Predicting Brain Metastasis Response To Stereotactic Radiosurgery Using Magnetic Resonance Imaging Radiomics And Machine Learning

David A. DeVries, Western University

Supervisor: Ward, Aaron D., The University of Western Ontario
Co-Supervisor: Hajdok, George, The University of Western Ontario

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Medical Biophysics

© David A. DeVries 2023

Follow this and additional works at: https://ir.lib.uwo.ca/etd

Part of the Medical Biophysics Commons

Recommended Citation
https://ir.lib.uwo.ca/etd/9376

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlsadmin@uwo.ca.
Abstract

Brain metastases (BMs) represent advanced cancer, and so BM patients must be treated quickly and effectively while minimizing treatment toxicities. Stereotactic radiosurgery (SRS) uses conformal, ablative radiation doses to treat BMs, but with failure rates up to 30%. A predictive model of BM progression post-SRS would therefore aid in BM treatment selection and SRS planning. Previous studies have used pre-treatment T1-weighted contrast-enhanced magnetic resonance imaging (T1w-CE MRI) to predict SRS outcomes, through quantitative radiomic analysis with machine learning (ML) and qualitative appearance analysis. Comparison of these methods is not well understood, and ML methods have not been studied for sensitivity to relevant clinical factors, robust model interpretability, or multi-centre external validation. To meet these needs, a dataset of 123 BMs across 99 SRS patients was used to develop T1w-CE MRI radiomics-based ML models to understand their sensitivity to clinical factors. A ML model using radiomic and clinical features obtained the highest area-under-the-receiver-operating-characteristic-curve (AUC) of 0.77, and this model was sensitive to primary cancer site, BM volume, and MRI scanner model. BM volume sensitivity was reduced by removing volume-correlated radiomic features. An observer study of BM qualitative analysis revealed high interobserver variability that in turn limited outcome prediction, while ML models provided enhanced stratification of BMs into risk groups for post-SRS progression (Kaplan-Meier rank-sum $p = 0.0003$). BM qualitative appearance was useful for ML model interpretation, revealing the post-SRS progression radiomic signature is tied to necrotic or heterogeneous BM appearance, indicating a potentially less radiosensitive, hypoxic environment. An additional dataset of 117 BMs across 62 SRS patients was collected at a second centre for external validation. Transferring a locked model between centres revealed poor performance, but limiting the model to use radiomic features important at both centres increased the AUC to 0.70. Retraining a model with the second centre’s dataset using a locked methodology developed with the first centre’s dataset achieved a higher AUC of 0.80. In conclusion, this work successfully char-
acterized radiomics-based ML model performance with respect to clinical factors and BM qualitative appearance, while also providing the ML model interpretability and external validation necessary to motivate future research and clinical translation.

**Keywords:** brain metastasis, magnetic resonance imaging, stereotactic radiosurgery, machine learning, radiomics, treatment outcome prediction
Summary for lay audience

Brain metastases (BMs) occur when a patient’s cancer spreads to their brain. BMs cause painful symptoms and even death, and so must be treated quickly and effectively, while minimizing side-effects. Stereotactic radiosurgery (SRS) uses radiation to destroy BMs, but SRS may fail. To aid in decision-making on how to best use SRS, a system using pretreatment BM data as input could be developed to predict SRS failure. Magnetic resonance imaging (MRI) provides images within the brain and BMs, which clinicians can use to qualitatively score BM appearance to predict SRS failure. BM MRI can also undergo computerized extraction of quantitative “radiomic” data which machine learning (ML) systems use to learn how to predict SRS failure. Both approaches are useful, but they need robust comparison. How ML systems make predictions is not well interpreted, along with their sensitivity to variability in the input data. ML systems must also be validated to work at multiple cancer centres. This research used a dataset of 123 BMs to create predictive ML systems. It was found that variability in where a patient’s cancer started, BM volume, and model of MRI scanner was important to consider. Next, each BM’s qualitative appearance was scored by multiple clinicians. High scoring variability was found, showing the difficulty of using qualitative analysis. Since ML systems are computerized, they avoid this same variability, and they also better divided BMs into groups at risk for SRS failure. BM qualitative appearance was then used to better interpret how the ML systems were making decisions. It was found that the BMs the ML systems predicted would fail SRS were also the BMs labeled with qualitative appearances that indicate low BM oxygen levels, which negatively impacts SRS effectiveness. Lastly, we gathered another dataset of 117 BMs from a different cancer centre. The ML systems were the most accurate when systems were built for each centre, showing that our method for ML system building is validated across multiple centres. These advances in further understanding, comparing, and validating predictive ML systems are all important in moving such systems closer to helping patients receive the best care possible.
Co-authorship statement

