOR threshold. In multivariate analysis with preoperative age and KPS included, the hazard ratio was decreased slightly to 1.42, Table 5.4. In summary, we were able to duplicate the results from Lacroix et al [14], using a machine based segmentation approach, giving us confidence in the algorithm. The hazard ratios of 1.7 (multivariate 1.4) serve as benchmark values to compare the effectiveness of postoperative and EOR image features on the high grade subset. Features that achieve comparable or greater

hazard ratios potentially have a stronger prognostic effect than the traditional extent of resection measure based on contrast enhancement.

We also looked at the effect of contrast enhanced EOR for the combined subset. The optimal threshold was also 93% corresponding to a hazard ratio of 0.812 for gross-total resection. The corresponding multivariate analysis against age, KPS, and grade is given in Table 5.5 and surprisingly shows EOR based on T1C as non-significant in the combined cohort. We also evaluated reduction in overall tumor volume based on T2-FLAIR size. A reduction of 67% or more in total tumor volume was significantly associated with survival in univariate analysis ($HR = 2.0$) and multivariate analysis ($HR = 1.55$) against age, KPS, and grade, Table 5.6. This value of 1.55 is a benchmark to compare the prognostic power of postoperative features.

	HR	CI	p	
KPS < 70	1.526	[1.231, 1.892]	1.18e-04	***
Age > 55	1.831	[1.445, 2.321]	5.55e-07	***
T1C EOR < 93%	1.420	[1.073, 1.880]	1.42e-02	*

TABLE 5.4: Multivariate Cox proportional hazards model for **WHO IV** cases using CLARA segmented enhancing volume EOR measurements and clinical factors that strongly influence survival. Confidence intervals are 95%

	HR	CI	p	
KPS < 70	1.879	[1.514, 2.332]	1.01e-08	***
Age > 55	2.678	[2.139, 3.352]	8.05e-18	***
WHO III/IV	3.511	[2.295, 5.371]	7.10e-09	***
T1C EOR < 93%	1.232	[0.963, 1.576]	9.75e-02	.

TABLE 5.5: Multivariate Cox proportional hazards model for all WHO grades (II III IV) cases using CLARA segmented enhancing volume EOR measurements and clinical factors that strongly influence survival. Confidence intervals are 95%

	HR	CI	p	
KPS < 70	1.861	[1.500, 2.309]	1.67e-08	***
Age > 55	2.582	[2.062, 3.233]	1.38e-16	***
WHO III/IV	3.521	[2.302, 5.385]	6.41e-09	***
T2 EOR < 67%	1.551	[1.154, 2.084]	3.63e-03	**

TABLE 5.6: Multivariate Cox proportional hazards model for all WHO grades (II III IV) cases using CLARA segmented total tumor volume EOR measurements and clinical factors that strongly influence survival. Confidence intervals are 95%

This study of extent of resection by Lacroix et al. [14] was further extended by Li et al. [15] by evaluating extent of resection of FLAIR hyperintensity in addition to enhancing volume. The main result was that resection 53.2% of the T2-FLAIR volume

in addition to 100% of enhancing volume conferred a survival benefit. We attempted to reproduce this result, but with some adjustments. Very few high grade cases have identically zero enhancing volume segmented, so the total resection requirement of 100% enhancement removed was relaxed. For the T2-FLAIR analysis, we used the subset of patients with enhancing volume resection $>93\%$: the GTR threshold in Table 5.4. The effect of further FLAIR resection is shown in Figure 5.13. We found an optimal threshold of 75% resection of T2-FLAIR volume in addition to total resection of enhancing volume led to a significant survival difference.

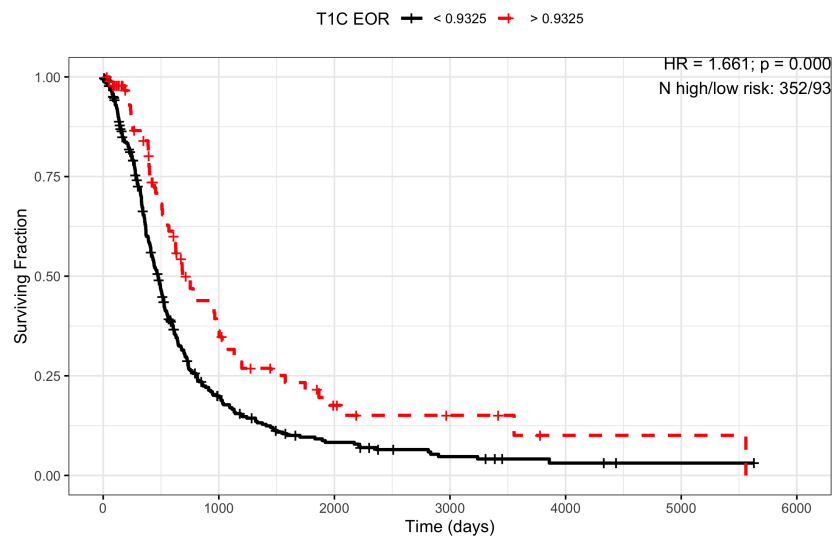


FIGURE 5.12: Survival curves for WHO IV cases based on extent of resection for T1 enhancing volume. Left: using automatically segmented tumor volumes. Right: using reference values. The cutoffs are selected to optimize the hazard ratio between groups. (p-value from log-rank test)

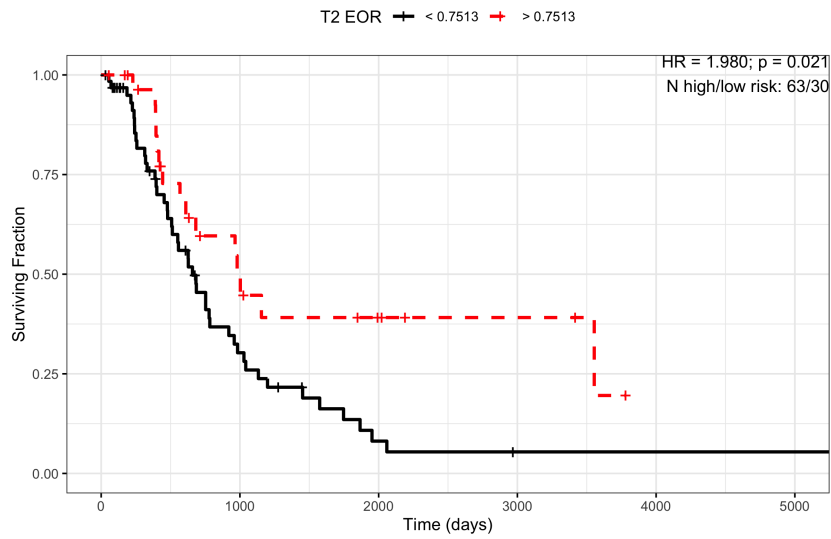


FIGURE 5.13: Survival curves based on T2 FLAIR resection for WHO IV cases with complete resection of enhancing volume. Left: using automatic volume measurements. Right: using reference values. Interestingly, the literature value is not reproduced. Note that the patient populations in each plot are not necessarily the same.

5.3.5 Extent of Resection Based on Image Features

The survival analysis proceeded in the same way as in Chapter 4 with the EOR features in place of the preoperative image features. For each feature we computed the C-index and optimal cross-validated threshold. Then, the prognostic information was compared with clinical factors using multivariate cox modeling while adjusting for covariates of age (>55 years), KPS (<80), and tumor grade. Table 5.7 lists the best feature for each base image and synthetic pathology map. For this analysis, we excluded TotalSum and TotalEnergy features since they all show very strong correlation with reduction in tumor volume and we are focused on potential features that do not simply recapitulate this known result. Overall, we found similar results to the preoperative survival. EOR features based on estimated pathology, especially proliferation and cell density, showed significant cross-validated survival stratification across all glioma grades. The results reported in this section are over the combined cohort of all WHO grades (II III IV). The corresponding results for individual grade subsets are listed in the appendix, Section A.4.4.

Panel A: Univariate analysis										
image	region	feature	IS		Univariate					
			C	Cut	HR	95% CI	p			
CD	wholetumor	EORfrac-10Percentile	0.535	0.205	1.894	[1.46, 2.46]	4e-05	***		
T2	wholetumor	EORreduc-Median	0.577	0.233	1.861	[1.44, 2.41]	5e-05	***		
-	nonenh	EORfrac-VoxelVolume	0.417	0.931	1.810	[1.41, 2.32]	5e-05	***		
ERG	wholetumor	EORreduc-IntensityAtVolume100	0.515	0.479	1.751	[1.38, 2.23]	9e-05	***		
Ki67	wholetumor	EORfrac-10Percentile	0.546	0.446	1.616	[1.29, 2.03]	4e-04	***		
FL	wholetumor	EORfrac-Median	0.479	0.21	1.549	[1.22, 1.96]	2e-03	**		
T1	enhanc	EORreduc-10Percentile	0.529	0.286	1.446	[1.15, 1.82]	7e-03	**		
Panel B: Multivariate analysis										
image	region	feature	Multivariate (age+KPS)				Multivariate (age+KPS+grade)			
			HR	95% CI	p		HR	95% CI	p	
CD	wholetumor	EORfrac-10Percentile	1.597	[1.23, 2.08]	5e-04	***	1.640	[1.26, 2.13]	2e-04	***
T2	wholetumor	EORreduc-Median	1.539	[1.19, 2.00]	1e-03	**	1.568	[1.21, 2.03]	7e-04	***
-	nonenh	EORfrac-VoxelVolume	1.278	[0.99, 1.65]	6e-02	.	1.218	[0.94, 1.57]	1e-01	
ERG	wholetumor	EORreduc-IntensityAtVolume100	1.426	[1.12, 1.82]	4e-03	**	1.452	[1.14, 1.85]	3e-03	**
Ki67	wholetumor	EORfrac-10Percentile	1.455	[1.16, 1.83]	1e-03	**	1.574	[1.25, 1.98]	1e-04	***
FL	wholetumor	EORfrac-Median	1.388	[1.10, 1.76]	6e-03	**	1.442	[1.14, 1.83]	2e-03	**
T1	enhanc	EORreduc-10Percentile	1.349	[1.07, 1.70]	1e-02	*	1.317	[1.05, 1.66]	2e-02	*

TABLE 5.7: Best extent of resection features from each image type among patients with all WHO grades (II III IV) in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate (Panel A) and multivariate (Panel B) models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. For local grade (GR) and T1 post-contrast (TC) no remaining features were univariate significant so they are omitted from the table. IS = in sample.

Cell Density: Overall Reduction in Cellularity For preoperative image features, we found that cell density (CD) was highly prognostic for the combined cohort and especially for low grade (WHO II) tumors. With respect to EOR, a 21% reduction in 10th percentile CD in the whole tumor ROI was also strongly prognostic with a hazard ratio of 1.894, Figure 5.14. The 21% threshold, which was on the higher end of the range of values observed in the cohort, was quite stable in cross-validated analysis and we found the same average cutoff among the 10 folds and in-sample. The correlation between reduction in 10th percentile CD and overall survival was smaller than many other EOR features though at just 0.535.

The hazard ratio was reduced in multivariate analysis alongside age and KPS to 1.597 but remained significant. Interestingly, the hazard ratio then increased slightly to 1.640 when WHO grade was included as a covariate. This means that the reduction in overall CD is an independent prognostic measure of prognosis relative to clinical factors and grade. We can interpret this feature (reduction in 10th percentile CD) as a reduction in the overall cellularity of the tumor. The 10th percentile CD is reduced when voxels

with values greater than the 10th percentile are preferentially removed, i.e. by targeting highly cellular regions for resection.

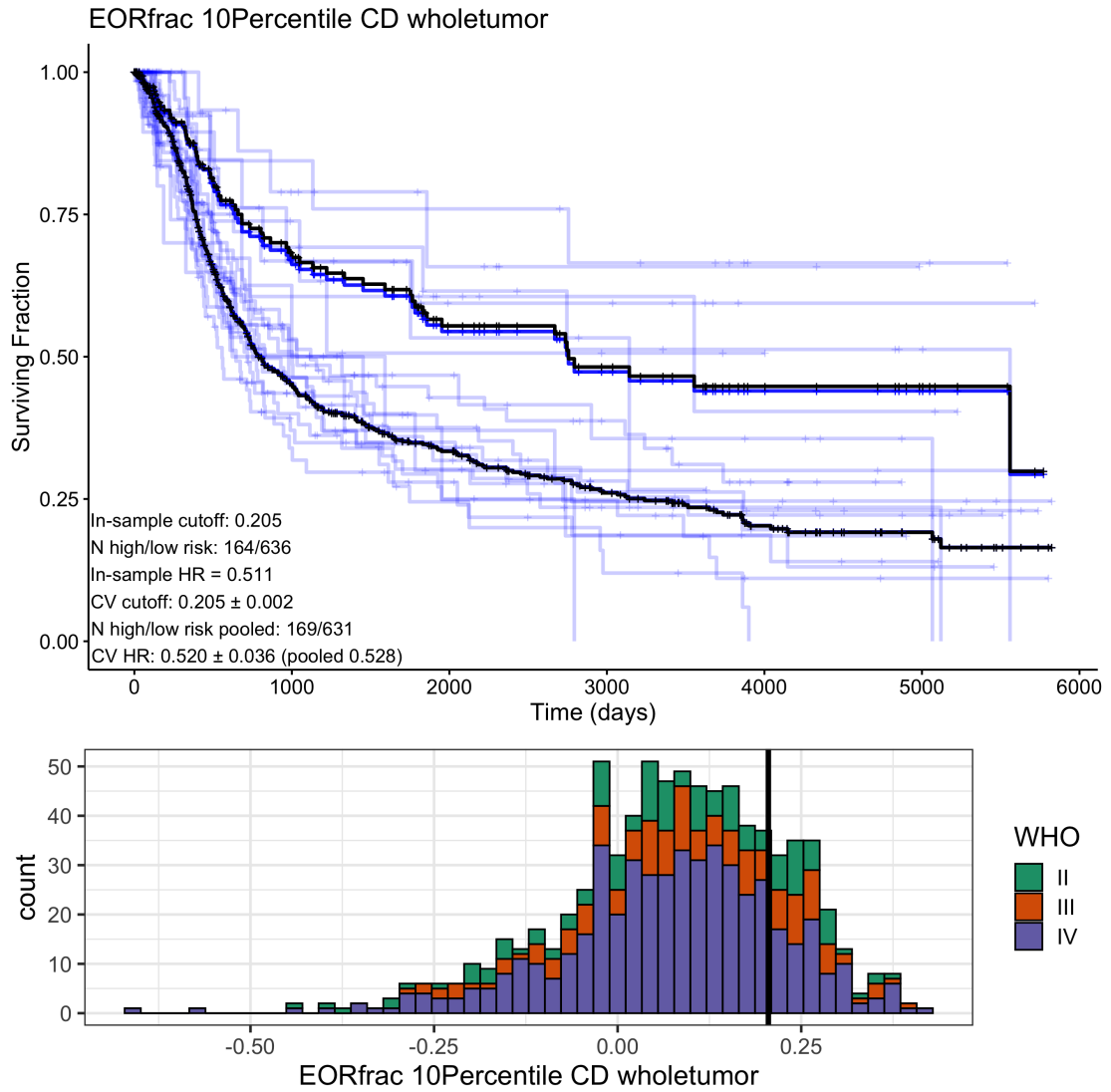


FIGURE 5.14: Best extent of resection feature for stratifying survival for WHO grades II, III, and IV cases for CD map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

T2: Reduction in Median T2-Weighted Brightness A reduction of 0.233 (on the brain - CSF scale) in median T2-weighted image intensity was associated with a hazard ratio of 1.861, Figure 5.15. Although the hazard ratio was slightly smaller than the best overall EOR feature (based on CD), the C-index was slightly higher at 0.577. We found the optimal threshold of 0.233 was very stable in cross-validation and the

average among the 10 folds was less than 3% different from the in-sample threshold. The survival curves in Figure 5.15 also showed good agreement.

The hazard ratio associated with reduction in median T2w intensity was reduced slightly with the inclusion of age and KPS in multivariate analysis to 1.539, but interestingly recovered slightly to 1.568 when WHO grade was also included. For all models, reduction in T2w intensity was a significant independent predictor of overall survival. The independent hazard ratio of 1.568 is comparable to the hazard ratio for reduction in tumor volume of 1.6. Reduction in median T2w intensity has a straightforward interpretation. Reducing the median intensity is accomplished by preferentially reducing the bright T2w tumor regions. This is consistent with the current practice of targeting T2-bright tumor for resection. Indeed the histogram in Figure 5.15 shows that a majority of cases have a reduction in median T2w intensity between 0 and 0.5.

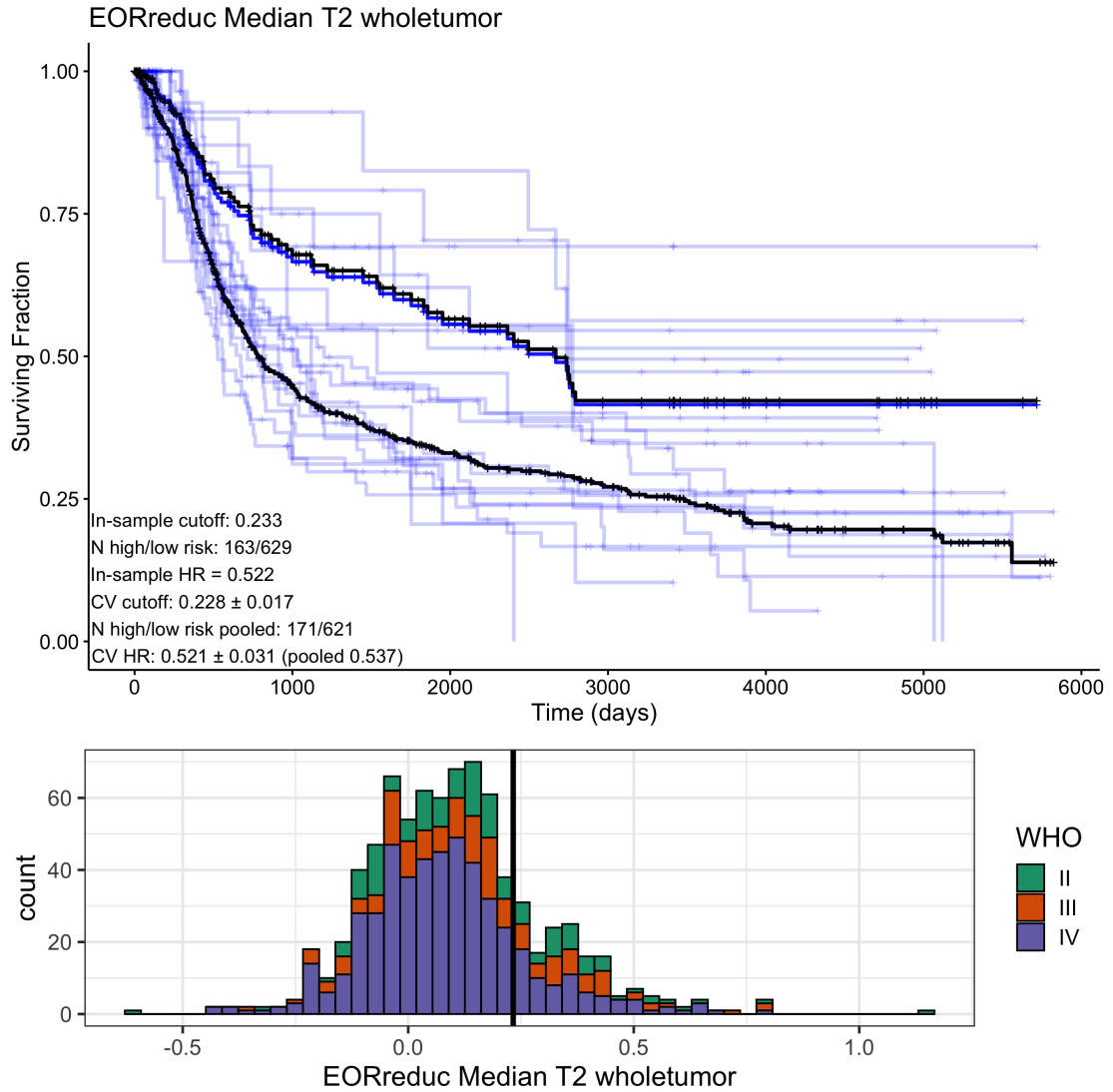


FIGURE 5.15: Best extent of resection feature for stratifying survival for WHO grades II, III, and IV cases for T2 image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

Shape: Fractional Reduction in Non-Enhancing Tumor Volume The best EOR shape feature by hazard ratio in the combined patient data set was for fractional reduction in non-enhancing tumor volume, Figure 5.16. We found that less than 93% reduction in non-enhancing tumor volume was associated with a worse prognosis and hazard ratio of 1.81. Although the 93% cutoff was quite stable in cross-validation the survival curves in Figure 5.16 show some possible overfitting since the black and blue curves separate slightly.

Conventional extent of resection is based on reduction of contrast enhancement. However, we found a larger hazard ratio for non-enhancing volume reduction. This is likely because the lower grade tumors are non-enhancing so a reduction in enhancing volume is not measurable. The difference in survival was not significant in multivariate analysis when controlled for age and KPS (or grade).

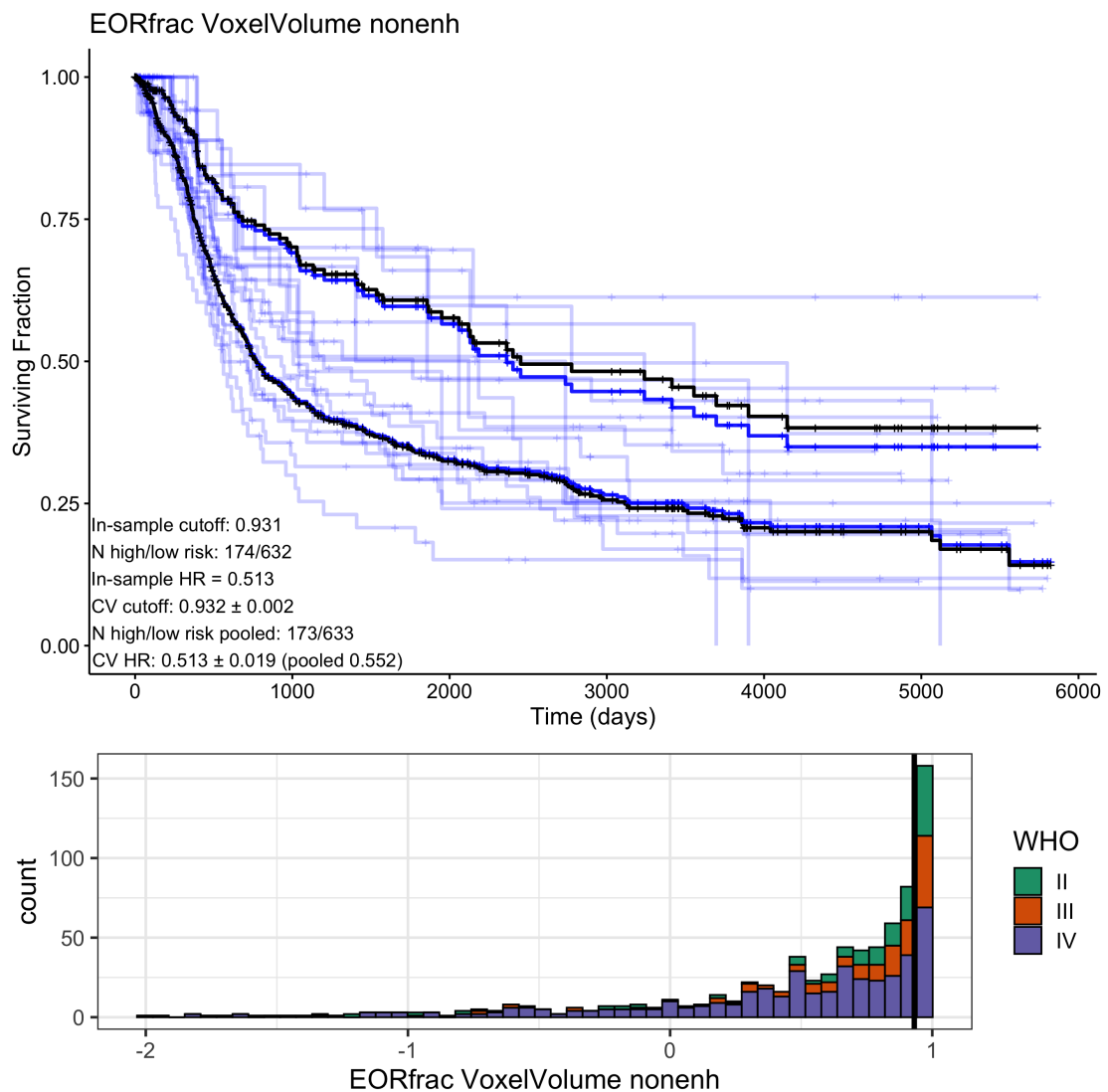


FIGURE 5.16: Best extent of resection shape feature for stratifying survival for all WHO grades (II III IV). Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

ERG: Reduction in Maximum ERG Expression Reduction in maximum ERG (0.1 cc constraint) over the whole-tumor VOI by less than 0.479 percentage points was

associated with a worse prognosis and hazard ratio of 1.751, Figure 5.17. The C-index was modest C-index at 0.515. The cutoff was identical between all 10 folds and the in-sample which means it was extremely stable. A majority of patients had the maximum ERG expression change by 1 percentage point or less which is why the optimal threshold is so small.

The hazard ratio remained significant in multivariate analysis alongside age and KPS and reduced slightly to 1.426 while remaining significant. This trend continued when high WHO grade was also included in the multivariate model. The hazard ratio for reduction in maximum ERG then was 1.452. This means that then small reduction in maximum ERG by less than 0.5 percentage points is still an independent prognostic factor to tumor grade with a comparable hazard ratio to the clinical factors.

Although the magnitude of reduction associated with the greatest hazard ratio is small, A reduction in ERG expression can be interpreted as targeting vascular tissue which is likely more aggressive. However, the prognostic value of the EOR feature was not significant in multivariate analysis which means some of the effect is redundant with preoperative mental status or advanced age. Again, these features could be detection biologically relevant extent of resection measures or may just be picking up on differences in tumor volume. Additional scrutiny is required.

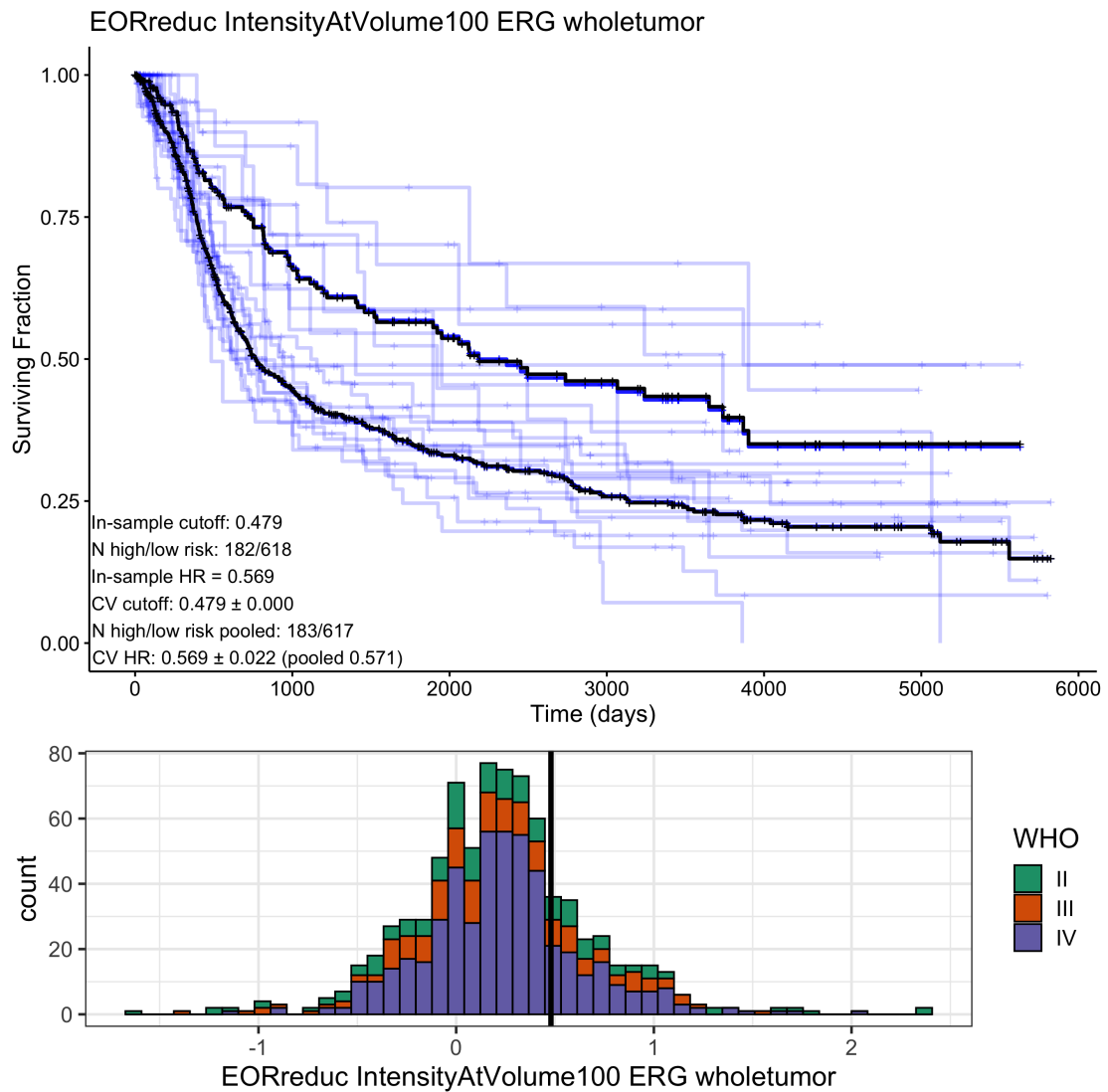


FIGURE 5.17: Best extent of resection feature for stratifying survival for all WHO grades (II III IV) for ERG map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

Ki67: Reduction in Overall Proliferation A reduction of 10th percentile estimated Ki67 of less than 45% inside the whole residual tumor VOI tumor region was associated with worse overall survival and a hazard ratio of 1.616, Figure 5.18. The concordance with overall survival was one of the highest among EOR features as well at 0.546. The optimal cutoff showed a small amount of variability in cross-validation (standard deviation 0.040); but there was still good agreement between the average cutoff and the in-sample value.

Reduction in Ki67 remained significant in multivariate analysis with age and KPS included, the multivariate hazard ratio was 1.455. With WHO grade included as well, the hazard ratio was actually slightly higher at 1.574. This means reduction in 10th percentile estimated Ki67 is an independent prognostic factor with a hazard ratio comparable to reduction in tumor volume (Table 5.6). Like the reduction in cell density, this feature measuring reduction in the lower quantiles of Ki67 is detecting cases where the high Ki67 tumor has been selectively removed, shifting the overall distribution of values down. It is reasonable why this would correlate with a better prognosis since a smaller overall proliferative activity should reduce the aggressiveness of residual tumor.

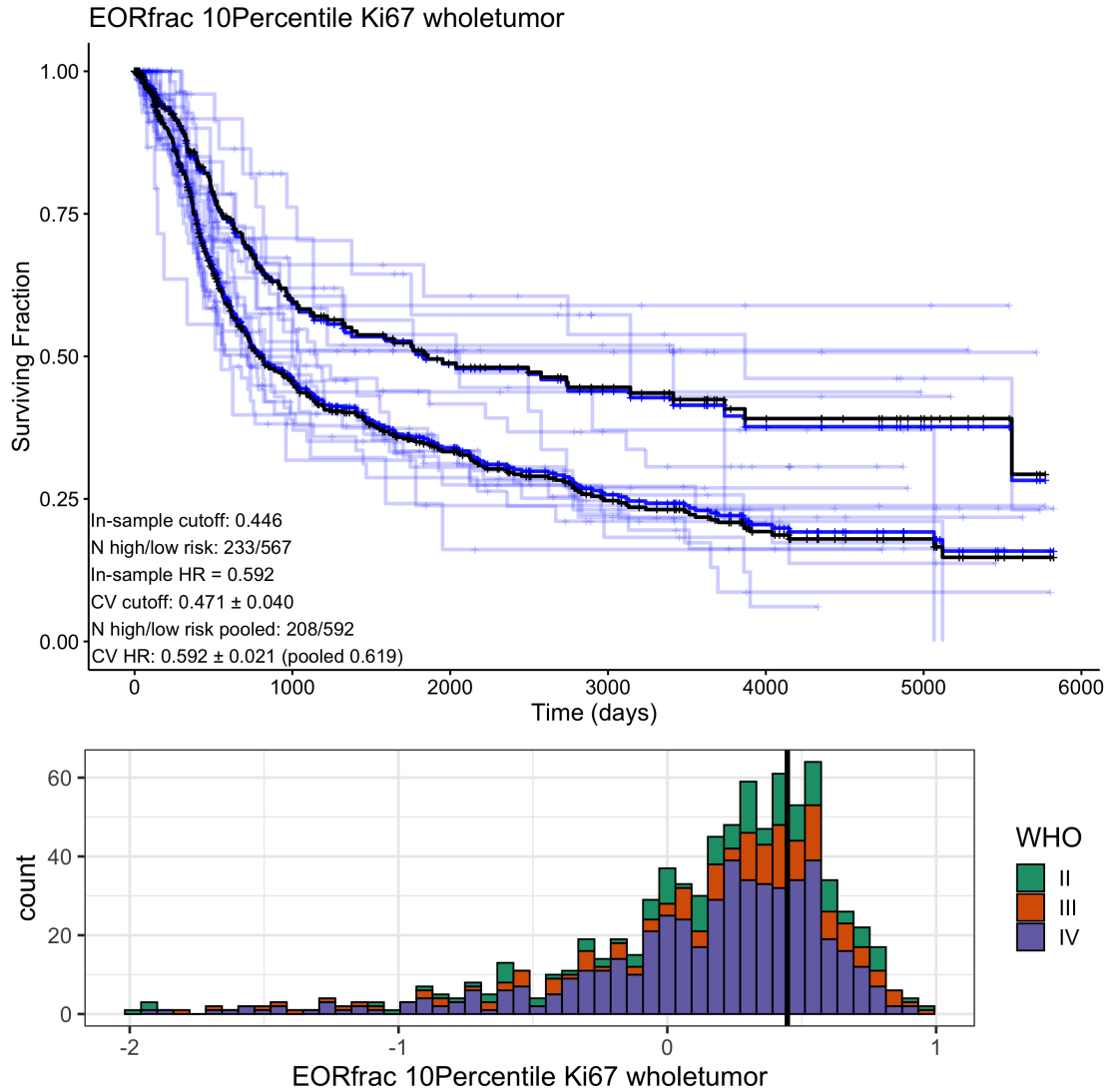


FIGURE 5.18: Best extent of resection feature for stratifying survival for all WHO grades (II III IV) for Ki67 map. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

FLAIR: Reduction in Median FLAIR Intensity A reduction in median FLAIR intensity of less than 21% (on the scale CSF to WM) was associated with a worse prognosis and a univariate hazard ratio of 1.549. This is very similar to the best EOR feature based on T2-weighted imaging (threshold 0.23, HR 1.861) which makes sense due to the similarity between T2-weighted and FLAIR images. Overall, there was very good stability in the optimal threshold among folds of cross-validation and a majority of cases had a median change of less than 50% in either direction, see the histogram in Figure 5.19.

Reduction in median FLAIR intensity remained an independent prognostic factor in multivariate analysis. When age and KPS were accounted for, the hazard ratio for reduction in median FLAIR reduced slightly to 1.388, which is comparable to the hazard ratios for age and KPS themselves. When WHO grade was also included in the covariates, the hazard ratio increased slightly back to 1.442. Overall, this means that reduction in median FLAIR intensity is an independent prognostic factor with a hazard ratio slightly less than the hazard ratio for reduction in tumor volume (1.6).

Reduction in median FLAIR intensity can be achieved by preferentially targeting the high-intensity FLAIR regions. This is consistent with current surgical treatment that focuses on T2 and T2-FLAIR hyperintensity in addition to contrast enhancement. Reduction in FLAIR hyperintense volume has also been established as a favorable prognostic indicator in high-grade gliomas [15].

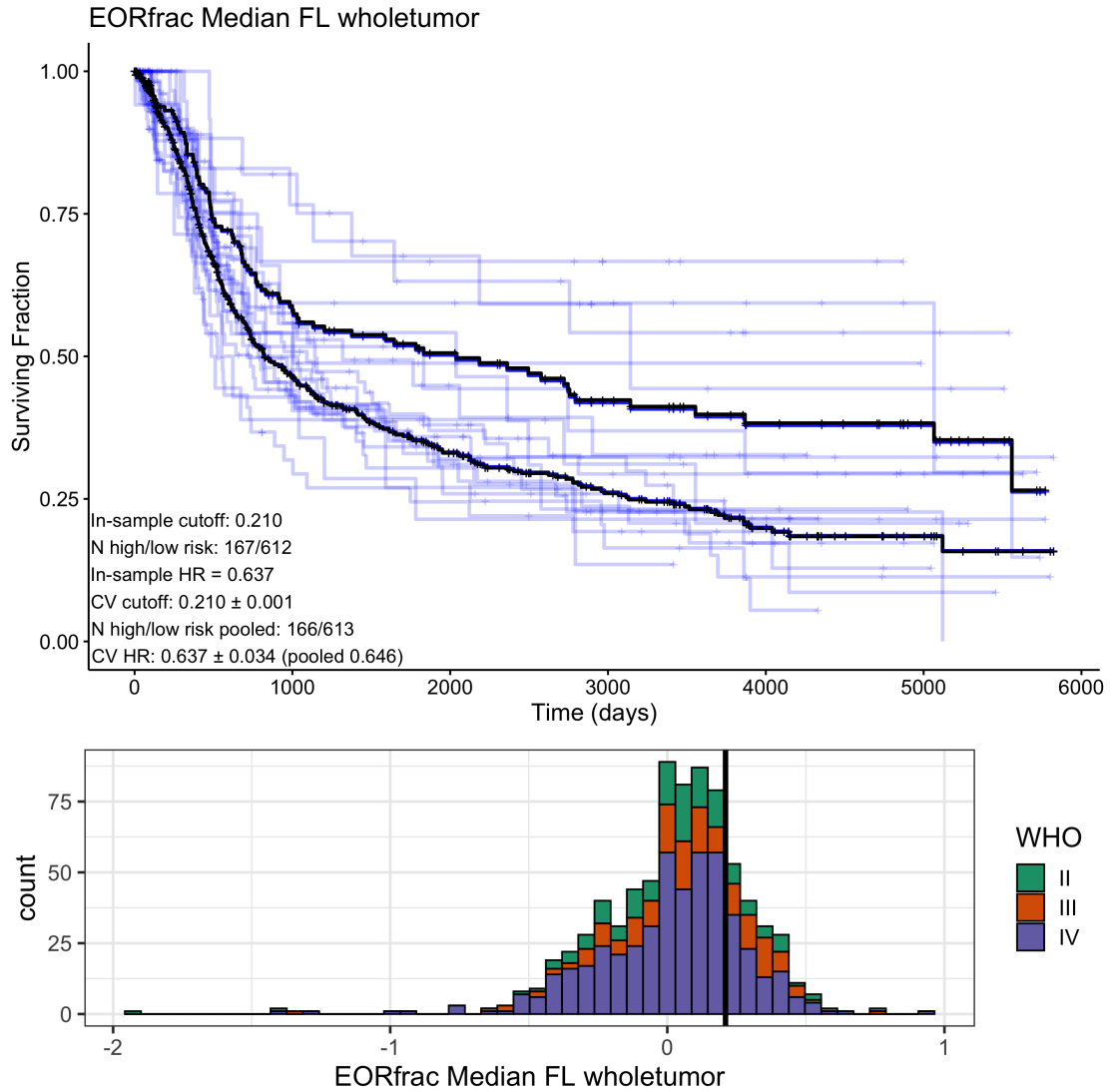


FIGURE 5.19: Best extent of resection feature for stratifying survival for all WHO grades (II III IV) cases for FLAIR image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

T1: Reduction in T1 Pre-Contrast Intensity A reduction in overall (10th percentile) T1 pre-contrast intensity over the enhancing sub-region by less than 0.286 (CSF to brain scale) was associated with a worse prognosis and hazard ratio of 1.446, Figure 5.20. The C-index was reasonable at 0.529, comparable to other EOR features. The in-sample optimal threshold of 0.286 was very stable too and equal to the mean cross-validated threshold.

In multivariate analysis compared to age and KPS, the hazard ratio decreased slightly

to 1.349 and remained significant. This is still comparable to the hazard ratios for age and KPS themselves. With WHO grade included in the analysis as well, the hazard ratio decreased minimally to 1.317 and remained significant. So, while T1 pre-contrast intensity was still an independent prognostic factor, its effect is smaller than reduction in bulk tumor volume (hazard ratio 1.6).

Interpretation of this feature is difficult since it measures an overall brightness of the pre-contrast intensity over the enhancing sub-region. The change preoperative to post-operative is also centered around zero (histogram in Figure 5.20) which means the T1 brightness was not systematically reduced on average by surgery. Further investigation is needed to examine how the reduction in T1 brightness relates to overall survival.

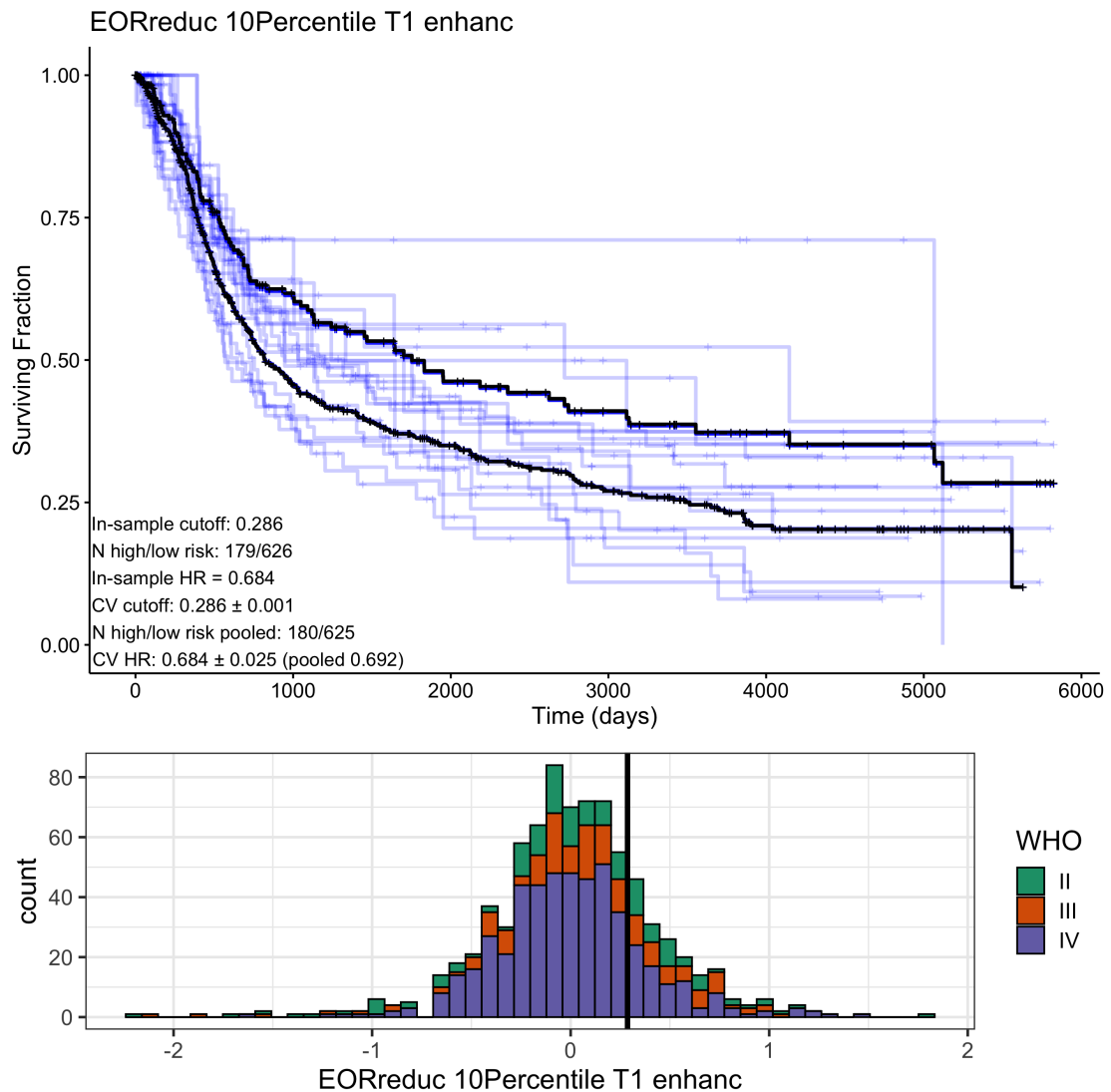


FIGURE 5.20: Best extent of resection feature for stratifying survival for all WHO grades (II III IV) for T1 image. Transparent blue curves are the predictions from each of the 10 folds of cross-validation. Solid blue curves are the pooled cross-validation predictions. Black curves are in-sample survival predictions. Ideally, the cross-validation and in-sample predictions coincide perfectly and appear superimposed. Bottom: Histogram of the feature values with the in-sample cutoff value shown with a black line.

Subset Analysis For WHO II subset, reduction in median cell density in the peritumoral edema was associated with markedly improved prognosis (hazard ratio 4.37, Table A.14). However, statistical significant was not reached which may be due to the low number of events in each group. For these tumors, the edema class makes a large majority of the visible tumor since there is little to no contrast enhancement. This means that targeting highly cellular tumor may be beneficial for these low-grade cases. Similarly to the preoperative data, more investigation is warranted. Table A.14 also

shows how similar features from ERG, Ki67, and predicted grade all potentially outperform conventional image features for this low-grade subset with results that are trending towards multivariate significance.

5.4 Discussion

In this chapter we evaluated postoperative image features and extent of resection (EOR) measurements based on the difference between preoperative and postoperative features. We confirmed the existing relationships between reduction in contrast enhancing tumor and improved survival after surgery. We also found several EOR features based on differences (deltas) between preoperative and postoperative features were significantly associated with survival in cross-validation. For postoperative features alone, conventional image features from T1 pre-contrast and FLAIR images showed the strongest prognostic value. This may be simply due to the fact that treatment effect on postoperative data adds noise to the comparable preoperative feature measurements that may otherwise be more prognostic. However, when we computed the differences between preoperative and postoperative features we found several features from estimated cell density maps and T2-weighted images that correlated best with survival among the combined patient cohort. The feature with the largest hazard ratio was a reduction in overall cellularity. This is consistent with the overall hypothesis that removing high grade malignancy (e.g. highly cellular disease) can improve prognosis. We also found some promising and strongly prognostic postoperative and EOR features for the low-grade WHO II subset based on cellularity but statistical significance was not achieved.

While extensive work has been done to correlate preoperative image features with survival in glioma patients, much fewer studies have looked at postoperative imaging or compared preoperative and postoperative measurements. These sorts of analyses are generally referred to as “delta radiomics” since they examine changes in image features to quantify treatment. Previous studies looking at non-small cell lung cancer [157, 158], colorectal liver metastases [159], and renal cell cancer [160] have found good correlation between delta radiomic features and survival.

In brain, Zhang et al. used a delta-radiomics approach to differentiate progression from radiation necrosis after radiosurgery. They found that delta radiomics features

had increased predictive value than traditional radiomics, although they examined the (assumed linear) difference over time after treatment rather than before and after treatment [161]. A few studies have used postoperative imaging features to correlate with survival specifically in glioblastoma patients. A recent study by Garcia-Ruiz et al. found enhancement thickness was a significant predictor of survival in 144 glioblastoma patients [162]. The authors also found a radiomics signature composed of 12 radiomics features (10 from texture) stratified short- and long-term survivors. Peeken et al. also found contrast enhancing thickness and other postoperative features were predictive of survival [163] and a large study by Ellingson et al. found a threshold of 12 ml of residual enhancing volume was significantly associated with glioblastoma patient survival [164]. However, none of these studies examined the difference between preoperative and postoperative features. Machine learning models have also been previously proposed to guide surgical treatment. Rathore et al. proposed estimated maps of tumor infiltration to predict the location of recurrence [165]. They found that areas of future recurrence showed lower T2w and T2-FLAIR intensity as well as higher T1 and T1CE intensity.

5.4.1 Limitations

One limitation of this work is the use of immediate or early postoperative imaging. The presence of swelling, bleeding, or inflammation of brain tissues due to surgery may manifest as treatment effects that can be mistaken for residual tumor. These effects are maximized on early (within 48 hours) postoperative images [124, 126]. As a consequence, tumor volume measurements using FLAIR volume overestimate the actual residual disease [123]. We see this reflected in our data as well. One method for handling treatment effect related to enhancement is to use T1-subtraction images [164, 166]. However this would need to be performed manually. A better alternative would be to use intraoperative or late (weeks) postoperative imaging assess residual tumor but each of these carries its own disadvantages. Intraoperative MRI does not have the same resolution or variety of sequences present in diagnostic postoperative MRI and furthermore it is only available at a select few academic centers. Late postoperative MRI may be a better measurement but cannot be used to inform a patient's extent of resection immediately following surgery. Also, it is possible for some tumor recurrence or progressing between surgery and imaging.

Another solution to this problem would be to use deformable image registration between preoperative and postoperative imaging. By mapping the preoperative tumor segmentation to the postoperative image it is possible to see what tumor remained versus what was removed (i.e. mapped to the operative cavity). However, this registration problem is difficult and ill-posed due to the necessary missing correspondence of brain tissues.

Finally, we defined potential EOR measures in a way that mimics conventional measurements i.e. as an absolute or fractional reduction. However, fractional reduction produces values between 0 and 1 only for features that are strictly positive and that decrease between preoperative and postoperative measurements. Our normalization scheme based on whole brain modal intensities and cerebrospinal fluid creates some negative image intensities. This means some features like minimum intensity also take on negative values and therefore provide extent of resection measurements greater than 1. Furthermore, the non-linear nature of the fractional EOR mapping for values around 0 creates a potentially unstable mapping between actual disease reduction and measured EOR. This complicates survival analysis since discovered cutoff values may not be physically meaningful. A solution would be to restrict analysis to only features that take on positive values and are decrease by resection so that EOR lies in $[0, 1]$.

Chapter 6

Discussion

The three goals of this work were to:

1. Establish the degree of correspondence between local image characteristics and tissue pathology from biopsy samples. This also includes generation of predictive models to map tumor biological heterogeneity using imaging.
2. Explore the impact of preoperative biological heterogeneity on survival, as predicted by our models, and compare with the information in raw image features.
3. Determine the ability for prognostic imaging predictions to be used to guide surgical intervention by examining new possible definitions of extent of resection.

In Chapter 2 we found that MR imaging was strongly predictive of cellular proliferation, cellular density, vascularity, and local grade using advanced MRI techniques. Many of these models have already been published [49–51]. Techniques like DCE and diffusion weighted imaging present functional physiologic data that machine learning models like random forest can leverage to predict pathology in new patients. We also found that reasonable predictive accuracy could be used to predict pathology using just routine MR sequences like T1 weighted, T2 weighted, T1 post contrast, and FLAIR. These routine sequences do not directly provide quantitative or functional information but have been developed over the years to increase lesion conspicuity and help treat glioma patients. So, it is not surprising that the intensity characteristics are also predictive of underlying pathology.

In Chapter 3 we curated a large historical data set consisting of clinical data, imaging data, and estimated pathology mappings. This data served as the bases for survival and extent of resection analysis. This curation effort overcomes tremendous entry barriers to large scale image analysis. Namely, the organization and labeling of diagnostic images from the hospital PACS system and translation into a format readily usable by research software and feature extractors. Through this curation we not only enabled our own analysis but provide a valuable resource to future research efforts who can use the retrospective data to generate and explore hypotheses. We also established a robust data review procedure to ensure that the final products of image processing were high quality. Without this manual review, it is nearly certain that incorrect data and measurements would be included in final analysis. This would undermine the credibility and impact of results.

In Chapter 4 we addressed the prognostic ability of simple image features to designate prognostic risk groups. Our cross-validated radiomics analysis confirmed the known relationships between contrast enhancement and survival. Although, we found quantitative features based on intensity were prognostic, not just total enhancing volume. This possibly suggests ways to better define enhancing tumor and target it for treatment. We also analyzed measurements related to the predicted biological heterogeneity. Several such features were still prognostic independently of clinical factors. For low grade (WHO II) tumors, estimated cellularity was by far the best prognostic feature and outperformed all the raw image features. These low-grade tumors usually have a homogenous imaging appearance which presents no focal target for therapy. This means estimated cellularity may be a novel way to focus treatment for these patients.