This thesis is presented in an integrated format, consisting of three chapters (2–4) which each represent a journal article either in press or under review. The journal articles are presented as published or submitted, except for a short paragraph before each chapter’s introduction describing the chapter’s context within the broader thesis. The supplementary materials from each journal article are presented as published or submitted within appendices A–C. Details of the copyright permissions used to reproduce these journal articles within this thesis are provided in the “Copyright permissions” section at the end of the thesis. No part of this thesis was written using generative large-language models such as ChatGPT.

Each journal article relied heavily upon the use of a machine learning analysis software framework built jointly by myself, Salma Dammak, Carol Johnson, Ryan Alfano, and Aaron Ward for our common use (see appendix D). My contributions to this software framework included providing leadership in software design and writing software code and associated tests. The other team members were acknowledged in each journal article for their contributions to the software framework, but they were not included as co-authors unless otherwise merited.

Chapter 2: “Performance sensitivity analysis of brain metastasis stereotactic radiosurgery outcome prediction using magnetic resonance imaging radiomics”


My contributions to this work included formulating the experimental design, curating the original dataset, creating and performing all machine experiments, performing all analysis, interpreting results, producing all figures and tables, and writing of the manuscript. My contributions represent approximately 95% of the work completed for the study, as the dataset used in this work pre-existed at the start of the study, and so collecting the
Chapter 3: “Assessment of brain metastasis qualitative appearance interobserver variability and comparison to magnetic resonance imaging radiomics for stereotactic radiosurgery outcome prediction”

David DeVries, Terence Tang, Ali Albweady, Andrew Leung, Joanna Laba, Carol Johnson, Frank Lagerwaard, Jaap Zindler, George Hajdok, Aaron Ward. Scientific Reports (under review, May 2023)

My contributions to this work included formulating the experimental design, curating the observer study dataset, creating and performing all machine experiments, performing all analysis, interpreting results, producing all figures and tables, and writing of the manuscript. My contributions represent approximately 85% of the work completed for the study, as the dataset used in this work pre-existed at the start of the study, and so collecting the original dataset is not included in this estimation. Terence Tang, Ali Albweady, Andrew Leung, and Joanna Laba contributed to collecting the observer study dataset. Carol Johnson contributed to building the application used for collecting the observer study dataset. Frank Lagerwaard and Jaap Zindler contributed to collecting the original dataset. Aaron Ward contributed to formulating the experimental design, interpreting results, and in-depth editing of the manuscript. All authors contributed to reviewing the article. Andrew Warner is not a coauthor, but was acknowledged in the journal article for his support in conducting the statistical analyses. Aaron Ward and George Hajdok provided supervision of the project.
Chapter 4: “Dual-centre validation of using magnetic resonance imaging radiomics to predict stereotactic radiosurgery outcomes”


My contributions to this work included formulating the experimental design, securing study research ethics board approval, engineering data collection and storage infrastructure, collecting imaging and treatment planning system data, curating the external dataset, creating and performing all machine experiments, performing all analysis, interpreting results, producing all figures and tables, and writing of the manuscript. My contributions represent approximately 75% of the work completed for the study. Terence Tang, Ghada Alqaidy, and Ali Albweady contributed to collecting clinical data. Andrew Leung contributed to formulating the experimental design, collecting clinical data, and curating the external dataset. Joanna Laba contributed to formulating the experimental design and curating the external dataset. Frank Lagerwaard and Jaap Zindler contributed to formulating the experimental design and collecting the original dataset. George Hajdok contributed to formulating the experimental design and providing guidance during data collection. Aaron Ward contributed to formulating the experimental design, securing study research ethics board approval, interpreting results, and in-depth editing of the manuscript. All authors contributed to reviewing the article. Aaron Ward and George Hajdok provided supervision of the project.
This thesis is dedicated to my parents,

Terry and Glenda DeVries,

who have ceaselessly supported me in all I do.