Interestingly, no single biological feature out-performed contrast enhancement. Since the T1 contrast image is an input to the random forests that predict grade, Ki67, CD, and ERG, we expected to see an increase in prognostic power using the model predictions. It is possible that the random forest had insufficient training data to fully capture the information in contrast enhancement. We admit a limitation of the biopsy training data was that only 7 of the 52 samples were collected from contrast enhancing regions. More training data covering a wider range of enhancement values may help fill the gap in prognostic ability. A more detailed subset analysis, for example in non-enhancing tumors where imaging is more ambiguous, may identify areas where estimated biology adds additional information over raw imaging.

In Chapter 5 we showed how postoperative image features and changes in image features between preoperative and postoperative MRIs showed differences in survival. We found multiple extent of resection based features that were prognostic, including the known features like reduction in contrast enhancing volume. Using raw image features we also discovered several other features based on T2-weighted or FLAIR images that could be alternative extent of resection measurements. Finally, we also found that features based on reducing overall cellularity, ERG or Ki67 were also strongly correlated with prognosis.

It is reasonable to hypothesize that removing highly cellular, vascular, or proliferative tumor would improve prognosis. But, this hypothesis was previously untestable since these quantities could only be measured on tissue samples that was actually removed. Not to mention infeasible to perform on every piece of a surgical specimen. Our models provide a way to estimate those biological quantities throughout a glioma before and after surgery, thus finally presenting evidence to support a hypothesis that removing histologically malignant tumor improves survival.

To the best of our knowledge, there has been no study of preoperative vs postoperative radiomics features in brain, likely due to the challenge of working with postoperative data. We avoided the image registration problem by measuring features separately on preoperative and postoperative images. However, we saw a lot of interference from treatment effect due to the early time point of the acquisitions. Nonetheless, we found cross-validated measures of extent of resection that significantly correlated with survival. Several of which, like reduction in maximum intensity, could be used to guide surgical intervention.

In summary, this thesis developed models to synthesize MR imaging into predictive maps of tumor pathology. We validated model predictions as prognostic biomarkers and targets for intervention. These results can be applied to prospectively guide intervention, identify heightened malignancy, and provide useful information in patients where gold-standard tissue data is otherwise unavailable. This work greatly improves our understanding of intratumoral heterogeneity and builds on familiar, well known histologic characteristics that are interpretable and meaningful to clinicians.

6.1 Future Work

Immediate future work can start validating predictive models in their ability to estimate the Ki67 expression, cell density, ERG expression, and local grade of biopsy samples in new patients. After comparing predictions to actual values, the new pathology data could be incorporated into the training data to improve the model overall. This is a key step towards moving these models to clinical use. Another key area of future work would be expanding the target variables to include key genetic mutations like IDH, p53, ATRX, and modifications like 1p19q-codeletion. Some of these are heterogeneously expressed and would add tremendous value to our understanding of clonal evolution of gliomas. Some work using image guided biopsies to examine genetic heterogeneity has already been performed [167].

The extensive database of diagnostic images can be further improved in several ways. First, there are several images that were tabulated but not analyzed including T2*-weighted images and diffusion weighted imaging. These could be incorporated into the feature extraction pipeline and correlated with outcomes in the same way as routine imaging. Diffusion features were a key element of the pathology prediction models so they will likely be strong correlates of prognosis. Second, the overall quality of the postoperative data was lower and we had to exclude relatively more data due to processing failures. Postoperative imaging has several treatment-related effects and abnormalities that can confuse algorithms developed in preoperative brain images. In particular, CLARA's tumor segmentation model systematically overestimated tumor volume due to treatment effect. Future work could help these algorithms by training on postoperative imaging directly. Another option is to explicitly accounting for treatment effect using T1-subtraction [164, 166]. Finally, annotation of postoperative data could be avoided by using deformable image registration to the preoperative scans. This is a particularly challenging proposition though due to the missing correspondence caused by resection and the subsequently large deformations in the brain.

We found strong prognostic biomarkers using relatively simple first-order intensity and shape features. This was intentional in order to produce interpretable features. Future work could expand on the feature set by including high order texture or deep-filter features that may capture local neighborhood information. These features have been successfully used to develop imaging signatures in previous studies [35, 149]. In addition

to higher-order features, more advanced predictive modeling techniques like generalized linear models or LASSO [168] could be used to combine features together and produce a risk score that takes advantage of raw image features and estimated pathology.

Appendix A

Appendix A

A.1 Additional Figures

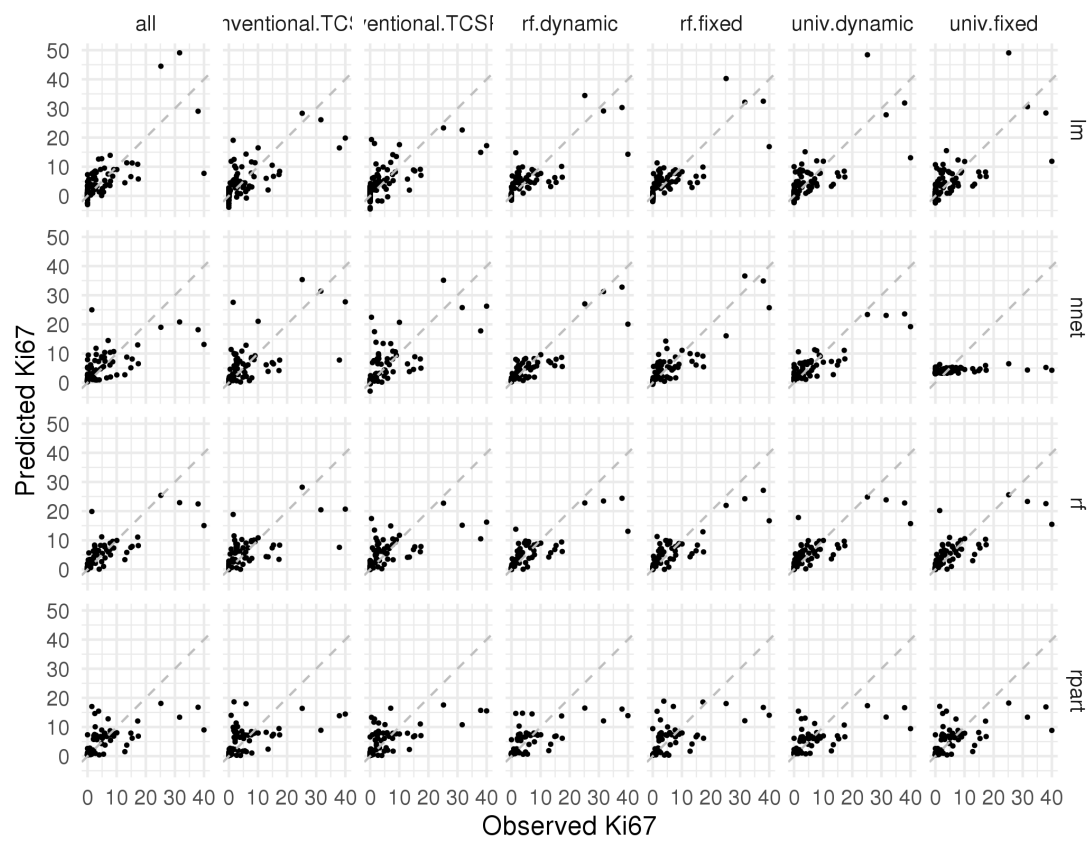


FIGURE A.1: Average predicted and observed Ki67 values

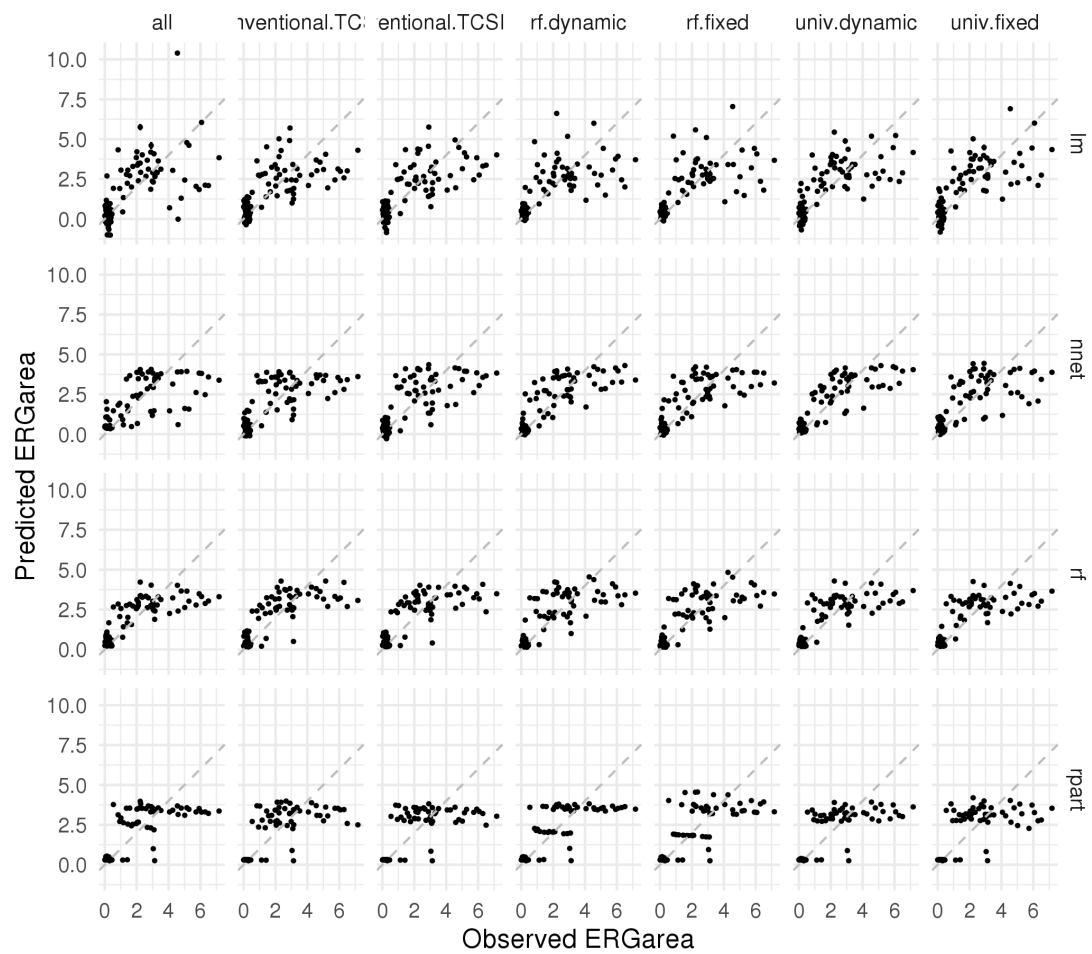


FIGURE A.2: Average predicted and observed ERG values

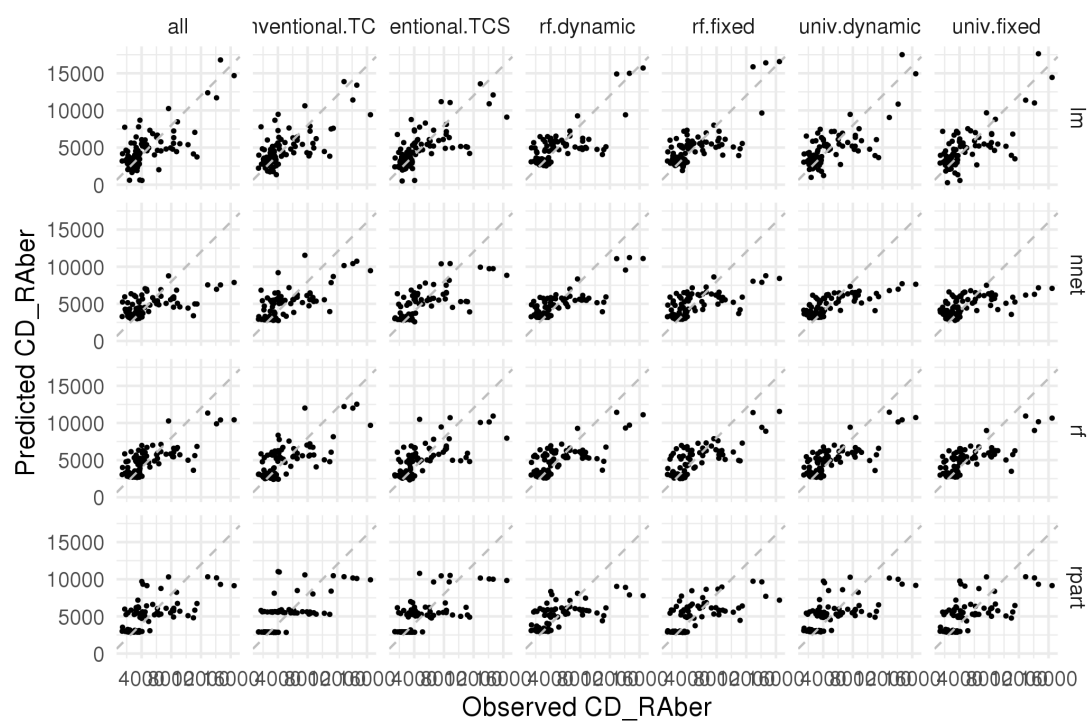


FIGURE A.3: Average predicted and observed CD values

A.2 Reference Tissue and MCSF Normalization Comparison

Overall, there is a good correlation between the original NLM values (as used in [49–51]) and the mode-CSF normalized images used in this work. See Table A.1 for the correlations between biopsy VOI values under the two schemes. The biggest changes are in the T1 post-contrast values, probably because they were previously normalized using the un-reliable putamen gray matter ROIs. Figure A.4 shows the model performance in repeated cross-validation for the old (reference tissue NLM) and new (MCSF) method. The “old_conv” values are for the NLM normalization and the MCSF values are for the new normalization scheme. For the old scheme, only T1C SE was used and in the new scheme there are two models depending on T1C type. The double-cluster nature of the prediction may be artificially inflating the R^2 results, however the RMSE is still smaller which is good. Overall, the results are comparable or even better with MCSF normalization.

	image	Pearson	Spearman
1	T1	0.84	0.83
2	T2	0.98	0.92
3	T1C	0.81	0.76
4	FLAIR	0.90	0.89
5	T1C SPGR	0.89	0.71
6	T2*	0.86	0.86
7	SWAN	0.74	0.75

TABLE A.1: Correlation between NLM and Mode-CSF normalized image intensities.

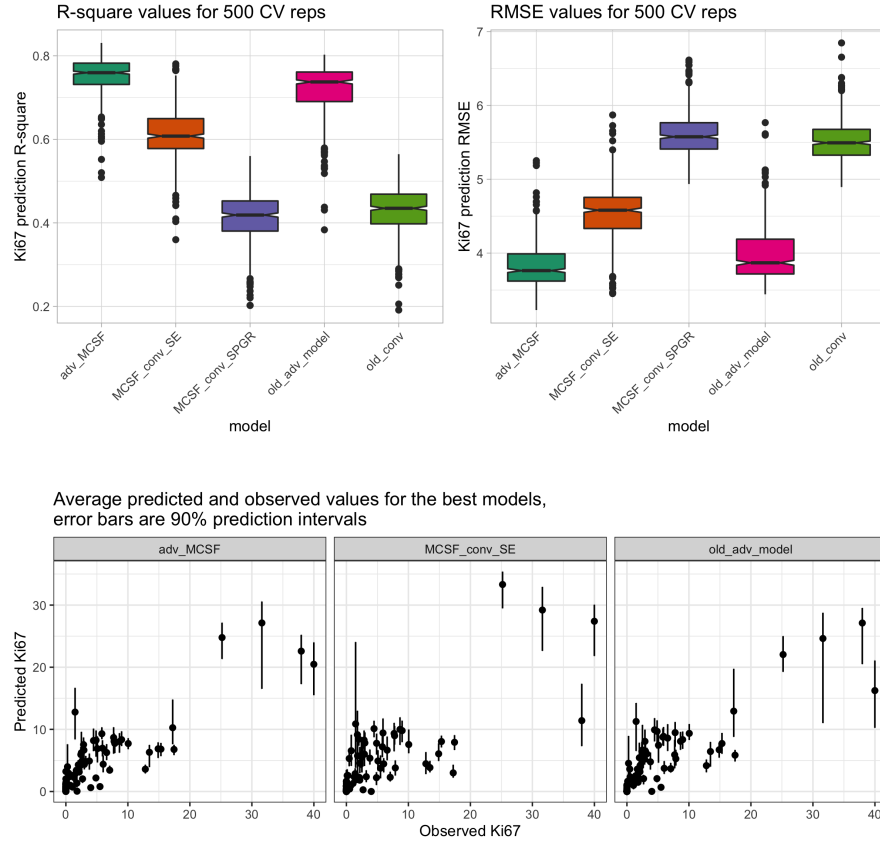


FIGURE A.4: Modeling results for Ki67 using the original NLM normalized and new brain-CSF mode normalized data. Top: R^2 and RMSE values for 500 rounds of 5-fold cross-validation. The conventional model with MCSF variables performs better than the old conventional model but not as well as the advanced variable model. The predicted and observed values for the three best models are shown on the bottom. The conventional model still has a high Ki67 sample that it struggles with, limiting the R^2 value.

A.3 Stereology and Normal Brain White Matter Cellularity

There are several methods that could be used to estimate of normal white matter cell density in order to impute values for virtual biopsies. An excellent review of all these methods is written in 2016 by von Bartheld, Bahney, and Herculano-Houzel [171]

1. The **isotropic fractionator method** presented by Herculano-Houzel [172, 173] determines the total number of cortical white matter nuclei present in the brain by dissolving tissue and counting nuclei in a homogeneous sample. A disadvantage is that tissue structure is not preserved. Azevedo et al. [174] used this method to estimate that there are approximately 21.17 ± 2.88 billion cells in half the brain's

white matter. This can be combined with total white matter volume given by Lüders [1] ($0.42 \pm 0.06 \text{ dm}^3$ for the whole brain) to estimate the nuclei per unit volume. As a note, the patient demographics were quite similar between these references so the values are comparable. A final assumption is that the nuclei per area as measured with a microscope will be equal to the nuclei per volume times the section thickness. While it sounds reasonable, *a better understanding of the stereology involved and references are needed to confirm this is the case*. The isotropic fractionator method also gives uncomfortable low cell density estimates around 403 nuclei per square mm.

2. Another method is **histology**. Recently a work by Roetzer [89] scanned $6 \mu\text{m}$ coronal sections of a whole brain and measured the cell density in the white matter directly. This is a much closer method to ours since it involves fixing, section, and straining tissue. The cell density estimates for white matter, cortex, and tumor, seem to agree with our measurements but again we have to assume *cell density (per area) is proportional to section thickness*. Again, references are needed to confirm this is the case.
3. Finally, **stereology** claims to be unbiased and uses uniform random sampling and some basic assumptions along with Cavalieri's principle to estimate cell numbers in sectioned samples. The primary references for this method are by Gunderson (1985, 1986, 1988).

Summary: much work has been done to quantify the total cell numbers in the brain but Roetzer seems to be one of the best references for white matter cell density. A comparison of 2D (i.e. stereology) and 3D methods is given in [179].

A.4 Survival Analysis on Individual Grade Subsets

A.4.1 Preoperative Tumor Volume

	HR	CI	p	
KPS < 70	3.145	[1.300, 7.609]	1.10e-02	*
Age > 55	2.477	[1.130, 5.429]	2.35e-02	*

TABLE A.2: Multivariate cox proportional hazards model hazard ratios and significance for grade II cases using only clinical factors that strongly influence survival. Confidence intervals are 95%

	HR	CI	p	
KPS < 70	3.219	[1.786, 5.802]	9.99e-05	***
Age > 55	2.638	[1.652, 4.212]	4.84e-05	***

TABLE A.3: Multivariate cox proportional hazards model hazard ratios and significance for grade III cases using only clinical factors that strongly influence survival. Confidence intervals are 95%

	HR	CI	p	
KPS < 70	1.587	[1.343, 1.875]	5.63e-08	***
Age > 55	1.859	[1.557, 2.219]	6.96e-12	***

TABLE A.4: Multivariate cox proportional hazards model hazard ratios and significance for grade IV cases using only clinical factors that strongly influence survival. Confidence intervals are 95%

	HR	CI	p		multivariate HR	multivariate CI	multivariate p	
CLARA_PREVOL2 > 73.07	1.910	[0.981, 3.719]	5.71e-02	.	1.964	[1.006, 3.835]	4.80e-02	*

TABLE A.5: Cox proportional hazards model hazard ratios and significance for tumor volumes among grade II cases, multivariate ratios include controls for age, and KPS

	HR	CI	p		multivariate HR	multivariate CI	multivariate p	
CLARA_PREVOL1 > 42.83	1.676	[1.051, 2.673]	3.01e-02	*	1.594	[0.975, 2.608]	6.32e-02	.
CLARA_PREVOL2 > 109.17	2.302	[1.468, 3.610]	2.82e-04	***	1.860	[1.140, 3.035]	1.30e-02	*

TABLE A.6: Cox proportional hazards model hazard ratios and significance for tumor volumes among grade III cases, multivariate ratios include controls for age, and KPS

	HR	CI	p		multivariate HR	multivariate CI	multivariate p	
CLARA_PREVOL1 > 9.55	1.542	[1.260, 1.887]	2.57e-05	***	1.421	[1.156, 1.747]	8.60e-04	***
CLARA_PREVOL2 > 110.63	0.871	[0.726, 1.045]	1.38e-01		0.771	[0.640, 0.930]	6.59e-03	**

TABLE A.7: Cox proportional hazards model hazard ratios and significance for tumor volumes among grade IV cases, multivariate ratios include controls for age, and KPS

A.4.2 Preoperative Image Features

image	region	feature	IS		Univariate			Multivariate (age+KPS)			
			C	Cut	HR	95% CI	p	HR	95% CI	p	
CD	wholetumor	10Percentile	0.594	3.51e+03	10.237	[1.41, 74.57]	2e-01	10.757	[1.47, 78.81]	2e-02	*
Ki67	wholetumor	10Percentile	0.572	1.52	6.279	[1.51, 26.06]	2e-01	6.323	[1.51, 26.43]	1e-02	*
T2	nonenh	Maximum	0.585	0.758	4.592	[1.11, 19.06]	3e-01	4.818	[1.16, 20.05]	3e-02	*
T1	nonenh	Maximum	0.531	1.83	4.197	[2.11, 8.34]	2e-02	4.521	[2.13, 9.61]	9e-05	***
ERG	nonenh	RootMeanSquared	0.552	2.5	2.240	[0.87, 5.74]	3e-01	1.973	[0.76, 5.09]	2e-01	
TC	wholetumor	Maximum	0.572	1.78	2.127	[0.98, 4.64]	3e-01	1.892	[0.86, 4.18]	1e-01	
GR	higher	Maximum3DDiameter	0.529	52.2	1.772	[0.94, 3.36]	3e-01	1.761	[0.93, 3.34]	8e-02	.
-	brainNoCSF	VoxelVolume	0.500	1.24e+03	1.730	[0.76, 3.93]	4e-01	1.696	[0.74, 3.89]	2e-01	
FL	brainNoCSF	IntensityAtVolume10	0.582	2.71	1.685	[0.82, 3.46]	4e-01	1.589	[0.77, 3.27]	2e-01	

TABLE A.8: Best preoperative image features from each image type among patients with WHO grade(s) II in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample

image	region	feature	IS		Univariate			Multivariate (age+KPS)			
			C	Cut	HR	95% CI	p	HR	95% CI	p	
TC	wholetumor	IntensityAtVolume100	0.554	1.36	2.745	[1.42, 5.31]	4e-02	2.655	[1.37, 5.16]	4e-03	**
T1	edema	Mean	0.522	0.844	2.500	[1.29, 4.84]	5e-02	2.702	[1.39, 5.25]	3e-03	**
Ki67	brainNoCSF	92Percentile	0.565	6.26	2.282	[1.18, 4.42]	7e-02	2.124	[1.08, 4.20]	3e-02	*
-	wholetumor	VoxelVolume	0.577	109	2.037	[1.30, 3.19]	4e-02	1.672	[1.03, 2.71]	4e-02	*
CD	brainNoCSF	Minimum	0.588	2.27e+03	2.008	[1.29, 3.13]	4e-02	1.949	[1.24, 3.05]	4e-03	**
FL	brainNoCSF	91Percentile	0.577	1.56	1.843	[1.16, 2.93]	6e-02	1.692	[1.04, 2.74]	3e-02	*
T2	edema	TotalSum	0.576	1.68e+04	1.809	[1.15, 2.84]	6e-02	1.218	[0.74, 2.02]	4e-01	
ERG	nonenh	TotalEnergy	0.516	3.83e+05	1.771	[1.13, 2.77]	7e-02	1.903	[1.19, 3.04]	7e-03	**
GR	higher	Maximum3DDiameter	0.517	59.1	1.442	[0.90, 2.31]	3e-01	1.334	[0.82, 2.17]	2e-01	

TABLE A.9: Best preoperative image features from each image type among patients with WHO grade(s) III in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample

image	region	feature	IS		Univariate				Multivariate (age+KPS)			
			C	Cut	HR	95% CI	p		HR	95% CI	p	
TC	wholetumor	RootMeanSquared	0.572	1.07	1.825	[1.47, 2.26]	1e-05	***	1.616	[1.30, 2.01]	1e-05	***
T1	wholetumor	10Percentile	0.547	0.476	1.563	[1.27, 1.92]	3e-04	***	1.479	[1.20, 1.82]	2e-04	***
-	enhanc	Maximum3DDiameter	0.590	40.5	1.553	[1.28, 1.89]	2e-04	***	1.465	[1.20, 1.79]	2e-04	***
FL	nonenh	10Percentile	0.527	0.972	1.528	[1.23, 1.90]	2e-03	**	1.659	[1.33, 2.06]	5e-06	***
CD	brainNoCSF	98Percentile	0.513	7.07e+03	1.452	[1.17, 1.81]	6e-03	**	1.271	[1.02, 1.59]	4e-02	*
Ki67	brainNoCSF	92Percentile	0.583	8.19	1.445	[1.23, 1.70]	2e-04	***	1.219	[1.03, 1.45]	2e-02	*
ERG	enhanc	IntensityAtVolume100	0.552	3.81	1.435	[1.18, 1.74]	2e-03	**	1.221	[1.00, 1.49]	5e-02	*
T2	enhanc	Maximum	0.515	0.925	1.284	[1.05, 1.57]	4e-02	*	1.159	[0.94, 1.42]	2e-01	
GR	higher	Maximum3DDiameter	0.535	43.8	1.128	[0.95, 1.34]	2e-01		1.043	[0.88, 1.24]	6e-01	

TABLE A.10: Best preoperative image features from each image type among patients with WHO grade(s) IV in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample

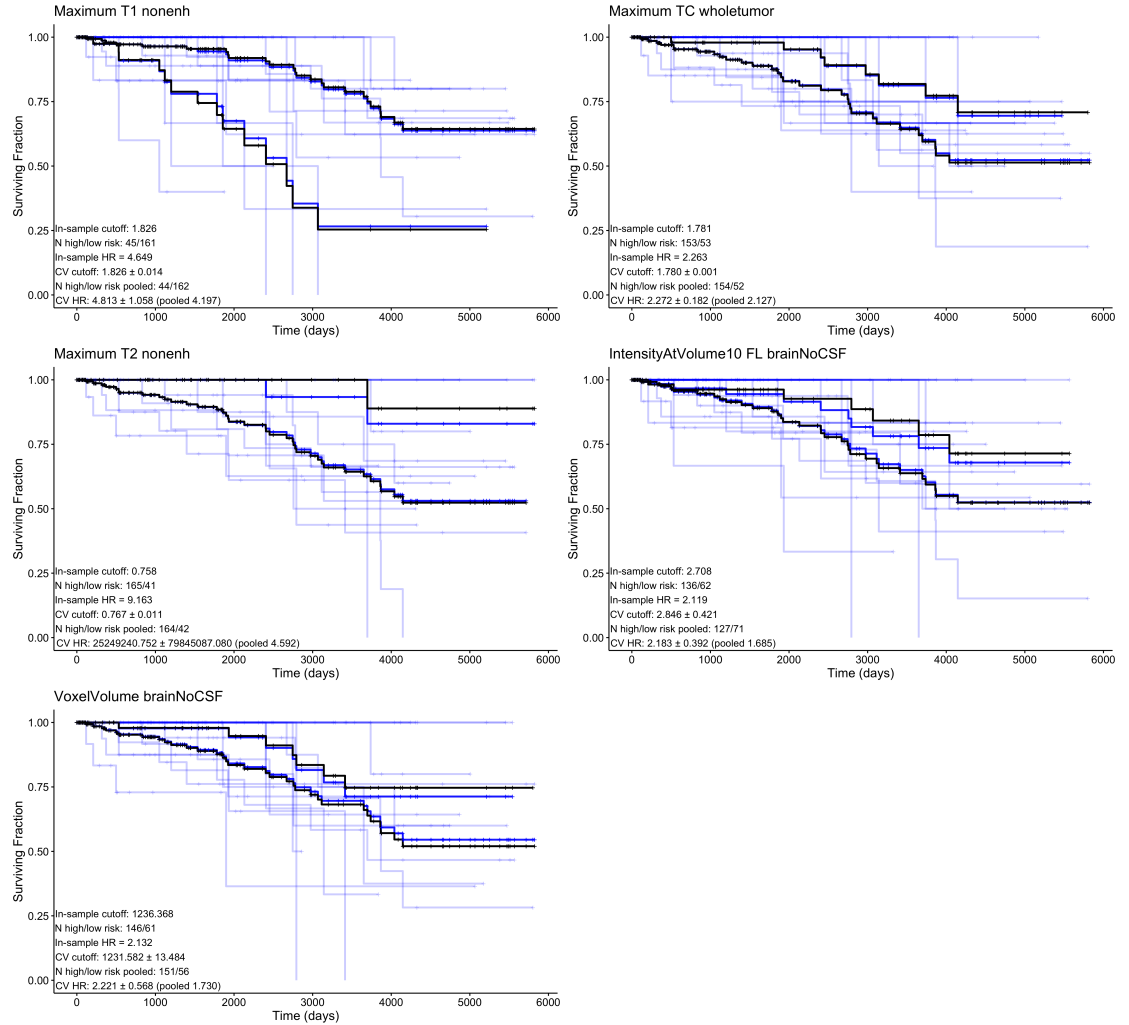


FIGURE A.5: Best features for stratifying survival for WHO grade II cases for each raw image type (T1, TC, T2, FLAIR)

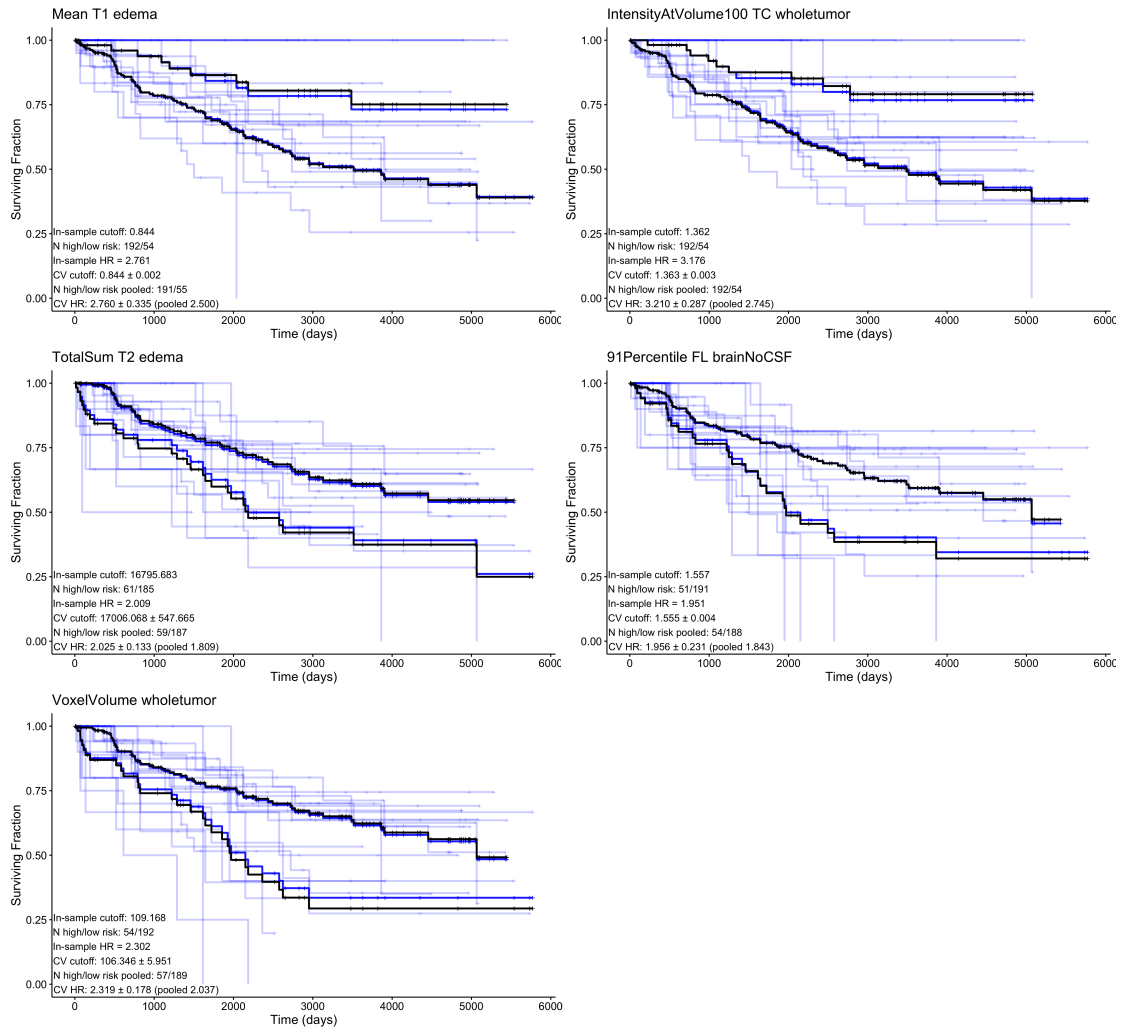


FIGURE A.6: Best features for stratifying survival for WHO grade III cases for each raw image type (T1, TC, T2, FLAIR)

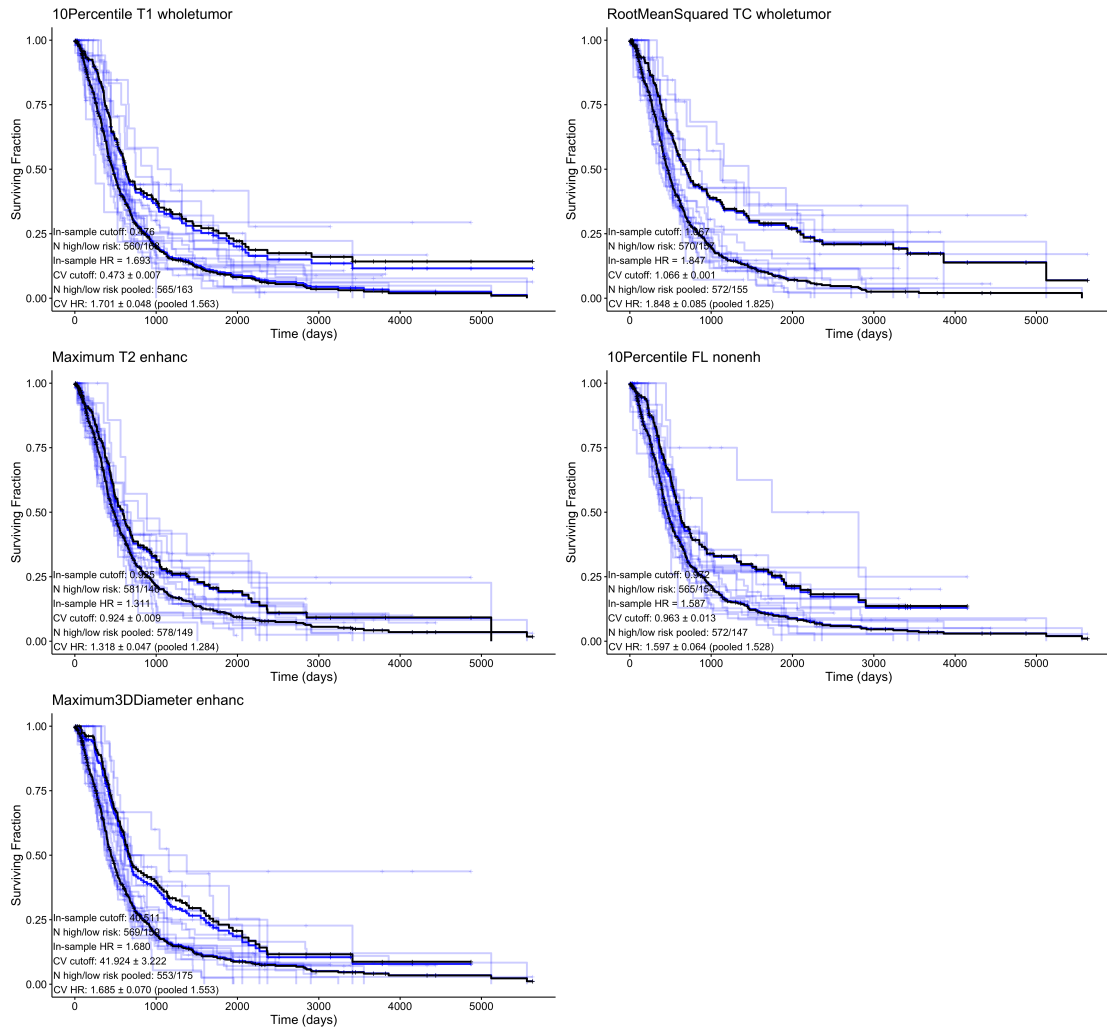


FIGURE A.7: Best features for stratifying survival for WHO grade IV cases for each raw image type (T1, TC, T2, FLAIR)

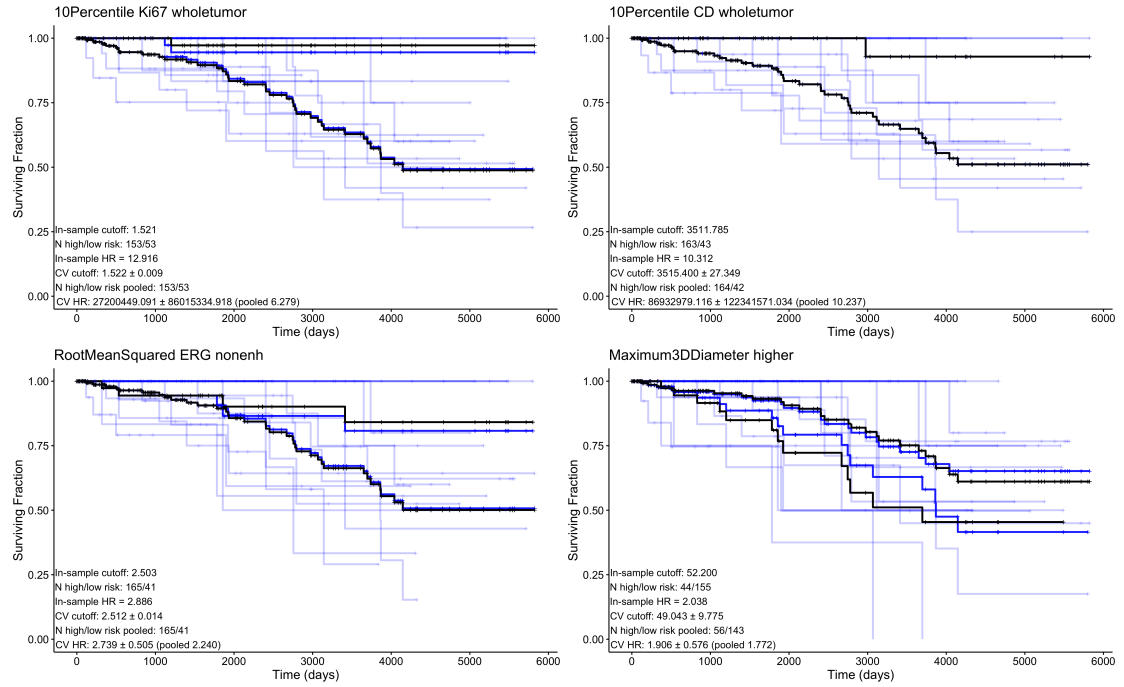


FIGURE A.8: Best features for stratifying survival for WHO grade II cases for each pathology map estimate (Ki67, CD, ERG)

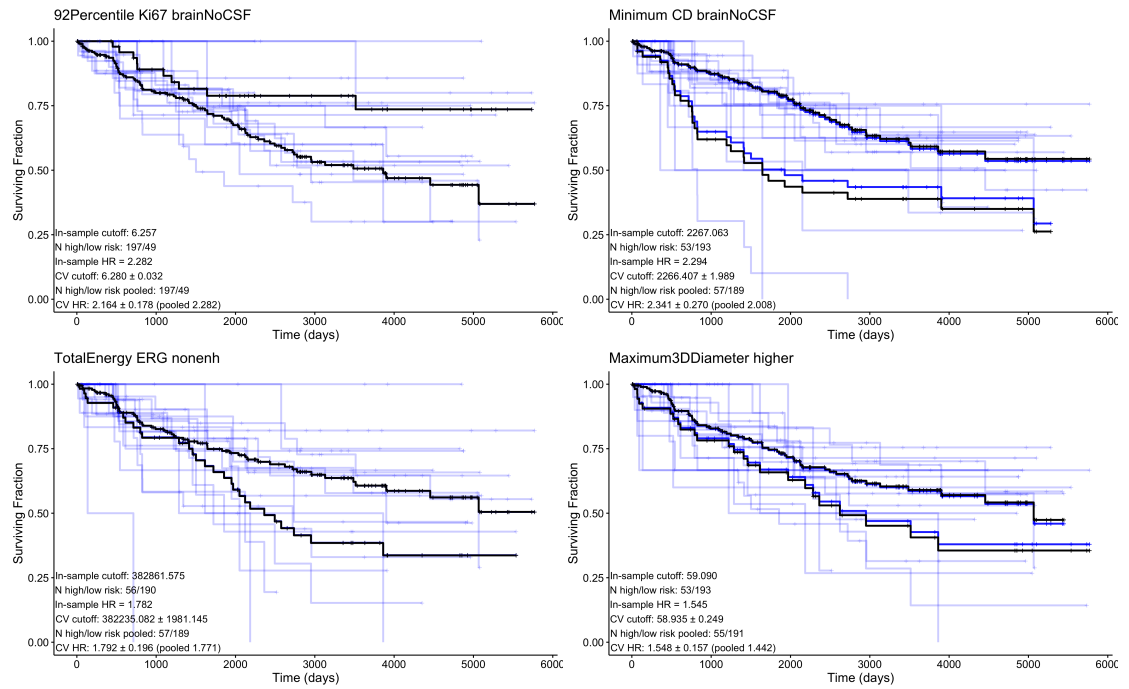


FIGURE A.9: Best features for stratifying survival for WHO grade III cases for each pathology map estimate (Ki67, CD, ERG)

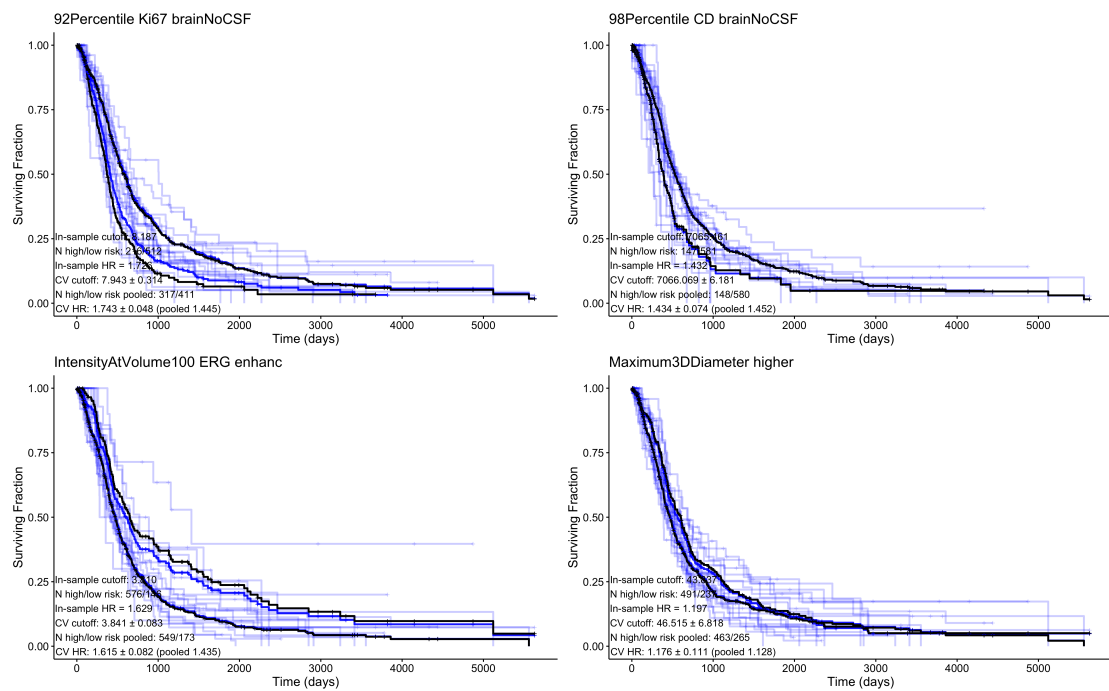


FIGURE A.10: Best features for stratifying survival for WHO grade IV cases for each pathology map estimate (Ki67, CD, ERG)

A.4.3 Postoperative Image Features

image	region	feature	IS		Univariate			Multivariate (age+KPS)		
			C	Cut	HR	95% CI	p	HR	95% CI	p
CD	edema	Median	0.490	4.65e+03	NA	[0.00, Inf]	1e+00	NA	[0.00, Inf]	1e+00
ERG	edema	Maximum	0.569	4.17	7.192	[1.70, 30.51]	2e-01	7.292	[1.72, 30.97]	7e-03 **
TC	nonenh	91Percentile	0.542	1.64	3.191	[1.40, 7.26]	2e-01	3.334	[1.36, 8.16]	8e-03 **
-	wholetumor	VoxelVolume	0.666	28.8	2.723	[1.27, 5.86]	2e-01	2.735	[1.25, 5.99]	1e-02 *
Ki67	wholetumor	IntensityAtVolume1000	0.603	9.31	2.664	[1.28, 5.55]	2e-01	2.602	[1.24, 5.47]	1e-02 *
T1	nonenh	91Percentile	0.525	1.38	2.075	[0.90, 4.76]	3e-01	2.042	[0.89, 4.70]	9e-02 .
GR	higher	Maximum3DDiameter	0.615	57.6	1.822	[0.80, 4.14]	4e-01	1.748	[0.75, 4.07]	2e-01
T2	nonenh	TotalSum	0.568	2.1e+03	1.350	[0.64, 2.86]	6e-01	1.294	[0.60, 2.81]	5e-01
FL	nonenh	TotalEnergy	0.536	1.06e+03	1.220	[0.59, 2.54]	7e-01	1.176	[0.56, 2.49]	7e-01

TABLE A.11: Best postoperative image features from each image type among patients with WHO grade(s) II in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample

image	region	feature	IS		Univariate			Multivariate (age+KPS)		
			C	Cut	HR	95% CI	p	HR	95% CI	p
FL	edema	10Percentile	0.665	1.27	3.912	[2.27, 6.73]	2e-04 ***	4.106	[2.30, 7.33]	2e-06 ***
CD	wholetumor	10Percentile	0.607	3.47e+03	3.211	[1.61, 6.39]	9e-03 **	4.482	[2.15, 9.36]	6e-05 ***
T1	wholetumor	IntensityAtVolume1000	0.616	1.53	2.876	[1.63, 5.08]	6e-03 **	2.428	[1.35, 4.37]	3e-03 **
ERG	edema	Mean	0.563	2.02	2.265	[1.07, 4.80]	1e-01	2.532	[1.16, 5.52]	2e-02 *
TC	enhanc	10Percentile	0.555	1.1	2.109	[1.22, 3.64]	3e-02 *	2.237	[1.30, 3.86]	4e-03 **
T2	nonenh	Mean	0.535	0.301	2.102	[0.90, 4.91]	2e-01	2.596	[1.09, 6.20]	3e-02 *
Ki67	edema	RootMeanSquared	0.538	4.87	1.980	[1.02, 3.84]	1e-01	1.907	[0.97, 3.74]	6e-02 .
-	enhanc	VoxelVolume	0.565	2.23	1.638	[0.96, 2.80]	2e-01	1.428	[0.83, 2.46]	2e-01
GR	higher	VoxelVolume	0.520	0.111	1.106	[0.62, 1.97]	8e-01	0.873	[0.48, 1.60]	7e-01

TABLE A.12: Best postoperative image features from each image type among patients with WHO grade(s) III in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample

image	region	feature	IS		Univariate			Multivariate (age+KPS)		
			C	Cut	HR	95% CI	p	HR	95% CI	p
T2	nonenh	92Percentile	0.534	0.554	1.476	[1.12, 1.94]	2e-01	1.503	[1.14, 1.98]	4e-03 **
FL	enhanc	IntensityAtVolume100	0.508	1.9	1.386	[1.08, 1.78]	2e-01	1.436	[1.11, 1.85]	5e-03 **
ERG	enhanc	IntensityAtVolume100	0.545	3.3	1.373	[1.06, 1.78]	2e-01	1.346	[1.04, 1.74]	2e-02 *
TC	wholetumor	94Percentile	0.537	1.2	1.361	[1.05, 1.76]	2e-01	1.374	[1.06, 1.78]	2e-02 *
-	enhanc	Maximum3DDiameter	0.546	70.3	1.344	[1.06, 1.71]	2e-01	1.280	[1.00, 1.63]	5e-02 *
T1	enhanc	RootMeanSquared	0.508	0.83	1.317	[1.05, 1.66]	2e-01	1.192	[0.94, 1.50]	1e-01
CD	wholetumor	IntensityAtVolume1000	0.548	7.09e+03	1.305	[1.03, 1.65]	2e-01	1.250	[0.99, 1.58]	6e-02 .
GR	higher	VoxelVolume	0.524	0.0192	1.242	[1.01, 1.52]	3e-01	1.165	[0.95, 1.43]	1e-01
Ki67	enhanc	Maximum	0.510	10.3	1.217	[0.97, 1.53]	4e-01	1.153	[0.91, 1.46]	2e-01

TABLE A.13: Best postoperative image features from each image type among patients with WHO grade(s) IV in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample

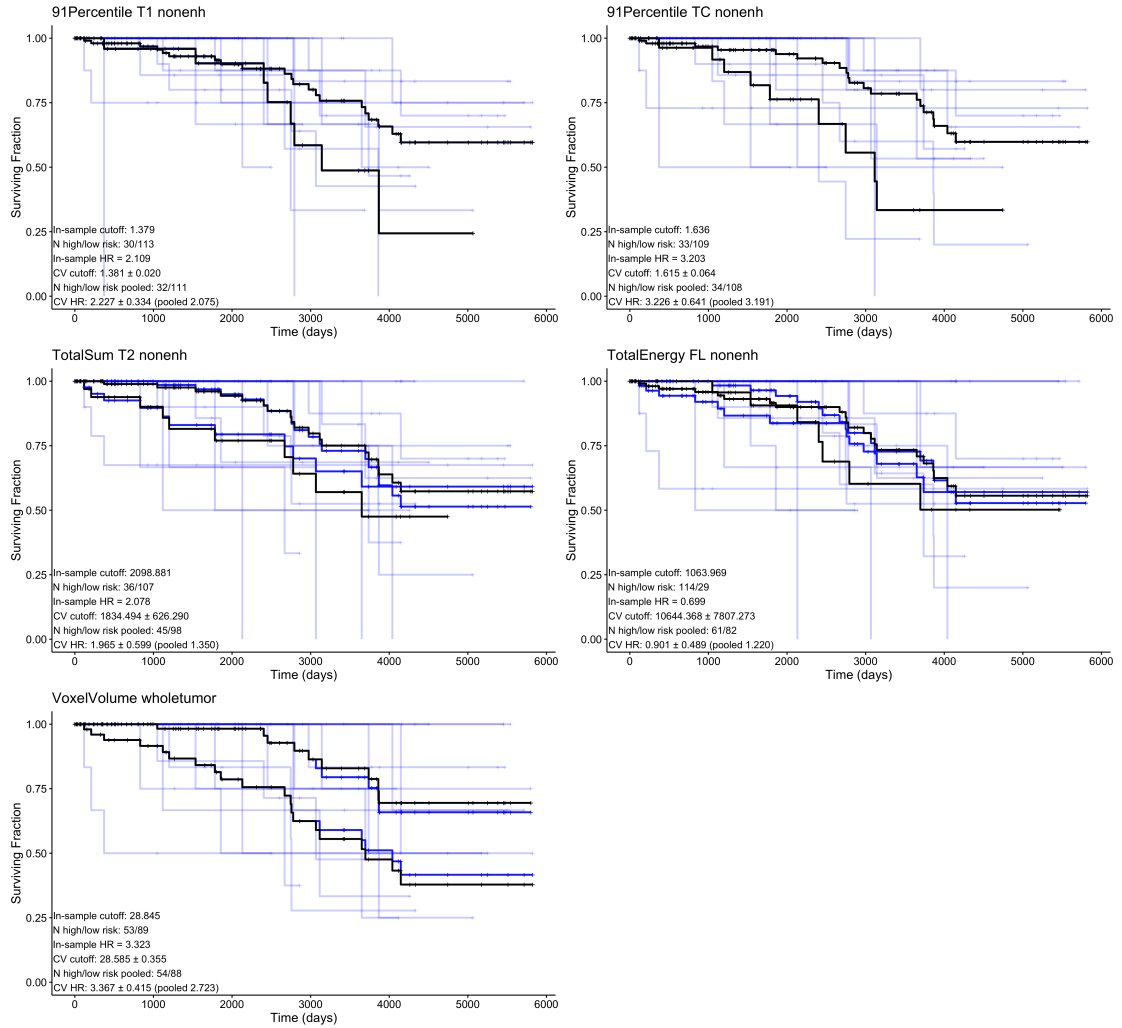


FIGURE A.11: Best extent of resection (EOR) measures for stratifying survival for WHO grade II cases for each raw image type (T1, TC, T2, FLAIR)