Thank you mom and dad.
Acknowledgments

This thesis would not have been possible without the support of many people and organizations. First, thank you to my supervisors, Dr. George Hajdok and Dr. Aaron Ward. George, thank you for your support and mentorship over these past years. Your commitment to ensuring my research excelled, even when you relocated cities, is incredibly appreciated. Aaron, thank you most of all for caring deeply about my personal well-being throughout my entire PhD. Thank you as well for mentoring me in becoming a better scientist, leader, colleague, and teacher. I have grown in more ways that I could have imagined from when I started, so thank you for carefully cultivating the environment I grew in.

I would also like to thank my clinical collaborators that made this research possible. Dr. Frank Lagerwaard, Dr. George Rodrigues, Dr. Timothy Pok Chi Yeung, and Dr. Jaap Zindler were instrumental in getting my research started and provided expert guidance along the way. Dr. Ali Albweady, Dr. Ghada Alqaidy, and Dr. Terence Tang not only collected massive amounts of data, but also provided key feedback to improve this research. Thank you to Dr. Joanna Laba, Dr. Andrew Leung, and Dr. Jason Vickress for not only being members of my advisory committee, but for also being such dedicated and invested collaborators. I deeply appreciate your commitment to my research’s success and for generously taking the time to answer my questions, big or small.

Thank you as well to the members of my examining committee: Dr. Glenn Bauman, Dr. Issam El Naqa, Dr. Aaron Fenster, and Dr. Terry Peters. I appreciate you all taking the time and energy to review this thesis and to perform my thesis examination.

I would then like to thank the administrative staff, graduate executive committee members, department chair (Dr. Jeff Frisbee), and graduate program chairs (Dr. Aaron Ward and Dr. Charlie McKenzie) at the Department of Medical Biophysics who provided a firm foundation for my research. The Lawson Health Research Institute, London Health Sciences Centre, London Regional Cancer Program, and Gerald C. Baines Centre and their donors, staff, patients, and patient families also provided key support and countless op-
portunities to learn and perform research. Thank you especially to the medical physics department for their teaching and mentorship during my career development. Thank you to Mary Lu Lacasse as well for her daily energy and joy that would always brighten our lab. The Natural Sciences and Engineering Research Council, Government of Ontario, and Western University generously provided graduate student funding.

Next, thank you to my labmates. Dr. Wenchao Han, Laurie Huang, Dr. Ryan Alfano, and Dr. Chris Smith, I thoroughly enjoyed our time together and you each taught me so much. To Ryan Au, Vignesh Chakravarthy, Alana Lopes, and Robert Policelli, it has been a wonderful experience to welcome you into the lab and to work and learn together. I wish you all the best as you complete your degrees. Thank you as well to Andrew Warner and Dr. David Palma, for your friendship and willingness to answer questions about statistics, radiation oncology, and everything in between. Finally, thank you to Carol Johnson and Salma Dammak, who have been with me through it all. Carol, thank you for being our “lab mother” and caring so deeply for us all. Thank you for all the times you helped out in a pinch, even if that meant venturing into the unknown. Salma, thank you for always being willing to jump onto a new idea, debate a good topic, or play another hand of euchre. Your commitment to learning and research, but also to celebrating success and pushing through failure, has been an massive encouragement to me. I miss or will miss getting to work with you all, so thank you again for being the best of teams to work with.