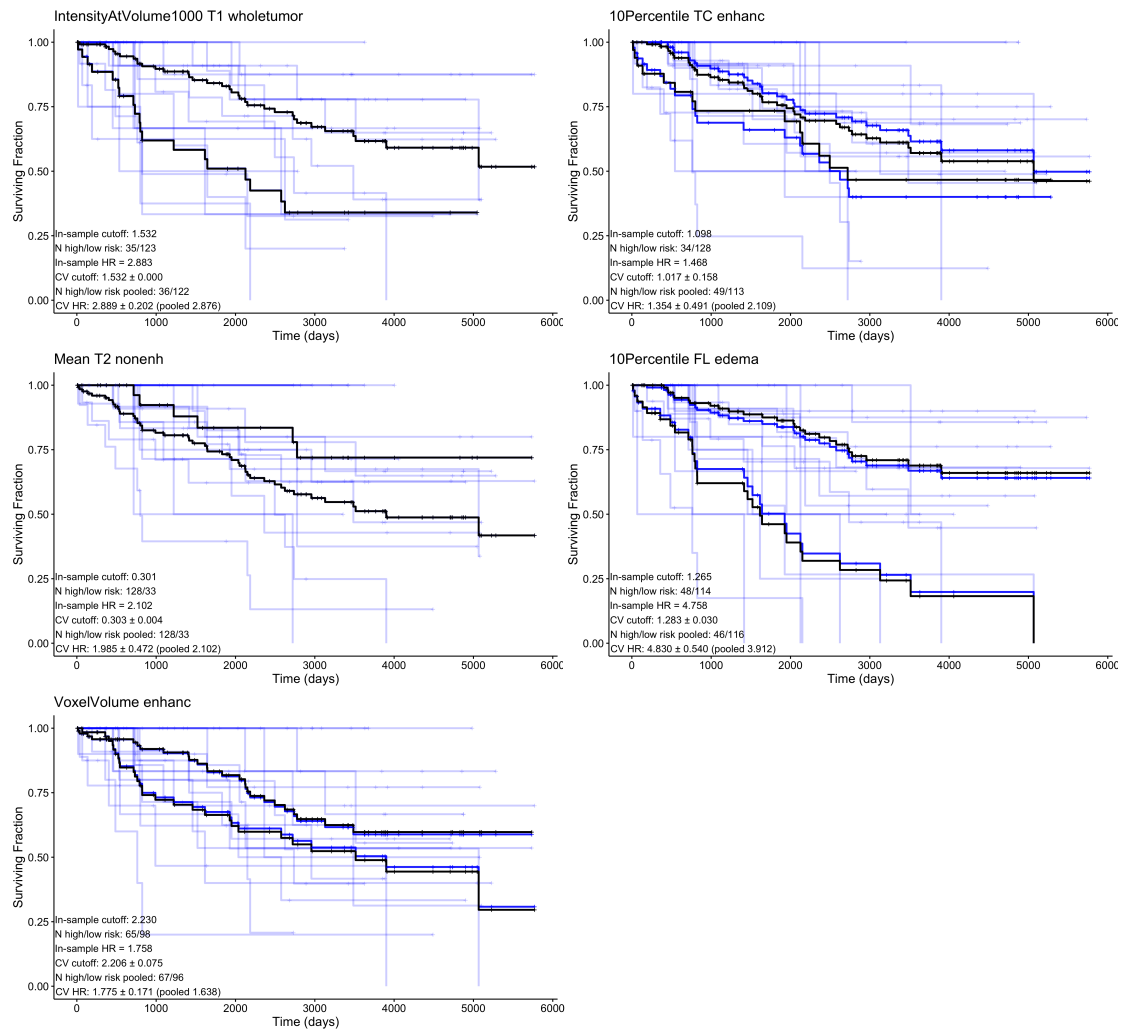


FIGURE A.12: Best extent of resection (EOR) measures for stratifying survival for WHO grade III cases for each raw image type (T1, TC, T2, FLAIR)

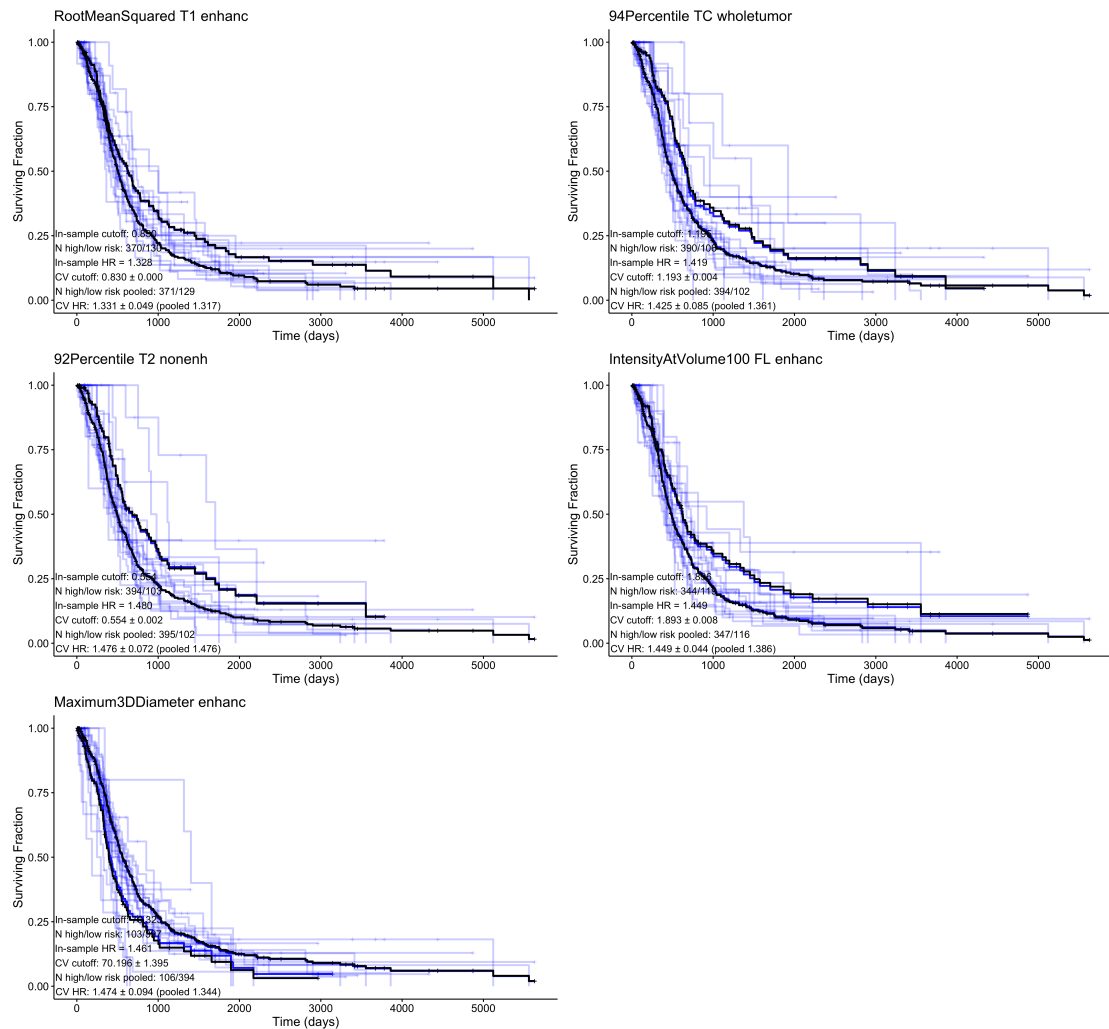


FIGURE A.13: Best extent of resection (EOR) measures for stratifying survival for WHO grade IV cases for each raw image type (T1, TC, T2, FLAIR)

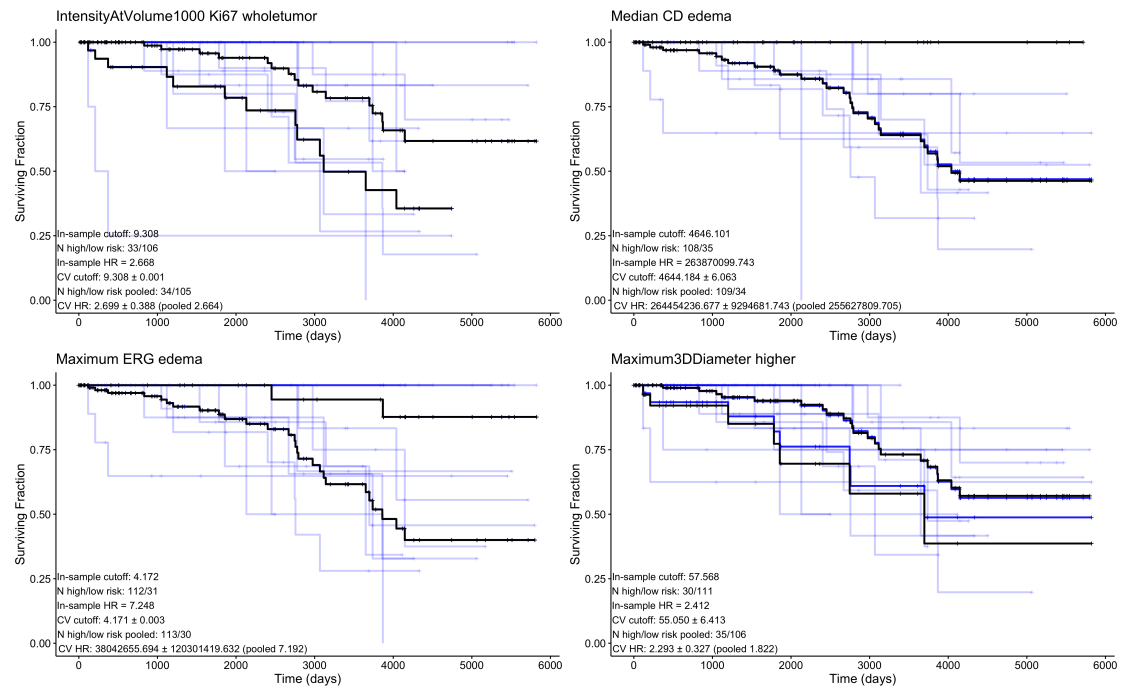


FIGURE A.14: Best extent of resection (EOR) measures for stratifying survival for WHO grade II cases for each pathology map estimate (Ki67, CD, ERG)

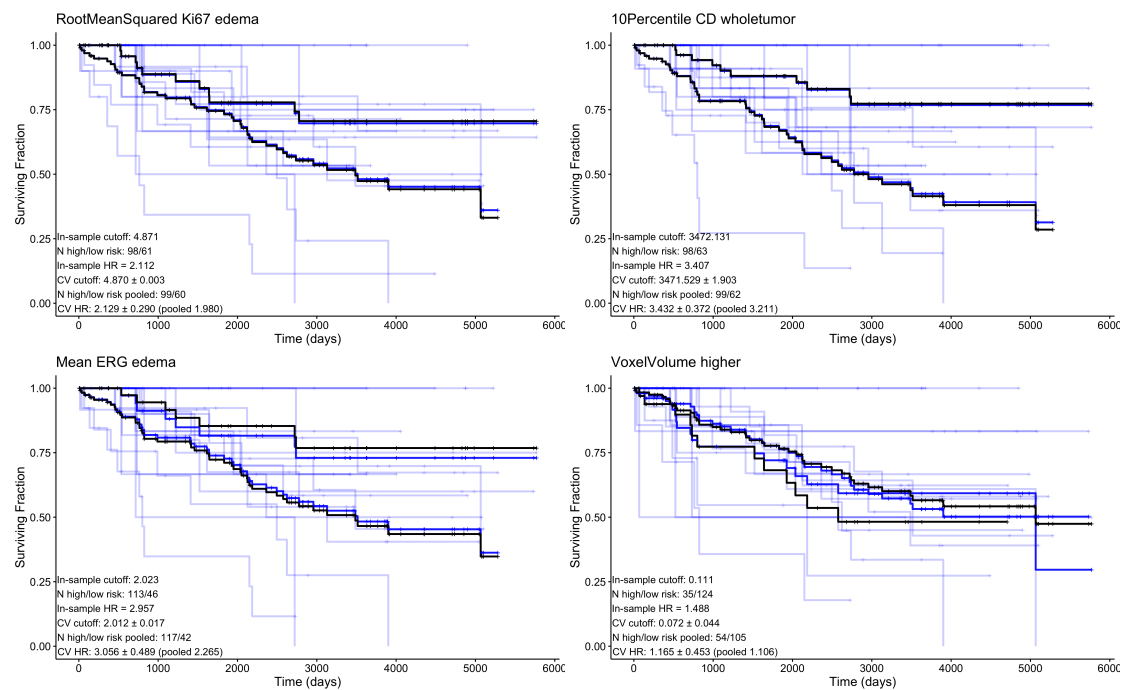


FIGURE A.15: Best extent of resection (EOR) measures for stratifying survival for WHO grade III cases for each pathology map estimate (Ki67, CD, ERG)

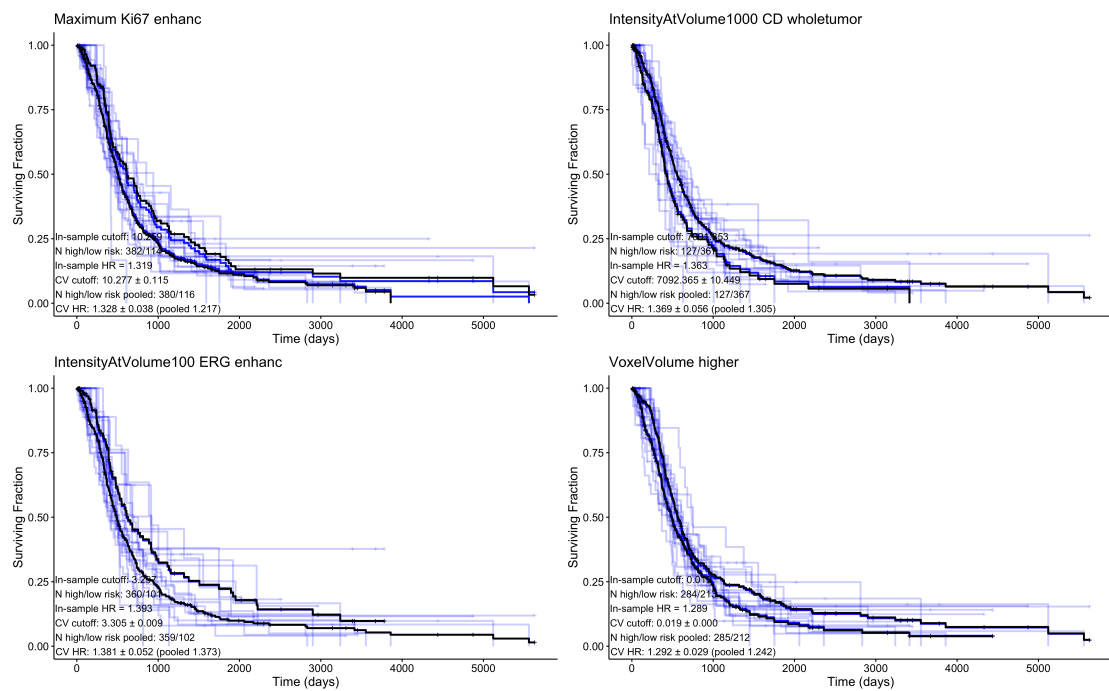


FIGURE A.16: Best extent of resection (EOR) measures for stratifying survival for WHO grade IV cases for each pathology map estimate (Ki67, CD, ERG)

A.4.4 Extent of Resection

image	region	feature	IS		Univariate			Multivariate (age+KPS)		
			C	Cut	HR	95% CI	p	HR	95% CI	p
CD	edema	EORreduc-Median	0.537	504	4.374	[0.59, 32.20]	9e-01	4.478	[0.61, 33.09]	1e-01
ERG	edema	EORreduc-98Percentile	0.585	0.174	3.134	[0.95, 10.36]	9e-01	3.069	[0.92, 10.20]	7e-02
GR	lower	EORfrac-Maximum3DDiameter	0.492	0.0704	2.520	[0.60, 10.62]	9e-01	2.485	[0.59, 10.54]	2e-01
Ki67	edema	EORfrac-98Percentile	0.595	0.0173	2.292	[1.02, 5.18]	9e-01	2.318	[1.03, 5.24]	4e-02
T1	wholetumor	EORfrac-10Percentile	0.497	0.151	1.822	[0.78, 4.28]	9e-01	2.064	[0.84, 5.04]	1e-01
FL	wholetumor	EORfrac-10Percentile	0.496	0.183	1.747	[0.77, 3.95]	9e-01	1.722	[0.76, 3.91]	2e-01
T2	edema	EORfrac-Mean	0.448	0.309	1.628	[0.56, 4.71]	9e-01	1.668	[0.57, 4.85]	3e-01
-	edema	EORfrac-Maximum3DDiameter	0.496	0.0739	1.591	[0.48, 5.29]	9e-01	1.555	[0.47, 5.19]	5e-01

TABLE A.14: Best EOR image features from each image type among patients with WHO grade(s) II in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample

image	region	feature	IS		Univariate			Multivariate (age+KPS)		
			C	Cut	HR	95% CI	p	HR	95% CI	p
CD	wholetumor	EORfrac-10Percentile	0.672	0.194	6.091	[2.20, 16.89]	3e-02	7.097	[2.52, 19.99]	2e-04
ERG	wholetumor	EORreduc-10Percentile	0.664	0.828	5.836	[2.10, 16.18]	4e-02	6.224	[2.20, 17.57]	6e-04
T2	edema	EORreduc-Median	0.605	0.12	5.467	[1.70, 17.54]	1e-01	7.491	[2.23, 25.13]	1e-03
Ki67	wholetumor	EORfrac-10Percentile	0.635	0.526	4.729	[1.88, 11.90]	4e-02	4.632	[1.82, 11.79]	1e-03
FL	enhanc	EORfrac-Median	0.575	0.321	3.526	[1.39, 8.92]	1e-01	3.250	[1.28, 8.28]	1e-02
-	nonenh	EORfrac-VoxelVolume	0.462	0.944	2.670	[1.20, 5.94]	1e-01	2.144	[0.95, 4.86]	7e-02
T1	enhanc	EORreduc-RootMeanSquared	0.566	0.262	1.871	[0.88, 3.97]	3e-01	1.947	[0.91, 4.15]	8e-02
GR	higher	EORreduc-VoxelVolume	0.570	3.99	1.703	[0.84, 3.45]	3e-01	1.345	[0.65, 2.78]	4e-01
TC	enhanc	EORfrac-95Percentile	0.534	0.124	1.501	[0.87, 2.60]	3e-01	2.139	[1.16, 3.93]	1e-02

TABLE A.15: Best EOR image features from each image type among patients with WHO grade(s) III in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample

image	region	feature	IS		Univariate			Multivariate (age+KPS)		
			C	Cut	HR	95% CI	p	HR	95% CI	p
ERG	edema	EORfrac-IntensityAtVolume100	0.541	0.0897	1.551	[1.20, 2.00]	4e-02	1.550	[1.20, 2.00]	8e-04
T2	enhanc	EORreduc-10Percentile	0.519	0.174	1.538	[1.19, 1.98]	4e-02	1.478	[1.15, 1.91]	3e-03
-	wholetumor	EORfrac-VoxelVolume	0.538	0.61	1.463	[1.12, 1.90]	8e-02	1.305	[1.00, 1.71]	5e-02
Ki67	wholetumor	EORreduc-10Percentile	0.525	0.695	1.435	[1.17, 1.77]	4e-02	1.386	[1.12, 1.71]	2e-03
FL	edema	EORfrac-98Percentile	0.527	0.107	1.423	[1.14, 1.78]	5e-02	1.385	[1.11, 1.73]	4e-03
CD	edema	EORfrac-10Percentile	0.518	0.107	1.318	[1.05, 1.66]	1e-01	1.276	[1.01, 1.61]	4e-02
GR	higher	EORreduc-Maximum3DDiameter	0.528	20.4	1.306	[1.05, 1.63]	1e-01	1.254	[1.01, 1.56]	4e-02
TC	enhanc	EORfrac-98Percentile	0.511	0.36	1.183	[0.96, 1.46]	3e-01	1.194	[0.97, 1.47]	9e-02
T1	enhanc	EORreduc-94Percentile	0.500	0.16	1.128	[0.91, 1.40]	6e-01	1.063	[0.85, 1.32]	6e-01

TABLE A.16: Best EOR image features from each image type among patients with WHO grade(s) IV in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. Uncorrected p-values are listed along with significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1). For Univariate p-values only, significance is corrected for multiple comparisons. IS = in sample

FIGURE A.17: Best extent of resection (EOR) measures for stratifying survival for WHO grade II cases for each raw image type (T1, TC, T2, FLAIR)

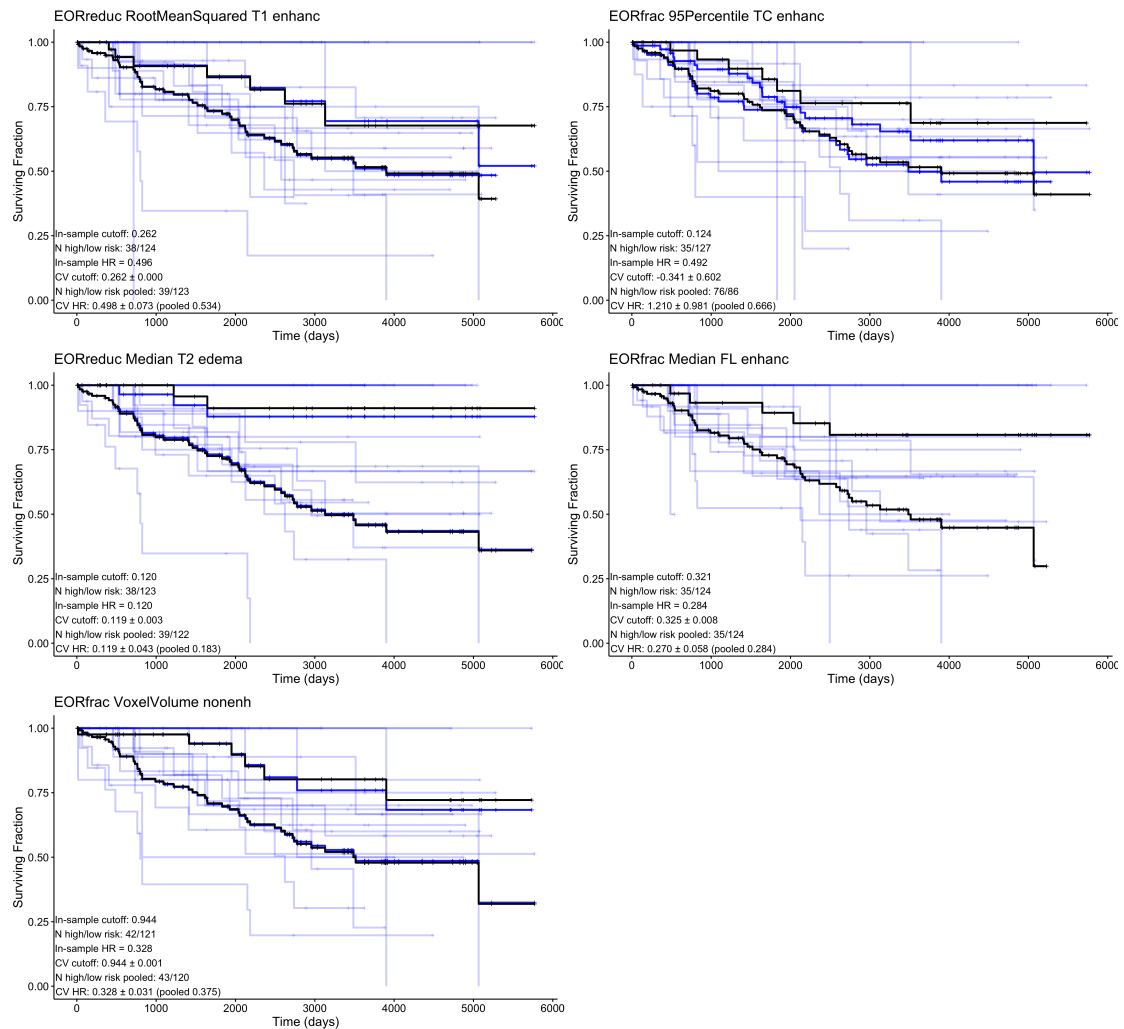


FIGURE A.18: Best extent of resection (EOR) measures for stratifying survival for WHO grade III cases for each raw image type (T1, TC, T2, FLAIR)

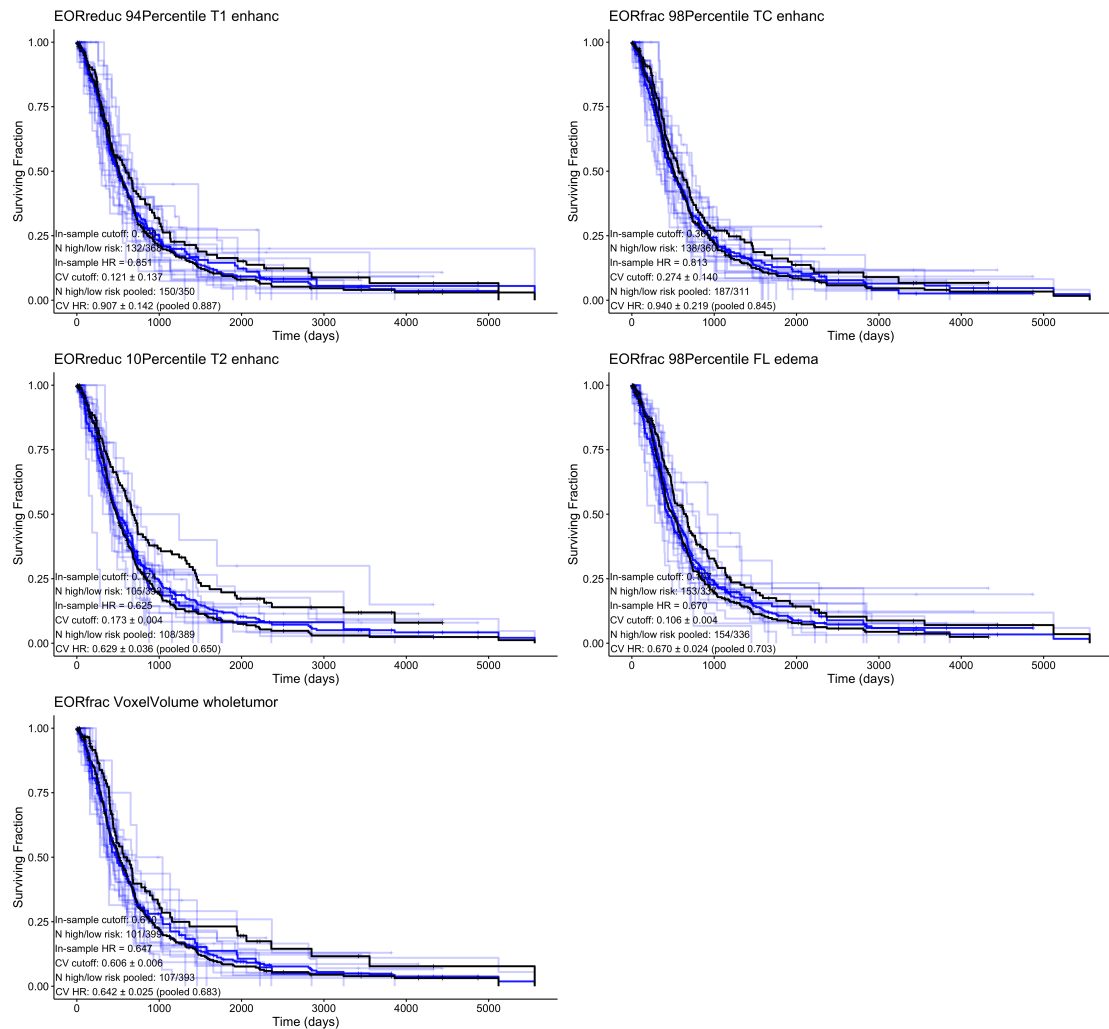


FIGURE A.19: Best extent of resection (EOR) measures for stratifying survival for WHO grade IV cases for each raw image type (T1, TC, T2, FLAIR)

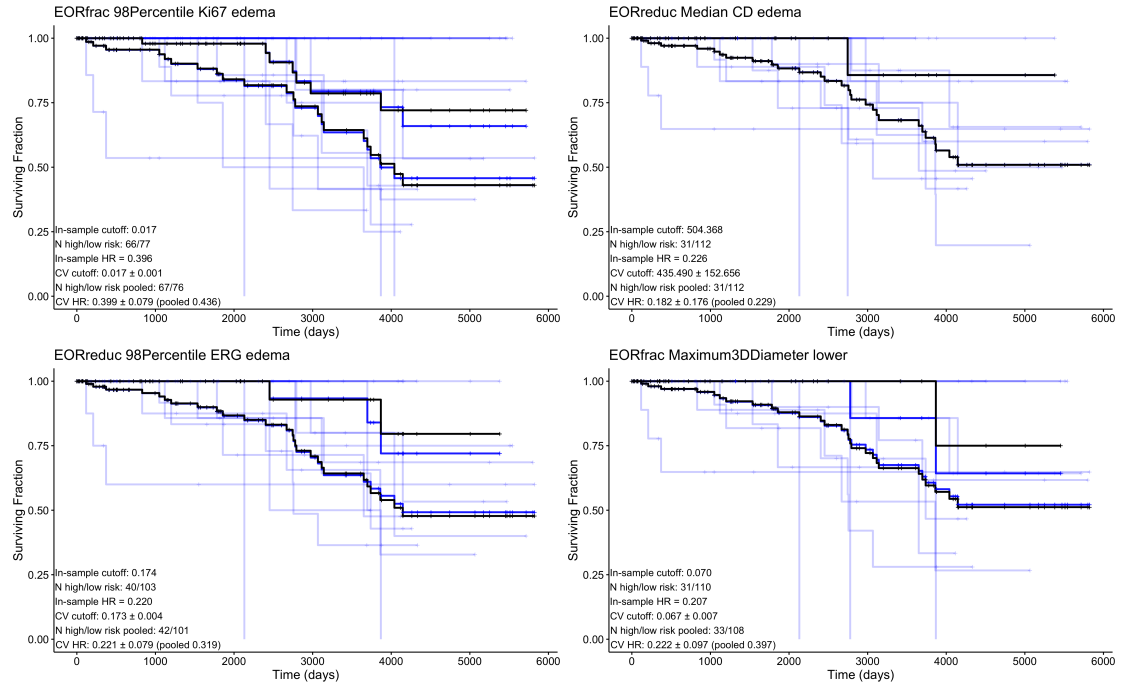


FIGURE A.20: Best extent of resection (EOR) measures for stratifying survival for WHO grade II cases for each pathology map estimate (Ki67, CD, ERG)

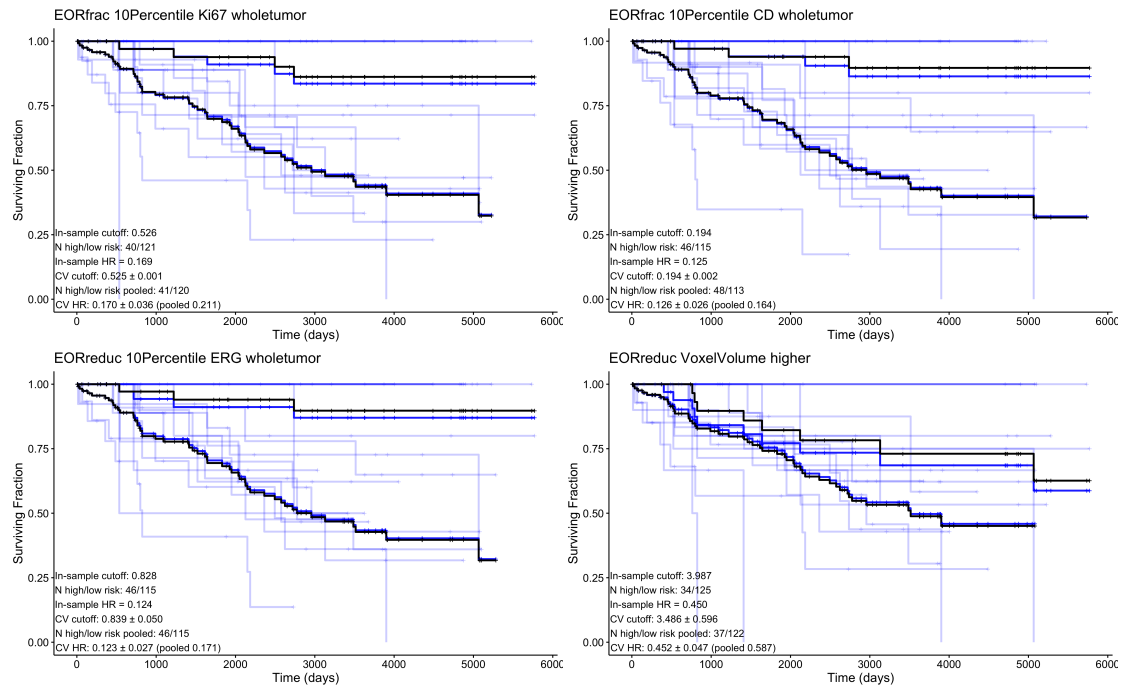


FIGURE A.21: Best extent of resection (EOR) measures for stratifying survival for WHO grade III cases for each pathology map estimate (Ki67, CD, ERG)

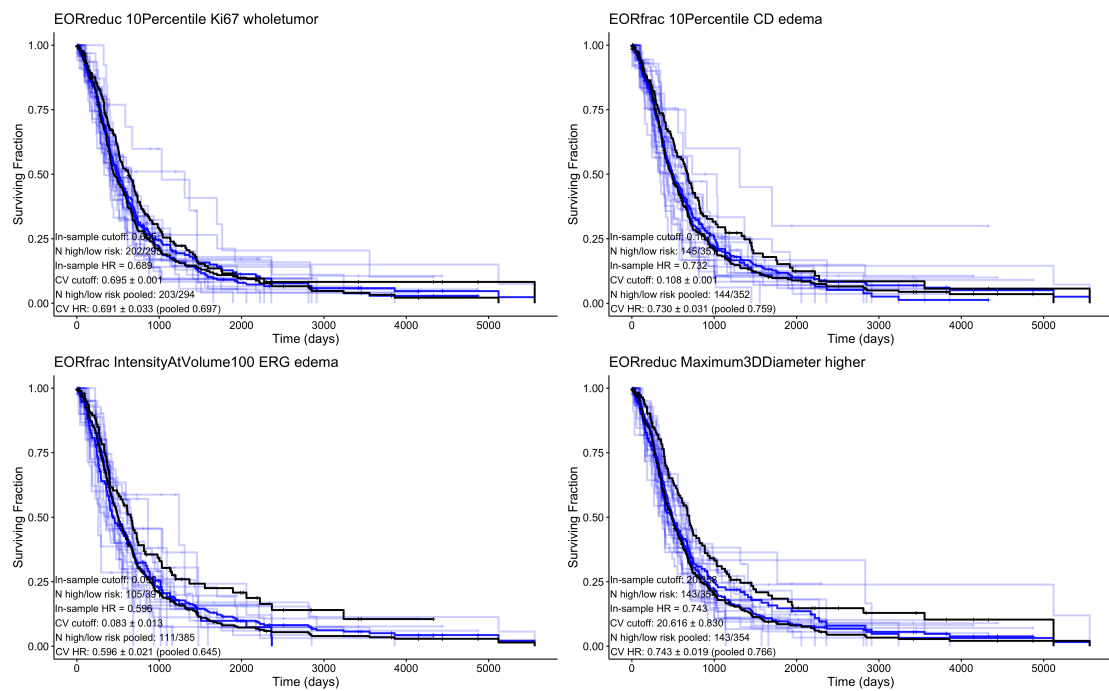


FIGURE A.22: Best extent of resection (EOR) measures for stratifying survival for WHO grade IV cases for each pathology map estimate (Ki67, CD, ERG)

A.5 Survival Analysis on Known IDH Mutation Subset

For the cases with known IDH mutation status, we redid the multivariate survival analysis including now age, KPS, grade, and IDH mutation status. The resulting univariate and multivariate hazard ratios are listed in the following tables: Table A.17 (preop), Table A.18, (postop) and Table A.19 (extent of resection). The results on postoperative generally suffered due to a lack of numbers of patients overall.

image	region	feature	IS		Univariate			Multivariate (age+KPS+IDH)			Multivariate (age+KPS+grade+IDH)		
			C	Cut	HR	95% CI	p	HR	95% CI	p	HR	95% CI	p
TC	wholetumor	IntensityAtVolume10	0.706	1.81	6.185	[3.49, 10.97]	4e-09 ***	4.574	[2.55, 8.20]	3e-07 ***	2.859	[1.55, 5.28]	8e-04 ***
CD	enhanc	IntensityAtVolume100	0.700	7.99e+03	5.648	[3.77, 8.46]	2e-15 ***	4.084	[2.67, 6.25]	1e-10 ***	2.548	[1.61, 4.02]	6e-05 ***
Ki67	enhanc	IntensityAtVolume10	0.682	13.1	3.732	[2.71, 5.15]	2e-14 ***	2.709	[1.92, 3.82]	1e-08 ***	1.833	[1.28, 2.62]	9e-04 ***
-	enhanc	VoxelVolume	0.616	0.645	3.197	[2.05, 4.99]	1e-06 ***	2.574	[1.63, 4.06]	5e-05 ***	1.759	[1.10, 2.81]	2e-02 *
T1	wholetumor	Median	0.608	0.72	3.137	[1.97, 4.98]	5e-06 ***	2.368	[1.47, 3.81]	4e-04 ***	2.395	[1.46, 3.92]	5e-04 ***
ERG	enhanc	92Percentile	0.599	3.38	2.846	[1.93, 4.20]	7e-07 ***	2.333	[1.57, 3.47]	3e-05 ***	1.870	[1.25, 2.79]	2e-03 **
FL	brainNoCSF	IntensityAtVolume100	0.617	2.44	2.462	[1.62, 3.75]	8e-05 ***	1.777	[1.15, 2.74]	9e-03 **	1.571	[1.01, 2.45]	5e-02 *
T2	edema	Maximum	0.606	1.2	2.447	[1.76, 3.39]	4e-07 ***	1.577	[1.12, 2.22]	9e-03 **	1.361	[0.97, 1.90]	7e-02 .

TABLE A.17: Best image features from each image type among patients with WHO grade(s) II III IV with known IDH mutation status in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. No local grade features had significant univariate hazard ratios on this subset. Uncorrected p-values are listed along with corrected significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1).

image	region	feature	IS		Univariate			Multivariate (age+KPS+IDH)			Multivariate (age+KPS+IDH+grade)		
			C	Cut	HR	95% CI	p	HR	95% CI	p	HR	95% CI	p
T1	nonenh	RootMeanSquared	0.587	0.721	2.353	[1.44, 3.85]	1e-02 *	2.063	[1.26, 3.39]	4e-03 **	0.697	[0.49, 1.00]	5e-02 *
FL	enhanc	91Percentile	0.577	1.8	2.038	[1.25, 3.33]	3e-02 *	1.966	[1.19, 3.24]	8e-03 **	0.729	[0.52, 1.02]	7e-02 .
-	wholetumor	VoxelVolume	0.601	50.6	1.969	[1.35, 2.87]	9e-03 **	1.233	[0.82, 1.86]	3e-01	0.846	[0.59, 1.21]	4e-01
TC	enhanc	10Percentile	0.572	0.615	1.931	[1.19, 3.12]	4e-02 *	1.759	[1.09, 2.85]	2e-02 *	1.008	[0.69, 1.47]	1e+00

TABLE A.18: Best postoperative image features from each image type among patients with WHO grade(s) II III IV in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. No local grade, T2, or synthetic pathology features had significant univariate hazard ratios on this subset. Uncorrected p-values are listed along with corrected significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1).

image	region	feature	IS		Univariate			Multivariate (age+KPS+IDH)			Multivariate (age+KPS+IDH+grade)		
			C	Cut	HR	95% CI	p	HR	95% CI	p	HR	95% CI	p
ERG	nonenh	EOReduc-IntensityAtVolume10	0.602	0.554	2.196	[1.40, 3.45]	9e-03 **	1.643	[1.03, 2.63]	4e-02 *	1.654	[1.01, 2.70]	4e-02 *
-	nonenh	EORfrac-VoxelVolume	0.354	0.699	2.021	[1.40, 2.91]	3e-03 **	1.450	[0.99, 2.12]	6e-02 .	1.255	[0.85, 1.84]	2e-01
CD	wholetumor	EORfrac-10Percentile	0.549	0.213	2.007	[1.25, 3.21]	4e-02 *	1.422	[0.88, 2.30]	2e-01	1.609	[0.99, 2.63]	6e-02 .
T2	nonenh	EOReduc-10Percentile	0.545	0.249	1.732	[1.19, 2.51]	4e-02 *	1.577	[1.08, 2.29]	2e-02 *	1.529	[1.05, 2.23]	3e-02 *

TABLE A.19: Best EOR image features from each image type among patients with WHO grade(s) II III IV in terms of univariate hazard ratio. For each feature, the concordance index (C) is listed alongside hazard ratios between high-risk and low-risk groups using an optimized threshold. The same threshold is used for the univariate and multivariate models. No local grade, Ki67, CD, T1, TC, or FLAIR features had significant univariate hazard ratios on this subset. Uncorrected p-values are listed along with corrected significance levels (0 *** 0.001 ** 0.01 * 0.05 . 0.1).

A.5.1 Validation on BraTS 2018 Data

Among the 163 brain tumor segmentation challenge (BraTS) cases with known survival, 147 of them passed QA with acceptable data quality. We expanded the features from Table A.10 to include the best feature from each image type and attempted to validate the prognostic stratification on the BraTS cases. For each feature, we tuned an optimal cutoff on all the WHO grade IV historical cases then applied the cutoff to the BraTS cohort to identify high risk and low-risk groups. The results are shown in Table A.20.

Surprisingly, none of the raw image features showed significant survival differences between groups for the BraTS data. For features based on predicted pathology, one feature did pass the validation: 98th percentile CD over the whole brain minus CSF. The in-sample hazard ratio for the historical cases was 1.43 and the validated hazard ratio among BraTS cases was 1.55 which is very similar. The cutoff of 7.07E3 almost evenly divided the BraTS cases as well with 74 cases above the threshold and 89 below., although this was not the case for the historical data (147 cases above, 581 below cutoff). Overall, this suggests 98th percentile whole-brain CD may be a strong prognostic biomarker for grade IV gliomas.

feature	image	region	Cutoff	Training HR	Validation HR	95% CI	p
RootMeanSquared	TC	wholetumor	1.067e+00	1.847	1.190	[0.832, 1.702]	0.340
10Percentile	T1	wholetumor	4.765e-01	1.693	1.182	[0.800, 1.746]	0.400
Maximum3DDiameter	highres	enhanc	4.051e+01	1.680	1.199	[0.803, 1.790]	0.375
10Percentile	FL	nonenh	9.716e-01	1.587	0.962	[0.572, 1.620]	0.885
98Percentile	CD	brainNoCSF	7.065e+03	1.432	1.547	[1.110, 2.155]	0.010 *
92Percentile	Ki67	brainNoCSF	8.187e+00	1.726	1.423	[0.843, 2.400]	0.187
IntensityAtVolume100	ERG	enhanc	3.810e+00	1.629	1.392	[0.897, 2.160]	0.140
Maximum	T2	enhanc	9.247e-01	1.311	1.112	[0.787, 1.570]	0.547

TABLE A.20: Survival stratification using the in-sample cutoffs for raw image features trained on the historical data and applied to the BraTS 2018 data. Results are for univariate analysis. Validation HRs marked * are significantly different from 1 (log-rank $p < 0.05$)

A.6 Mathematical Methods

A.6.1 Proportional Hazard Model and Partial Likelihood Fitting

Here, a toy example for the Proportional Hazards model. The Cox model optimized the partial likelihood for failure given the observations. Suppose we have a set of survival times for n patients $X_i : i = 1, \dots, n$ and for each X_i there is an associated δ_i where

$\delta_i = 1$ if patient i has an event and $\delta_i = 0$ if patient i is censored at time X_i . Denote the set of covariate(s) of interest as \mathbf{Z}_i . The proportional hazard model assumed the form of the hazard ratio

$$h(X_i|\mathbf{Z}_i) = h_0(t) \exp(\beta^T \mathbf{Z}_i)$$

Denote the n (distinct) ordered event times $t_1 < \dots < t_n$. For any time t we have $R(t)$ patients at risk of failure within the time interval: $R(t) = \{i : X_i \geq t\}$. This is called the risk set. Since the Cox model is non-parametric, its parameters only depend on the order of the failures and not their explicit time values. The discrete failures give rise to a set of time intervals where only one failure occurs. We seek to find the parameters that maximize the failures occurring in that order. For every failure time X_j (note: the number of failures $\leq n$ due to censoring) there is a single time interval $[t_j, t_{j+1}]$ that contains the failure and no other failures (since we do not allow ties). So, we can compute the likelihood of observing this one failure in that interval:

$$L_j(\beta) = P(\text{patient } j \text{ fails} \mid \text{One failure among } R(X_j)) \quad (\text{A.1})$$

$$= \frac{h(X_j|\mathbf{Z}_j)}{\sum_{k \in R(X_j)} h(X_k|\mathbf{Z}_k)} \quad (\text{A.2})$$

The full partial likelihood is the product over the likelihoods for all uncensored patients. To use only uncensored patients, use the censoring value $\delta_i \in \{0, 1\}$.

$$L(\beta) = \prod_{j=1}^n \left(\frac{h_0(X_j) \exp(\beta^T \mathbf{Z}_j)}{\sum_{k \in R(X_j)} h_0(X_k) \exp(\beta^T \mathbf{Z}_k)} \right)^{\delta_j} \quad (\text{A.3})$$

$$= \prod_{j=1}^n \left(\frac{\exp(\beta^T \mathbf{Z}_j)}{\sum_{k \in R(X_j)} \exp(\beta^T \mathbf{Z}_k)} \right)^{\delta_j} \quad (\text{A.4})$$

The maximum partial likelihood can be obtained by maximizing the log partial likelihood to fit coefficients to the proportional hazard model.

$$\frac{\partial \log L(\beta)}{\partial \beta} = \frac{\partial}{\partial \beta} \sum_{j=1}^n \delta_j \left[\beta^T \mathbf{Z}_j - \log \left(\sum_{k=1}^n I(X_k \geq X_j) \exp(\beta^T \mathbf{Z}_k) \right) \right] \quad (\text{A.5})$$

$$= \sum_{j=1}^n \delta_j \left[\mathbf{Z}_j - \frac{\sum_{k=1}^n I(X_k \geq X_j) \mathbf{Z}_k \exp(\beta^T \mathbf{Z}_k)}{\sum_{k=1}^n I(X_k \geq X_j) \exp(\beta^T \mathbf{Z}_k)} \right] \quad (\text{A.6})$$

Where $I(X_k \geq X_j)$ is an indicator to sum over only the terms in the risk set $R(X_j)$.

The information matrix is then given by:

$$\begin{aligned} \frac{\partial^2 \log L(\beta)}{\partial \beta^2} &= \sum_{j=1}^n \delta_j \left[0 + \frac{(\sum_{k=1}^n I(X_k \geq X_j) \exp(\beta^T \mathbf{Z}_k)) (\sum_{k=1}^n I(X_k \geq X_j) \mathbf{Z}_k \exp(\beta^T \mathbf{Z}_k))}{(\sum_{k=1}^n I(X_k \geq X_j) \exp(\beta^T \mathbf{Z}_k))^2} \right. \\ &\quad \left. - \frac{(\sum_{k=1}^n I(X_k \geq X_j) \mathbf{Z}_k \exp(\beta^T \mathbf{Z}_k)) (\sum_{k=1}^n I(X_k \geq X_j) \mathbf{Z}_k \exp(\beta^T \mathbf{Z}_k))}{(\sum_{k=1}^n I(X_k \geq X_j) \exp(\beta^T \mathbf{Z}_k))^2} \right] \end{aligned} \quad (\text{A.7})$$

$$= \sum_{j=1}^n \delta_j \left[\frac{\sum_{k=1}^n I(X_k \geq X_j) |\mathbf{Z}_k|^2 \exp(\beta^T \mathbf{Z}_k)}{\sum_{k=1}^n I(X_k \geq X_j) \exp(\beta^T \mathbf{Z}_k)} - \frac{(\sum_{k=1}^n I(X_k \geq X_j) \mathbf{Z}_k \exp(\beta^T \mathbf{Z}_k))^2}{(\sum_{k=1}^n I(X_k \geq X_j) \exp(\beta^T \mathbf{Z}_k))^2} \right] \quad (\text{A.8})$$

$$= \sum_{j=1}^n \delta_j \left[\frac{(\sum_{k=1}^n w_k |\mathbf{Z}_k|^2) \sum_{k=1}^n w_k - (\sum_{k=1}^n w_k \mathbf{Z}_k)^2}{(\sum_{k=1}^n w_k)^2} \right] \quad (\text{A.9})$$

Where $w_k = I(X_k \geq X_j) \exp(\beta^T \mathbf{Z}_k)$ is non-negative. Using Equation A.9 we can see the information matrix is semi-positive definite and therefore $\log L(\beta)$ is a concave function.

Bibliography

- [1] Eileen Lüders, Helmuth Steinmetz, and Lutz Jäncke. Brain size and grey matter volume in the healthy human brain. *Neuroreport*, 13(17):2371–2374, 2002.
- [2] M M Oken, R H Creech, D C Tormey, J Horton, T E Davis, E T McFadden, and P P Carbone. Toxicity and response criteria of the Eastern Cooperative Oncology Group. *American journal of clinical oncology*, 5(6):649–655, dec 1982.
- [3] Quinn T Ostrom, Nirav Patil, Gino Cioffi, Kristin Waite, Carol Kruchko, and Jill S Barnholtz-Sloan. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2013–2017. *Neuro-Oncology*, 22(Supplement_1):iv1–iv96, oct 2020.
- [4] Roger Stupp, Warren P Mason, Martin J van den Bent, Michael Weller, Barbara Fisher, Martin J B Taphoorn, Karl Belanger, Alba A Brandes, Christine Marosi, Ulrich Bogdahn, Jürgen Curschmann, Robert C Janzer, Samuel K Ludwin, Thierry Gorlia, Anouk Allgeier, Denis Lacombe, J Gregory Cairncross, Elizabeth Eisenhauer, and René O. Mirimanoff, "the European Organisation for Research and Treatment of Cancer Brain Tumor and Radiotherapy Groups", and "the National Cancer Institute of Canada Clinical Trials Group". Radiotherapy plus Concomitant and Adjuvant Temozolomide for Glioblastoma. *N. Engl. J. Med.*, 352(10):987–996, 2005.
- [5] Yoshikazu Okamoto, Pier Luigi Di Patre, Christoph Burkhard, Sonja Horstmann, Benjamin Jourde, Michael Fahey, Danielle Schüler, Nicole M. Probst-Hensch, M. Gazi Yasargil, Yasuhiro Yonekawa, Urs M. Lütolf, Paul Kleihues, and Hiroko Ohgaki. Population-based study on incidence, survival rates, and genetic alterations of low-grade diffuse astrocytomas and oligodendrogliomas. *Acta Neuropathologica*, 108(1):49–56, 2004.

- [6] Hiroko Ohgaki and Paul Kleihues. Epidemiology and etiology of gliomas. *Acta Neuropathologica*, 109(1):93–108, 2005.
- [7] Peter C. Burger and Paul Kleihues. Cytologic composition of the untreated glioblastoma with implications for evaluation of needle biopsies. *Cancer*, 63(10):2014–2023, may 1989.
- [8] Dinorah Friedmann-Morvinski. Glioblastoma Heterogeneity and Cancer Cell Plasticity. *Critical Reviews in Oncogenesis*, 19(5):327–336, 2014.
- [9] Diane J Aum, David H Kim, Thomas L Beaumont, Eric C Leuthardt, Gavin P Dunn, and Albert H Kim. Molecular and cellular heterogeneity: the hallmark of glioblastoma. *Neurosurgical focus*, 37(6):E11, 2014.
- [10] Deborah A. Forst, Brian V. Nahed, Jay S. Loeffler, and Tracy T. Batchelor. Low-Grade Gliomas. *The Oncologist*, 19(4):403–413, apr 2014.
- [11] Akio Soeda, Akira Hara, Takahiro Kunisada, Shin-ichi Yoshimura, Toru Iwama, and Deric M. Park. The Evidence of Glioblastoma Heterogeneity. *Scientific Reports*, 5(1):7979, jul 2015.
- [12] Mona Meyer, Jüri Reimand, Xiaoyang Lan, Renee Head, Xueming Zhu, Michelle Kushida, Jane Bayani, Jessica C. Pressey, Anath C. Lionel, Ian D. Clarke, Michael Cusimano, Jeremy A. Squire, Stephen W. Scherer, Mark Bernstein, Melanie A. Woodin, Gary D. Bader, and Peter B. Dirks. Single cell-derived clonal analysis of human glioblastoma links functional and genomic heterogeneity. *Proceedings of the National Academy of Sciences*, 112(3):851–856, jan 2015.
- [13] C. Bouvier-Labit, O. Chinott, C. Ochit, D. Gambarelli, H. Dufourt, and D. Figarella-Branger. Prognostic significance of Ki67, p53 and epidermal growth factor receptor immunostaining in human glioblastomas. *Neuropathology and Applied Neurobiology*, 24(5):381–388, 1998.
- [14] Michel Lacroix, Dima Abi-Said, Daryl R Fourney, Ziya L Gokaslan, Weiming Shi, Franco DeMonte, Frederick F Lang, Ian E McCutcheon, Samuel J Hassenbusch, Eric Holland, Kenneth Hess, Christopher Michael, Daniel Miller, and Raymond Sawaya. A multivariate analysis of 416 patients with glioblastoma multiforme: prognosis, extent of resection, and survival. *Journal of Neurosurgery*, 95(2):190–198, aug 2001.

- [15] Yan Michael Li, Dima Suki, Kenneth Hess, and Raymond Sawaya. The influence of maximum safe resection of glioblastoma on survival in 1229 patients: Can we do better than gross-total resection? *Journal of Neurosurgery*, 124(4):977–988, apr 2016.
- [16] Kaisorn L. Chaichana, Ignacio Jusue-Torres, Rodrigo Navarro-Ramirez, Shaan M. Raza, Maria Pascual-Gallego, Aly Ibrahim, Marta Hernandez-Hermann, Luis Gomez, Xiaobu Ye, Jon D. Weingart, Alessandro Olivi, Jaishri Blakeley, Gary L. Gallia, Michael Lim, Henry Brem, and Alfredo Quinones-Hinojosa. Establishing percent resection and residual volume thresholds affecting survival and recurrence for patients with newly diagnosed intracranial glioblastoma. *Neuro-Oncology*, 16(1):113–122, 2014.
- [17] Matthew M Grabowski, Pablo F Recinos, Amy S Nowacki, Jason L Schroeder, Lilyana Angelov, Gene H Barnett, and Michael A Vogelbaum. Residual tumor volume versus extent of resection: predictors of survival after surgery for glioblastoma. *J Neurosurg*, 121(121):1115–1123, 2014.
- [18] Brian J Gill, David J Pisapia, Hani R Malone, Hannah Goldstein, Liang Lei, Adam Sonabend, Jonathan Yun, Jorge Samanamud, Jennifer S Sims, Matei Banu, Athanassios Dovas, Andrew F Teich, Sameer A Sheth, Guy M McKhann, Michael B Sisti, Jeffrey N Bruce, Peter A Sims, and Peter Canoll. MRI-localized biopsies reveal subtype-specific differences in molecular and cellular composition at the margins of glioblastoma. *Proceedings of the National Academy of Sciences of the United States of America*, 111(34):12550–12555, 2014.
- [19] Christian Hartmann, Bettina Hentschel, Wolfgang Wick, David Capper, Jörg Felsberg, Matthias Simon, Manfred Westphal, Gabriele Schackert, Richard Meyermann, Torsten Pietsch, Guido Reifenberger, Michael Weller, Markus Loeffler, and Andreas von Deimling. Patients with IDH1 wild type anaplastic astrocytomas exhibit worse prognosis than IDH1-mutated glioblastomas, and IDH1 mutation status accounts for the unfavorable prognostic effect of higher age: implications for classification of gliomas. *Acta Neuropathologica*, 120(6):707–718, 2010.
- [20] David N. Louis, Arie Perry, Guido Reifenberger, Andreas von Deimling, Dominique Figarella-Branger, Webster K. Cavenee, Hiroko Ohgaki, Otmar D. Wiestler, Paul

- Kleihues, and David W. Ellison. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathologica*, 131(6):803–820, 2016.
- [21] David N Louis, Hiroko Ohgaki, Otmar D Wiestler, Webster K Cavenee, Peter C Burger, Anne Jouvett, Bernd W Scheithauer, and Paul Kleihues. The 2007 WHO Classification of Tumours of the Central Nervous System. *Acta Neuropathologica*, 114(2):97–109, 2007.
- [22] Jason T Huse. Establishing a Robust Molecular Taxonomy for Diffuse Gliomas of Adulthood. *Surgical Pathology Clinics*, 9(3):379–390, 2016.
- [23] Guido Reifenberger, Hans-Georg Wirsching, Christiane B Knobbe-Thomsen, and Michael Weller. Advances in the molecular genetics of gliomas — implications for classification and therapy. *Nature Reviews Clinical Oncology*, 14(7):434–452, 2017.
- [24] Jason T Huse, Eli L Diamond, Lu Wang, and Marc K Rosenblum. Mixed glioma with molecular features of composite oligodendroglioma and astrocytoma: a true “oligoastrocytoma”? *Acta neuropathologica*, 129(1):151–153, 2015.
- [25] David N Louis, Caterina Giannini, David Capper, Werner Paulus, Dominique Figarella-Branger, M Beatriz Lopes, Tracy T Batchelor, J Gregory Cairncross, Martin van den Bent, Wolfgang Wick, and Pieter Wesseling. cIMPACT-NOW update 2: diagnostic clarifications for diffuse midline glioma, H3 K27M-mutant and diffuse astrocytoma/anaplastic astrocytoma, IDH-mutant. *Acta Neuropathologica*, 135(4):639–642, 2018.
- [26] Daniel J Brat, Kenneth Aldape, Howard Colman, Eric C Holland, David N Louis, Robert B Jenkins, B K Kleinschmidt-DeMasters, Arie Perry, Guido Reifenberger, Roger Stupp, Andreas von Deimling, and Michael Weller. cIMPACT-NOW update 3: recommended diagnostic criteria for “Diffuse astrocytic glioma, IDH-wildtype, with molecular features of glioblastoma, WHO grade IV”. *Acta Neuropathologica*, 136(5):805–810, 2018.
- [27] Daniel J Brat, Kenneth Aldape, Howard Colman, Dominique Figarella-Branger, Gregory N Fuller, Caterina Giannini, Eric C Holland, Robert B Jenkins, Bette

- Kleinschmidt-DeMasters, Takashi Komori, Johan M Kros, David N Louis, Catriona McLean, Arie Perry, Guido Reifenberger, Chitra Sarkar, Roger Stupp, Martin J van den Bent, Andreas von Deimling, and Michael Weller. cIMPACT-NOW update 5: recommended grading criteria and terminologies for IDH-mutant astrocytomas. *Acta Neuropathologica*, 139(3):603–608, 2020.
- [28] Andrea Sottoriva, Inmaculada Spiteri, Sara G M Piccirillo, Anestis Touloumis, V Peter Collins, John C Marioni, Christina Curtis, Colin Watts, and Simon Tavaré. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences*, 110(10):4009–4014, 2013.
- [29] Charles Swanton. Intratumor Heterogeneity: Evolution through Space and Time. *Cancer Research*, 72(19):4875, 2012.
- [30] Zoé Pedetour-Braccini, Fanny Burel-Vandenbos, Catherine Gozé, Coralie Roger, Audrey Bazin, Valérie Costes-Martineau, Hugues Duffau, and Valérie Rigau. Microfoci of malignant progression in diffuse low-grade gliomas: towards the creation of an intermediate grade in glioma classification? *Virchows Arch.*, 466(4):433–444, 2015.
- [31] Robert J Jackson, Gregory N Fuller, Dima Abi-Said, Frederick F Lang, Ziya L Gokaslan, Wei Ming Shi, David M Wildrick, and Raymond Sawaya. Limitations of stereotactic biopsy in the initial management of gliomas. *Neuro-Oncology*, 3(3):193–200, 2001.
- [32] Javier E. Villanueva-Meyer, Marc C. Mabray, and Soonmee Cha. Current clinical brain tumor imaging. *Clinical Neurosurgery*, 81(3):397–415, 2017.
- [33] J. N. Scott, P. M.A. Brasher, R. J. Sevick, N. B. Rewcastle, and P. A. Forsyth. How often are nonenhancing supratentorial gliomas malignant? A population study. *Neurology*, 59(6):947–949, sep 2002.
- [34] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-Andre Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Herve Delingette,

- Cagatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftikharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, Jose Antonio Mariz, Raphael Meier, Sergio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, oct 2015.
- [35] Hugo J W L Aerts, Emmanuel Rios Velazquez, Ralph T H Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M Rietbergen, C René Leemans, Andre Dekker, John Quackenbush, Robert J Gillies, and Philippe Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, (5):4006, 2014.
- [36] Joost J.M. Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77(21):e104–e107, 2017.
- [37] Timothy Van Meter, Catherine Dumur, Naiel Hafez, Carleton Garrett, Helen Fillmore, and William C Broaddus. Microarray Analysis of MRI-defined Tissue Samples in Glioblastoma Reveals Differences in Regional Expression of Therapeutic Targets. *Diagn Mol Pathol*, 15:195–205, 2006.
- [38] Ramon F Barajas Jr, Joanna J Phillips, Rupa Parvataneni, Annette Molinaro, Emma Essock-Burns, Gabriela Bourne, Andrew T Parsa, Manish K Aghi, Michael W McDermott, Mitchel S Berger, Soonmee Cha, Susan M Chang, and Sarah J Nelson. Regional variation in histopathologic features of tumor specimens from treatment-naïve glioblastoma correlates with anatomic and physiologic MR Imaging. *Neuro-Oncology*, 14(7):942–954, 2012.

- [39] Christian Ewelt, Frank W Floeth, Jörg Felsberg, Hans J Steiger, Michael Sabel, Karl-Josef Langen, Gabriele Stoffels, and Walter Stummer. Finding the anaplastic focus in diffuse gliomas: The value of Gd-DTPA enhanced MRI, FET-PET, and intraoperative, ALA-derived tissue fluorescence. *Clinical Neurology and Neurosurgery*, 113(7):541–547, 2011.
- [40] Leon Weninger, Christoph Haarbuerger, and Dorit Merhof. Robustness of Radiomics for Survival Prediction of Brain Tumor Patients Depending on Resection Status, 2019.
- [41] Asgeir S. Jakola, Kristin S. Myrnes, Roar Kloster, Sverre H. Torp, Sigurd Lindal, Geirmund Unsgård, and Ole Solheim. Comparison of a strategy favoring early surgical resection vs a strategy favoring watchful waiting in low-grade gliomas. *JAMA - Journal of the American Medical Association*, 308(18):1881–1888, 2012.
- [42] S J Price, R Jena, N G Burnet, P J Hutchinson, A F Dean, A Pena, J D Pickard, T A Carpenter, and J H Gillard. Improved Delineation of Glioma Margins and Regions of Infiltration with the Use of Diffusion Tensor Imaging: An Image-guided Biopsy Study. *Am. J. Neuroradiol.*, 27(9):1969–1974, 2006.
- [43] Hamed Akbari, Luke Macyszyn, Xiao Da, Michel Bilello, Ronald L Wolf, Maria Martinez-Lage, George Biros, Michelle Alonso-Basanta, Donald M O’Rourke, and Christos Davatzikos. Imaging Surrogates of Infiltration Obtained Via Multiparametric Imaging Pattern Analysis Predict Subsequent Location of Recurrence of Glioblastoma. *Neurosurg.*, 78(4):572–580, 2016.
- [44] R Jena, S J Price, C Baker, S J Jefferies, J D Pickard, J H Gillard, and N G Burnet. Diffusion Tensor Imaging: Possible Implications for Radiotherapy Treatment Planning of Patients with High-grade Glioma. *Clinical Oncology*, 17(8):581–590, 2005.
- [45] A Chaddad, M Toews, C Desrosiers, and T Niazi. Deep Radiomic Analysis Based on Modeling Information Flow in Convolutional Neural Networks. *IEEE Access*, 7:97242–97252, 2019.
- [46] C E Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, jul 1948.

- [47] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.
- [48] Mark Ringner. What is principal component analysis? *Nat. Biotechnol.*, 26(3):303–304, 2008.
- [49] Evan D H Gates, Jonathan S Lin, Jeffrey S Weinberg, Jackson Hamilton, Sujit S Prabhu, John D Hazle, Gregory N Fuller, Veera Baladandayuthapani, David Fuentes, and Dawid Schellingerhout. Guiding the first biopsy in glioma patients using estimated Ki-67 maps derived from MRI: conventional versus advanced imaging. *Neuro-Oncology*, 21(4):527–536, mar 2019.
- [50] E.D.H. Gates, J.S. Lin, J.S. Weinberg, S.S. Prabhu, J. Hamilton, J.D. Hazle, G.N. Fuller, V. Baladandayuthapani, D.T. Fuentes, and D. Schellingerhout. Imaging-Based Algorithm for the Local Grading of Glioma. *American Journal of Neuroradiology*, feb 2020.
- [51] E D H Gates, J S Weinberg, S S Prabhu, J S Lin, J Hamilton, J D Hazle, G N Fuller, V Baladandayuthapani, D T Fuentes, and D Schellingerhout. Estimating Local Cellular Density in Glioma Using MR Imaging Data. *American Journal of Neuroradiology*, nov 2020.
- [52] Jonathan Lin. *Predicting Tissue Characteristics in Brain Tumors Using Radiological-Pathological Correlations*. Doctor of philosophy, Rice University, 2016.
- [53] Linda Douw, Martin Klein, Selene SAA Fagel, Josje van den Heuvel, Martin JB Taphoorn, Neil K. Aaronson, Tjeerd J. Postma, W. Peter Vandertop, Jacob J. Mooij, Rudolf H. Boerman, Guus N. Beute, Jasper D. Sluimer, Ben J. Slotman, Jaap C. Reijneveld, and Jan J. Heimans. Cognitive and radiological effects of radiotherapy in patients with low-grade glioma: long-term follow-up. *The Lancet Neurology*, 8(9):810–818, 2009.
- [54] O Surma-aho, M Niemelä, J Vilkki, M Kouri, A Brander, O Salonen, A Paetau, M Kallio, LicPhil Pyykkönen J., and J Jääskeläinen. Adverse long-term effects of brain radiotherapy in adult low-grade glioma patients. *Neurology*, 56(10):1285 LP – 1290, may 2001.

- [55] Ramon Francisco Jr. Barajas and Soonmee Cha. Benefits of dynamic susceptibility-weighted contrast-enhanced perfusion MRI for glioma diagnosis and therapy. *CNS Oncol.*, 3(6):407–419, 2014.
- [56] G Çoban, S Mohan, F Kural, S Wang, D M O’Rourke, and H Poptani. Prognostic Value of Dynamic Susceptibility Contrast-Enhanced and Diffusion-Weighted MR Imaging in Patients with Glioblastomas. *American Journal of Neuroradiology*, 36(7):1247 LP – 1252, jul 2015.
- [57] Meng Law, Stanley Yang, James S Babb, Edmond A Knopp, John G Golfinos, David Zagzag, and Glyn Johnson. Comparison of cerebral blood volume and vascular permeability from dynamic susceptibility contrast-enhanced perfusion MR imaging with glioma grade. *American Journal of Neuroradiology*, 25(5):746–755, 2004.
- [58] Natalie R Boonzaier, Sara G M Piccirillo, Colin Watts, and Stephen J Price. Assessing and monitoring intratumor heterogeneity in glioblastoma: how far has multimodal imaging come? *CNS Oncol.*, 4(6):399–410, 2015.
- [59] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, 26(6):1045–1057, dec 2013.
- [60] Fred W. Prior, Ken Clark, Paul Commean, John Freymann, Carl Jaffe, Justin Kirby, Stephen Moore, Kirk Smith, Lawrence Tarbox, Bruce Vendt, and Guillermo Marquez. TCIA: An information resource to enable open science. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2013.
- [61] Stephen J Price, H A L Green, A F Dean, J Joseph, P J Hutchinson, and J H Gillard. Correlation of MR Relative Cerebral Blood Volume Measurements with Cellular Density and Proliferation in High-Grade Gliomas: An Image-Guided Biopsy Study. *Am. J. Neuroradiol.*, 32(3):501–506, 2011.

- [62] P. D. Chang, H. R. Malone, S. G. Bowden, D. S. Chow, B. J.A. Gill, T. H. Ung, J. Samanamud, Z. K. Englander, A. M. Sonabend, S. A. Sheth, G. M. McKhann, M. B. Sisti, L. H. Schwartz, A. Lignelli, J. Grinband, J. N. Bruce, and P. Canoll. A multiparametric model for mapping cellularity in glioblastoma using radiographically localized biopsies. *American Journal of Neuroradiology*, 38(5):890–898, 2017.
- [63] Adam Autry, Joanna J. Phillips, Stojan Maleschlijski, Ritu Roy, Annette M. Molinaro, Susan M. Chang, Soonmee Cha, Janine M. Lupo, and Sarah J. Nelson. Characterization of Metabolic, Diffusion, and Perfusion Properties in GBM: Contrast-Enhancing versus Non-Enhancing Tumor. *Translational Oncology*, 10(6):895–903, dec 2017.
- [64] Paul S Tofts, Gunnar Brix, David L Buckley, Jeffrey L Evelhoch, Elizabeth Henderson, Michael V Knopp, Henrik B W Larsson, Ting-Yim Lee, Nina a Mayr, Geoffrey J M Parker, Ruediger E Port, June Taylor, and Robert M Weisskoff. Estimating Kinetic Parameters From Dynamic Contrast-Enhanced T1-Weighted MRI of a Diffusible Tracer: Standardized Quantities and Symbols. *J Magn Reson Imag*, 10(July):223–232, 1999.
- [65] Huijun Chen, Feiyu Li, Xihai Zhao, Chun Yuan, Brian Rutt, and William S Kerwin. Extended graphical model for analysis of dynamic contrast-enhanced MRI. *Magnetic Resonance in Medicine*, 66(3):868–878, sep 2011.
- [66] Paul S Tofts. Modeling Tracer Kinetics in Dynamic Gd-DTPA MR Imaging. *J. Magn. Reson. Imaging*, 7(1):91–101, 1997.
- [67] Leif Østergaard, Robert M Weisskoff, David A Chesler, Carsten Gyldensted, and Bruce R Rosen. High Resolution Measurement of Cerebral Blood Flow Using Intravascular Tracer Bolus Passages. Part I: Mathematical Approach and Statistical Analysis. *Magn. Reson. Med.*, 36(5):715–725, 1996.
- [68] Leif Østergaard, Alma Gregory Sorenson, Kenneth K Kwong, Robert M Weiskoff, Carsten Gyldensted, and Bruce R Rosen. High Resolution Measurement of Cerebral Blood Flow Using Intravascular Tracer Bolus Passages. Part II: Experimental Comparison and Preliminary Results. *Magn. Reson. Med.*, 36(5):726–736, 1996.
- [69] Leif Østergaard. Principles of Cerebral Perfusion Imaging by Bolus Tracking. *J. Magn. Reson. Imaging*, 22(6):710–717, 2005.

- [70] Lori Arlinghaus and Thomas E Yankeelov. Diffusion-Weighted MRI. In T E Yankeelov, D R Pickens, and R R Price, editors, *Quantitative MRI in Cancer*, pages 81–98. Taylor & Francis, Boca Raton, FL, 2011.
- [71] Denis Le Bihan, Eric Breton, Denis Lallemand, Philippe Grenier, Emmanuel Cabanis, and Maurice Laval-Jeantet. MR Imaging of Intravoxel Incoherent Motions: Application to Diffusion and Perfusion in Neurologic Disorders. *Radiology*, 161(2):401–407, 1986.
- [72] Peter J Basser, James Mattiello, and Denis LeBihan. MR Diffusion Tensor Spectroscopy and Imaging. *Biophys. J.*, 66(1):259–267, 1994.
- [73] Peter J Basser and Carlo Pierpaoli. Microstructural and Physiological Features of Tissues Elucidated by Quantitative-Diffusion-Tensor MRI. *J. Magn. Reson. B*, 111:209–219, 1996.
- [74] Stephen M Smith. Fast Robust Automated Brain Extraction. *Hum. Brain Mapp.*, 17(3):143–155, 2002.
- [75] Brian B Avants, Nicholas J Tustison, Michael Stauffer, Gang Song, Baohua Wu, and James C Gee. The Insight ToolKit image registration framework. *Front. Neuroinf.*, 8:1–13, 2014.
- [76] Brian B Avants, Nicholas J Tustison, Gang Song, Philip A Cook, Arno Klein, and James C Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, 54(3):2033–2044, 2011.
- [77] Jonathan S Lin, David Fuentes, Adam Chandler, Sujit Prabhu, Jeffrey S Weinberg, Veera Baladandayuthapani, John D Hazle, and Dawid Schellingerhout. Performance Assessment for Brain Magnetic Resonance Imaging Registration Methods. *Am. J. Neuroradiol.*, 38(5), 2017.
- [78] Kelvin K Leung, Matthew J Clarkson, Jonathan W Bartlett, Shona Clegg, Clifford R Jack Jr, Michael W Weiner, Nick C Fox, Sébastien Ourselin, and the Alzheimer’s Disease Neuroimaging Initiative. Robust atrophy rate measurement in Alzheimer’s disease using multi-site serial MRI: Tissue-specific intensity normalization and parameter selection. *NeuroImage*, 50(2):516–523, 2010.

- [79] Florian Ringel, Dominik Ingerl, Stephanie Ott, and Bernhard Meyer. VARIOGUIDE: A NEW FRAMELESS IMAGE-GUIDED STEREOTACTIC SYSTEM—ACCURACY STUDY AND CLINICAL ASSESSMENT. *Operative Neurosurgery*, 64(suppl_5):ons365–ons373, may 2009.
- [80] Ondrej Bradac, Anna Steklacova, Katerina Nebrenska, Jiri Vrana, Patricia de Lacy, and Vladimir Benes. Accuracy of VarioGuide Frameless Stereotactic System Against Frame-Based Stereotaxy: Prospective, Randomized, Single-Center Study. *World Neurosurgery*, 104:831–840, aug 2017.
- [81] Ian J. Gerard, Marta Kersten-Oertel, Kevin Petrecca, Denis Sirhan, Jeffery A. Hall, and D. Louis Collins. Brain shift in neuronavigation of brain tumors: A review. *Medical Image Analysis*, 35:403–420, 2017.
- [82] P C Burger, T Shibata, and P Kleihues. The use of the monoclonal antibody Ki-67 in the identification of proliferating cells: application to surgical neuropathology. *The American journal of surgical pathology*, 10(9):611–617, 1986.
- [83] Charles S Parkins, John L Darling, Steven S Gill, Thomas Revesz, and David G Thomas. Cell proliferation in serial biopsies through human malignant brain tumours: measurement using Ki67 antibody labelling. *Br. J. Neurosurg.*, 5(3):289–298, 1991.
- [84] Anne J. Skjulsvik, Jørgen N. Mørk, Morten O. Torp, and Sverre H. Torp. Ki-67/MIB-1 immunostaining in a cohort of human gliomas. *International Journal of Clinical and Experimental Pathology*, 7(12):8905–8910, 2014.
- [85] Wen-Jie Chen, De-Shen He, Rui-Xue Tang, Fang-Hui Ren, and Gang Chen. Ki-67 is a valuable prognostic factor in gliomas: evidence from a systematic review and meta-analysis. *Asian Pacific Journal of Cancer Prevention*, 16(2):411–420, 2015.
- [86] Hai Yan, D Williams Parsons, Genglin Jin, Roger Mclendon, B Ahmed Rasheed, Weishi Yuan, Ivan Kos, Ines Batinic-Haberle, Siân Jones, Gregory J Riggins, Henry Friedman, Allan Friedman, David Reardon, James Herndon, Kenneth W Kinzler, Victor E Velculescu, Bert Vogelstein, and Darell D Bigner. IDH1 and IDH2 Mutations in Gliomas. *Radiation Oncology Neuro-Oncology N Engl J Med*, 360:765–73, 2009.

- [87] Takuya Watanabe, Sumihito Nobusawa, Paul Kleihues, and Hiroko Ohgaki. IDH1 Mutations Are Early Events in the Development of Astrocytomas and Oligodendrogliomas. *The American Journal of Pathology*, 174(4):1149–1153, 2009.
- [88] Matthew A Haber, Amir Iranmahboob, Cheddi Thomas, Mengling Liu, Amanda Najjar, and David Zagzag. ERG is a novel and reliable marker for endothelial cells in central nervous system tumors. *Clin. Neuropathol.*, 34(3):117–127, 2015.
- [89] Thomas Roetzer, Konrad Leskovaar, Nadine Peter, Julia Furtner, Martina Muck, Marco Augustin, Antonia Lichtenegger, Martha Nowosielski, Johannes A. Hainfellner, Bernhard Baumann, and Adelheid Woehrer. Evaluating cellularity and structural connectivity on whole brain slides using a custom-made digital pathology pipeline. *Journal of Neuroscience Methods*, 311(July 2018):215–221, jan 2019.
- [90] M Abercrombie. Estimation of nuclear population from microtome sections. *The Anatomical Record*, 94(2):239–247, feb 1946.
- [91] John C Hedreen. What was wrong with the Abercrombie and empirical cell counting methods? A review. *The Anatomical Record*, 250(3):373–380, mar 1998.
- [92] Eva Gagy, Bernadett Kormos, Karla J Castellanos, Klara Valyi-Nagy, Dennis Korneff, Patrizia LoPresti, Randy Woltjer, and Tibor Valyi-Nagy. Decreased Oligodendrocyte Nuclear Diameter in Alzheimer’s Disease and Lewy Body Dementia. *Brain Pathology*, 22(6):803–810, nov 2012.
- [93] Rasmus Krarup Sigaard, Majken Kjær, and Bente Pakkenberg. Development of the Cell Population in the Brain White Matter of Young Children. *Cerebral Cortex*, 26(1):89–95, aug 2014.
- [94] Jessica M. Snyder, Catherine E. Hagan, Brad Bolon, and C. Dirk Keene. Nervous System. In *Comparative Anatomy and Histology*, chapter 20, pages 403–444. Academic Press, 2 edition, jan 2018.
- [95] ”R Core Team”. R: A Language and Environment for Statistical Computing, 2013.
- [96] R Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, 2(0):1137–1143, 1995.

- [97] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, (9):1263–1284, 2008.
- [98] Max Kuhn. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5), 2008.
- [99] William S Noble. How does multiple testing correction work. *Nat. Biotechnol.*, 27(12):1135–1137, 2009.
- [100] Leo Breiman. Random Forests. *Mach. Learn.*, 45(1):5–32, 2001.
- [101] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.
- [102] Elizabeth Huynh, Ahmed Hosny, Christian Guthier, Danielle S Bitterman, Steven F Petit, Daphne A Haas-Kogan, Benjamin Kann, Hugo J W L Aerts, and Raymond H Mak. Artificial intelligence in radiation oncology. *Nature Reviews Clinical Oncology*, aug 2020.
- [103] Spyridon Bakas, Mauricio Reyes, Et Int., and Bjoern Menze. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation , Progression Assessment , and Overall Survival Prediction in the BRATS Challenge. *arXiv preprint arXiv:1811.02629*, 2018.
- [104] Meng Law, Stanley Yang, Hao Wang, James S Babb, Glyn Johnson, Soonmee Cha, Edmond A Knopp, and David Zagzag. Glioma grading: sensitivity, specificity, and predictive values of perfusion MR imaging and proton MR spectroscopic imaging compared with conventional MR imaging. *AJNR American journal of neuroradiology*, 24(10):1989–1998, 2003.
- [105] Christopher R Durst, Prashant Raghavan, Mark E Shaffrey, David Schiff, M Beatriz Lopes, Jason P Sheehan, Nicholas J Tustison, James T Patrie, Wenjun Xin, W Jeff Elias, Kenneth C Liu, Greg A Helm, A Cupino, and Max Wintermark. Multimodal MR imaging model to predict tumor infiltration in patients with gliomas. *Neuroradiology*, 56(2):107–115, 2014.
- [106] N Sadeghi, N D’Haene, C Decaestecker, M Levivier, T Metens, C Maris, D Wikler, D Baleriaux, I Salmon, and S Goldman. Apparent Diffusion Coefficient and Cerebral Blood Volume in Brain Gliomas: Relation to Tumor Cell Density and

- Tumor Microvessel Density Based on Stereotactic Biopsies. *Am. J. Neuroradiol.*, 29(3):476–482, 2008.
- [107] Ming Zhao, Li Li Guo, Ning Huang, Qiong Wu, Li Zhou, Hui Zhao, Jing Zhang, and Kuang Fu. Quantitative analysis of permeability for glioma grading using dynamic contrast-enhanced magnetic resonance imaging. *Oncology Letters*, 14(5):5418–5426, 2017.
- [108] Xin Zhang, Lin-Feng Yan, Yu-Chuan Hu, Gang Li, Yang Yang, Yu Han, Ying-Zhi Sun, Zhi-Cheng Liu, Qiang Tian, Zi-Yang Han, Le-De Liu, Bin-Quan Hu, Zi-Yu Qiu, Wen Wang, and Guang-Bin Cui. Optimizing a machine learning based glioma grading system using multi-parametric MRI histogram and texture features. *Oncotarget*, 8(29):47816–47830, 2017.
- [109] Ken Chang, Harrison X. Bai, Hao Zhou, Chang Su, Wenya Linda Bi, Ena Agbodza, Vasileios K. Kavouridis, Joekey T. Senders, Alessandro Boaro, Andrew Beers, Biqi Zhang, Alexandra Capellini, Weihua Liao, Qin Shen, Xuejun Li, Bo Xiao, Jane Cryan, Shakti Ramkissoon, Lori Ramkissoon, Keith Ligon, Patrick Y. Wen, Ranjit S. Bindra, John Woo, Omar Arnaout, Elizabeth R. Gerstner, Paul J. Zhang, Bruce R. Rosen, Li Yang, Raymond Y. Huang, and Jayashree Kalpathy-Cramer. Residual convolutional neural network for the determination of IDH status in low- and high-grade gliomas from mr imaging. *Clinical Cancer Research*, 24(5):1073–1081, 2018.
- [110] Falgun H Chokshi, Adam E Flanders, Luciano M Prevedello, and Curtis P Langlotz. Fostering a Healthy AI Ecosystem for Radiology: Conclusions of the 2018 RSNA Summit on AI in Radiology. *Radiology: Artificial Intelligence*, 1(2):190021, mar 2019.
- [111] Thomas R Mackie, Edward F Jackson, and Maryellen Giger. Opportunities and challenges to utilization of quantitative imaging: Report of the AAPM practical big data workshop. *Medical Physics*, 45(10):e820–e828, oct 2018.
- [112] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4:170117, sep 2017.

- [113] Brian B Avants, Nicholas J Tustison, Jue Wu, Philip A Cook, and James C Gee. An Open Source Multivariate Framework for n-Tissue Segmentation with Evaluation on Public Data. *Neuroinformatics*, 9(4):381–400, 2011.
- [114] Snehashis Roy, John A. Butman, and Dzung L. Pham. Robust skull stripping using multiple MR image contrasts insensitive to pathology. *NeuroImage*, 146(November 2016):132–147, 2017.
- [115] Andriy Myronenko. 3D MRI Brain Tumor Segmentation Using Autoencoder Regularization. In Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 311–320, Cham, 2019. Springer International Publishing.
- [116] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, jun 2010.
- [117] Evan Gates, J Gregory Pauloski, Dawid Schellingerhout, and David Fuentes. Glioma Segmentation and a Simple Accurate Model for Overall Survival Prediction BT - Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. pages 476–484, Cham, 2019. Springer International Publishing.
- [118] Rhian Jenkins, Nick C Fox, Alex M Rossor, Richard J Harvey, and Martin N Rossor. Intracranial Volume and Alzheimer Disease: Evidence Against the Cerebral Reserve Hypothesis. *Archives of Neurology*, 57(2):220–224, feb 2000.
- [119] D D Blatter, E D Bigler, S D Gale, S C Johnson, C V Anderson, B M Burnett, N Parker, S Kurth, and S D Horn. Quantitative volumetric analysis of brain MR: normative database spanning 5 decades of life. *American Journal of Neuroradiology*, 16(2):241 LP – 251, feb 1995.
- [120] Romane Gauriau, Christopher Bridge, Lina Chen, Felipe Kitamura, Neil A Tenenholtz, John E Kirsch, Katherine P Andriole, Mark H Michalski, and Bernardo C Bizzo. Using DICOM Metadata for Radiological Image Series Categorization: a Feasibility Study on Large Clinical Brain MRI Datasets. *Journal of Digital Imaging*, 33(3):747–762, 2020.

- [121] P. Kalavathi and V. B Surya Prasath. Methods on Skull Stripping of MRI Head Scan Images-a Review. *Journal of Digital Imaging*, 29(3):365—379, 2016.
- [122] Tim J Kruser, Walter R Bosch, Shahed N Badiyan, Joseph A Bovi, Amol J Ghia, Michelle M Kim, Abhishek A Solanki, Sean Sachdev, Christina Tsien, Tony J C Wang, Minesh P Mehta, and Kevin P McMullen. NRG brain tumor specialists consensus guidelines for glioblastoma contouring. *Journal of Neuro-Oncology*, 143(1):157–166, 2019.
- [123] Stefanie Bette, Johannes Kaesmacher, Thomas Huber, Claire Delbridge, Florian Ringel, Tobias Boeckh-Behrens, Bernhard Meyer, Claus Zimmer, Jan S Kirschke, and Jens Gempt. Value of Early Postoperative FLAIR Volume Dynamic in Glioma with No or Minimal Enhancement. *World Neurosurgery*, 91:548–559.e1, 2016.
- [124] Andrej Pala, Christine Brand, Thomas Kapapa, Michal Hlavac, Ralph König, Bernd Schmitz, Christian Rainer Wirtz, and Jan Coburger. The Value of Intraoperative and Early Postoperative Magnetic Resonance Imaging in Low-Grade Glioma Surgery: A Retrospective Study. *World Neurosurgery*, 93:191–197, 2016.
- [125] Moritz Scherer, Christine Jungk, Michael Götz, Philipp Kickingeder, David Reuss, Martin Bendszus, Klaus Maier-Hein, and Andreas Unterberg. Early postoperative delineation of residual tumor after low-grade glioma resection by probabilistic quantification of diffusion-weighted imaging. *Journal of Neurosurgery*, 130(6):2016–2024, jun 2019.
- [126] Sinan M. K. Belhawi, Friso W. A. Hoefnagels, Johannes C. Baaijen, Esther Sanchez Aliaga, Jaap C. Reijneveld, Jan J. Heimans, Frederik Barkhof, W. Peter Vandertop, and Philip C. De Witt Hamer. Early postoperative MRI overestimates residual tumour after resection of gliomas with no or minimal enhancement. *European Radiology*, 21(7):1526–1534, jul 2011.
- [127] Oscar Esteban, Daniel Birman, Marie Schaer, Oluwasanmi O Koyejo, Russell A Poldrack, and Krzysztof J Gorgolewski. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLOS ONE*, 12(9):e0184661, sep 2017.
- [128] Amir Reza Sadri, Andrew Janowczyk, Ren Zhou, Ruchika Verma, Niha Beig, Jacob Antunes, Anant Madabhushi, Pallavi Tiwari, and Satish E Viswanath. Technical

Note: MRQy — An open-source tool for quality control of MR imaging data. *Medical Physics*, 47(12):6029–6038, dec 2020.

- [129] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.
- [130] C C Schag, R L Heinrich, and P A Ganz. Karnofsky performance status revisited: reliability, validity, and guidelines. *Journal of Clinical Oncology*, 2(3):187–193, mar 1984.
- [131] Jeffrey D Rudie, Andreas M Rauschecker, R Nick Bryan, Christos Davatzikos, and Suyash Mohan. Emerging Applications of Artificial Intelligence in Neuro-Oncology. *Radiology*, 290(3):607–618, jan 2019.
- [132] Wenya Linda Bi, Ahmed Hosny, Matthew B Schabath, Maryellen L Giger, Nicolai J Birkbak, Alireza Mehrtash, Tavis Allison, Omar Arnaout, Christopher Abbosh, Ian F Dunn, Raymond H Mak, Rulla M Tamimi, Clare M Tempany, Charles Swanton, Udo Hoffmann, Lawrence H Schwartz, Robert J Gillies, Raymond Y Huang, and Hugo J W L Aerts. Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA: A Cancer Journal for Clinicians*, 69(2):127–157, mar 2019.
- [133] E Sala, E Mema, Y Himoto, H Veeraraghavan, J D Brenton, A Snyder, B Weigelt, and H A Vargas. Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. 2017.
- [134] Hairui Sun, Lianhu Yin, Showwei Li, Song Han, Guangrong Song, Ning Liu, and Changxiang Yan. Prognostic significance of IDH mutation in adult low-grade gliomas: a meta-analysis. *Journal of Neuro-Oncology*, 113(2):277–284, jun 2013.
- [135] Severina Leu, Stefanie von Felten, Stephan Frank, Erik Vassella, Istvan Vajtai, Elisabeth Taylor, Marianne Schulz, Gregor Hutter, Jürgen Hench, Philippe Schucht, Jean-Louis Boulay, and Luigi Mariani. IDH/MGMT-driven molecular classification of low-grade glioma is a strong predictor for long-term survival. *Neuro-Oncology*, 15(4):469–479, apr 2013.
- [136] Christian Hartmann, Jochen Meyer, Jörg Balss, David Capper, Wolf Mueller, Arne Christians, Jörg Felsberg, Marietta Wolter, Christian Mawrin, Wolfgang Wick,

- Michael Weller, Christel Herold-Mende, Andreas Unterberg, Judith W M Jeuken, Peter Wesseling, Guido Reifenberger, and Andreas von Deimling. Type and frequency of IDH1 and IDH2 mutations are related to astrocytic and oligodendroglial differentiation and age: a study of 1,010 diffuse gliomas. *Acta Neuropathologica*, 118(4):469–474, 2009.
- [137] Chase Robinson and B K Kleinschmidt-DeMasters. IDH1-Mutation in Diffuse Gliomas in Persons Age 55 Years and Over. *Journal of Neuropathology & Experimental Neurology*, 76(2):151–154, feb 2017.
- [138] Valeria Barresi, Albino Eccher, Michele Simbolo, Rekha Cappellini, Giuseppe K. Ricciardi, Francesca Calabria, Marco Cancedda, Renzo Mazzarotto, Bruno Bonetti, Giampietro Pinna, Francesco Sala, Claudio Ghimenton, and Aldo Scarpa. Diffuse gliomas in patients aged 55 years or over: A suggestion for IDH mutation testing. *Neuropathology : official journal of the Japanese Society of Neuropathology*, 40(1):68–74, feb 2020.
- [139] Annette M Molinaro, Shawn Hervey-Jumper, Ramin A Morshed, Jacob Young, Seunggu J Han, Pranathi Chunduru, Yalan Zhang, Joanna J Phillips, Anny Shai, Marisa Lafontaine, Jason Crane, Ankush Chandra, Patrick Flanigan, Arman Jahangiri, Gino Cioffi, Quinn Ostrom, John E Anderson, Chaitra Badve, Jill Barnholtz-Sloan, Andrew E Sloan, Bradley J Erickson, Paul A Decker, Matthew L Kosel, Daniel LaChance, Jeanette Eckel-Passow, Robert Jenkins, Javier Villanueva-Meyer, Terri Rice, Margaret Wrensch, John K Wiencke, Nancy Ann Oberheim Bush, Jennie Taylor, Nicholas Butowski, Michael Prados, Jennifer Clarke, Susan Chang, Edward Chang, Manish Aghi, Philip Theodosopoulos, Michael McDermott, and Mitchel S Berger. Association of Maximal Extent of Resection of Contrast-Enhanced and Non-Contrast-Enhanced Tumor With Survival Within Molecular Subgroups of Patients With Newly Diagnosed Glioblastoma. *JAMA Oncology*, 6(4):495–503, apr 2020.
- [140] Cecil Jr. Hastings, Frederick Mosteller, John W Tukey, and Charles P Winsor. Low Moments for Small Samples: A Comparative Study of Order Statistics. *The Annals of Mathematical Statistics*, 18(3):413–426, sep 1947.
- [141] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society*, 34(2):187–220, 1972.

- [142] Bernard Rosner. *Fundamentals of Biostatistics*. Cengage Learning, Boston, MA, 7 edition, 2010.
- [143] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, 57(1):289–300, 1995.
- [144] Richard M Simon, Jyothi Subramanian, Ming-Chung Li, and Supriya Menezes. Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Briefings in bioinformatics*, 12(3):203–214, may 2011.
- [145] Judah Folkman. Role of angiogenesis in tumor growth and metastasis. *Seminars in Oncology*, 29(6, Supplement 16):15–18, 2002.
- [146] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. Radiomics: Images Are More Than Pictures, They Are Data. *Radiology*, 278(2):563–577, 2016.
- [147] Sohi Bae, Yoon Seong Choi, Sung Soo Ahn, Jong Hee Chang, Seok-Gu Kang, Eui Hyun Kim, Se Hoon Kim, and Seung-Koo Lee. Radiomic MRI Phenotyping of Glioblastoma: Improving Survival Prediction. *Radiology*, 289(3):797–806, oct 2018.
- [148] Rajan Jain and Yvonne W Lui. How Far Are We from Using Radiomics Assessment of Gliomas in Clinical Practice? *Radiology*, 289(3):807–808, oct 2018.
- [149] Jiangwei Lao, Yinsheng Chen, Zhi-Cheng Li, Qihua Li, Ji Zhang, Jing Liu, and Guangtao Zhai. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. *Scientific Reports*, 7(1):10353, 2017.
- [150] Yiming Li, Zenghui Qian, Kaibin Xu, Kai Wang, Xing Fan, Shaowu Li, Xing Liu, Yinyan Wang, and Tao Jiang. Radiomic features predict Ki-67 expression level and survival in lower grade gliomas. *Journal of Neuro-Oncology*, 135(2):317–324, 2017.
- [151] Abraham Noorbakhsh, Jessica A Tang, Logan P Marcus, Brandon McCutcheon, David D Gonda, Craig S Schallhorn, Mark A Talamini, David C Chang, Bob S Carter, and Clark C Chen. Gross-total resection outcomes in an elderly population

- with glioblastoma: a SEER-based analysis. *Journal of Neurosurgery*, 120(1):31–39, jan 2014.
- [152] Timothy J Brown, Matthew C Brennan, Michael Li, Ephraim W Church, Nicholas J Brandmeir, Kevin L Rakszawski, Akshal S Patel, Elias B Rizk, Dima Suki, Raymond Sawaya, and Michael Glantz. Association of the Extent of Resection With Survival in Glioblastoma: A Systematic Review and Meta-analysis. *JAMA Oncology*, 2(11):1460–1469, nov 2016.
- [153] Audrey G Sinclair and Daniel J Scoffings. Imaging of the Post-operative Cranium. *RadioGraphics*, 30(2):461–482, mar 2010.
- [154] Adrian Celaya, Jonas Actor, Rajarajeswari Muthusivarajan, Evan Gates, Caroline Chung, Dawid Schellingerhout, Beatrice Riviere, and David Fuentes. PocketNet: A Smaller Neural Network for 3D Medical Image Segmentation. *arXiv preprint arXiv:2104.10745*, 2021.
- [155] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.
- [156] Paul A Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C Gee, and Guido Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage*, 31(3):1116–1128, 2006.
- [157] Xenia Fave, Lifei Zhang, Jinzhong Yang, Dennis Mackin, Peter Balter, Daniel Gomez, David Followill, Aaron Kyle Jones, Francesco Stingo, Zhongxing Liao, Radhe Mohan, and Laurence Court. Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. *Scientific Reports*, 7(1):588, 2017.
- [158] Geewon Lee, So Hyeon Bak, Ho Yun Lee, Joon Young Choi, Hyunjin Park, Seung-Hak Lee, Yoshiharu Ohno, Mizuki Nishino, Edwin J R van Beek, and Kyung Soo Lee. Measurement Variability in Treatment Response Determination for Non-Small Cell Lung Cancer: Improvements Using Radiomics. *Journal of Thoracic Imaging*, 34(2), 2019.

- [159] Sheng-Xiang Rao, Doenja M J Lambregts, Roald S Schnerr, Rianne C J Beckers, Monique Maas, Fabrizio Albarello, Robert G Riedl, Cornelis H C Dejong, Milou H Martens, Luc A Heijnen, Walter H Backes, Geerard L Beets, Meng-Su Zeng, and Regina G H Beets-Tan. CT texture analysis in colorectal liver metastases: A better way than size and volume measurements to assess response to chemotherapy? *United European Gastroenterology Journal*, 4(2):257–263, apr 2016.
- [160] Vicky Goh, Balaji Ganeshan, Paul Nathan, Jaspal K Juttla, Anup Vinayan, and Kenneth A Miles. Assessment of Response to Tyrosine Kinase Inhibitors in Metastatic Renal Cell Cancer: CT Texture as a Predictive Biomarker. *Radiology*, 261(1):165–171, oct 2011.
- [161] Zijian Zhang, Jinzhong Yang, Angela Ho, Wen Jiang, Jennifer Logan, Xin Wang, Paul D Brown, Susan L McGovern, Nandita Guha-Thakurta, Sherise D Ferguson, Xenia Fave, Lifei Zhang, Dennis Mackin, Laurence E Court, and Jing Li. A predictive model for distinguishing radiation necrosis from tumour progression after gamma knife radiosurgery based on radiomic features from MR images. *European Radiology*, 28(6):2255–2263, 2018.
- [162] Alonso Garcia-Ruiz, Pablo Naval-Baudin, Marta Ligeró, Albert Pons-Escoda, Jordi Bruna, Gerard Plans, Nahum Calvo, Monica Cos, Carles Majós, and Raquel Perez-Lopez. Precise enhancement quantification in post-operative MRI as an indicator of residual tumor impact is associated with survival in patients with glioblastoma. *Scientific reports*, 11(1):695, jan 2021.
- [163] Jan C Peeken, Josefine Hesse, Bernhard Haller, Kerstin A Kessel, Fridtjof Nüsslin, and Stephanie E Combs. Semantic imaging features predict disease progression and survival in glioblastoma multiforme patients. *Strahlentherapie und Onkologie*, 194(6):580–590, 2018.
- [164] Benjamin M Ellingson, Lauren E Abrey, Sarah J Nelson, Timothy J Kaufmann, Josep Garcia, Olivier Chinot, Frank Saran, Ryo Nishikawa, Roger Henriksson, Warren P Mason, Wolfgang Wick, Nicholas Butowski, Keith L Ligon, Elizabeth R Gerstner, Howard Colman, John de Groot, Susan Chang, Ingo Mellinghoff, Robert J Young, Brian M Alexander, Rivka Colen, Jennie W Taylor, Isabel Arrillaga-Romany, Arnav Mehta, Raymond Y Huang, Whitney B Pope, David Reardon, Tracy Batchelor, Michael Prados, Evanthia Galanis, Patrick Y Wen, and

- Timothy F Cloughesy. Validation of postoperative residual contrast-enhancing tumor volume as an independent prognostic factor for overall survival in newly diagnosed glioblastoma. *Neuro-Oncology*, 20(9):1240–1250, aug 2018.
- [165] Saima Rathore, Hamed Akbari M.D., Jimit Doshi, Gaurav Shukla M.D., Martin Rozycki, Michel Bilello M.D., Robert A Lustig M.D., and Christos A Davatzikos. Radiomic signature of infiltration in peritumoral edema predicts subsequent recurrence in glioblastoma: implications for personalized radiotherapy planning. *Journal of Medical Imaging*, 5(2):1–10, mar 2018.
- [166] Benjamin M Ellingson. Contrast-Enhanced T1-Weighted Digital Subtraction for Increased Lesion Conspicuity and Quantifying Treatment Response in Malignant Gliomas BT - Glioma Imaging: Physiologic, Metabolic, and Molecular Approaches. pages 49–60. Springer International Publishing, Cham, 2020.
- [167] Evan D H Gates, Jie Yang, Kazutaka Fukumura, Jonathan S Lin, Jeffrey S Weinberg, Sujit S Prabhu, Lihong Long, David Fuentes, Erik P Sulman, Jason T Huse, and Dawid Schellingerhout. Spatial Distance Correlates With Genetic Distance in Diffuse Glioma. *Frontiers in Oncology*, 9:676, 2019.
- [168] Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, jan 1996.
- [169] Sue E. Knoblaugh and Julie Randolph-Habecker. Necropsy and Histology. In *Comparative Anatomy and Histology*, chapter 3, pages 23–51. Academic Press, 2 edition, jan 2018.
- [170] Paul C. Goodwin, Brian Johnson, and Charles W. Frevert. Microscopy, Immunohistochemistry, Digital Imaging, and Quantitative Microscopy. In *Comparative Anatomy and Histology*, chapter 4, pages 53–66. Academic Press, 2 edition, jan 2018.
- [171] Christopher S von Bartheld, Jami Bahney, and Suzana Herculano-Houzel. The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting. *Journal of Comparative Neurology*, 524(18):3865–3895, dec 2016.

- [172] S. Herculano-Houzel and R. Lent. Isotropic Fractionator: A Simple, Rapid Method for the Quantification of Total Cell and Neuron Numbers in the Brain. *Journal of Neuroscience*, 25(10):2518–2521, 2005.
- [173] S. Herculano-Houzel, C. E. Collins, P. Wong, and J. H. Kaas. Cellular scaling rules for primate brains. *Proceedings of the National Academy of Sciences*, 104(9):3562–3567, 2007.
- [174] Frederico A C Azevedo, Ludmila R B Carvalho, Lea T Grinberg, José Marcelo Farfel, Renata E L Ferretti, Renata E P Leite, Wilson Jacob Filho, Roberto Lent, and Suzana Herculano-Houzel. Equal Numbers of Neuronal and Nonneuronal Cells Make the Human Brain an Isometrically Scaled-up Primate Brain. *J. Comp. Neurol.*, 513(5):532–541, 2009.
- [175] Charlotte Havelund Nykjær, Tomasz Brudek, Lisette Salvesen, and Bente Pakkenberg. Changes in the cell population in brain white matter in multiple system atrophy. *Movement Disorders*, 32(7):1074–1082, jul 2017.
- [176] D.P. Pelvig, H. Pakkenberg, A.K. Stark, and B. Pakkenberg. Neocortical glial cell numbers in human brains. *Neurobiology of Aging*, 29(11):1754–1762, nov 2008.
- [177] Bente Pakkenberg and Hans Jørgen G Gundersen. Neocortical neuron number in humans: Effect of sex and age. *Journal of Comparative Neurology*, 384(2):312–320, jul 1997.
- [178] Solveig Walløe, Bente Pakkenberg, and Katrine Fabricius. Stereological estimation of total cell numbers in the human cerebral and cerebellar cortex. *Frontiers in Human Neuroscience*, 8:508, 2014.
- [179] Francine M Benes and Nicholas Lange. Two-dimensional versus three-dimensional cell counting: a practical perspective. *Trends in Neurosciences*, 24(1):11–17, jan 2001.

Vita

Evan Donald Huckins Gates (born Evan D. H. Johnson) was raised in Bellingham Washington and completed his secondary education at Sehome High School and Whatcom Community College. After, he attended Western Washington University, receiving a Bachelor of Science degree in 2014 with a double major in Physics and Mathematics. He continued on to receive a Master of Science degree in Mathematics from Western Washington University in 2016 and was named distinguished graduate. That same year, Evan moved to Houston Texas and entered The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences to pursue the doctor of philosophy degree in medical physics.