Lastly, I would like to thank my family and friends, both near and far. You have supported me in uncountable ways, and have inspired me to go the distance in finishing this degree. And thank you to my little family: Rebecca, Levi, and Sadie. Levi and Sadie, it was a joy to welcome you into this world during my PhD, and thank you for helping me keep that world in perspective. Rebecca, thank you for being an endless source of love, encouragement, and fun. I always admire your selfless dedication to me, our family, and community, and it constantly inspires me to be the same. These past five years have been a journey, so thank you for being the best companion I could ever ask for on that journey.
# Table of contents

Abstract ii  
Summary for lay audience iv  
Co-authorship statement v  
Dedication viii  
Acknowledgments ix  
List of figures xviii  
List of tables xxiv  
List of abbreviations, symbols and nomenclature xxvii  

## 1 Introduction 1  
1.1 Brain metastases ........................................... 1  
1.2 Diagnosis ....................................................... 4  
1.2.1 Presenting neurological symptoms ......................... 5  
1.2.2 Medical imaging ............................................ 6  
  X-ray computed tomography ..................................... 6  
  Magnetic resonance imaging ................................... 7  
  Positron emission tomography .................................. 10  
1.2.3 Pathology .................................................... 11
1.3 Treatment techniques ............................................. 12
  1.3.1 Acute symptom management .............................. 14
    Corticosteroids ................................................. 14
  1.3.2 Systemic therapy ........................................... 14
  1.3.3 Surgical resection ........................................... 16
  1.3.4 Radiation Therapy ........................................... 17
    Whole brain radiation therapy ................................ 19
    Stereotactic radiosurgery and radiation therapy .......... 21
  1.3.5 Non-standard therapies .................................... 25
  1.3.6 Combinations of treatments ............................... 26
  1.3.7 Comparison and selection of treatments ................. 27
1.4 Prognostic and predictive models for brain metastasis patients .... 29
  1.4.1 Current prognostic models ................................ 29
  1.4.2 Current predictive models ................................ 30
  1.4.3 Machine learning techniques ............................. 33
  1.4.4 Machine learning models and radiomics ................. 36
  1.4.5 Current machine learning models ......................... 39
  1.4.6 Knowledge gaps and unmet needs in predicting brain metastasis response to stereotactic radiosurgery using machine learning .... 44
1.5 Thesis outline and research objectives ............................. 45
1.6 References .......................................................... 46

2 Performance sensitivity analysis of brain metastasis stereotactic radiosurgery outcome prediction using magnetic resonance imaging radiomics 67
  2.1 Introduction ....................................................... 68
  2.2 Methods .............................................................. 70
    2.2.1 Study sample ............................................... 70
    2.2.2 Clinical features and study endpoint ................... 71
3.2.2 Clinical features and study endpoint ........................................ 102
3.2.3 Imaging data and radiomic features ........................................ 103
3.2.4 Machine learning experimental design ..................................... 103
3.2.5 Qualitative appearance interobserver variability .......................... 104
3.2.6 Qualitative appearance machine learning models ......................... 105
3.2.7 Post-stereotactic radiosurgery progression machine learning models 106
3.2.8 Post-stereotactic radiosurgery progression machine learning model
interpretation ................................................................. 107
3.2.9 Ethics declaration .......................................................... 108
3.3 Results ....................................................................................... 109
3.3.1 Qualitative appearance interobserver variability ......................... 109
3.3.2 Qualitative appearance machine learning models ......................... 111
3.3.3 Post-stereotactic radiosurgery progression machine learning models 112
3.3.4 Post-stereotactic radiosurgery progression machine learning model
interpretation ................................................................. 113
3.4 Discussion ................................................................................. 115
3.4.1 Qualitative appearance interobserver variability ......................... 118
3.4.2 Qualitative appearance machine learning models ......................... 119
3.4.3 Post-stereotactic radiosurgery progression machine learning models 119
3.4.4 Post-stereotactic radiosurgery progression machine learning model
interpretation ................................................................. 120
3.4.5 Limitations and future work ..................................................... 123
3.5 Conclusions ............................................................................... 124
3.6 Data and code availability .......................................................... 124
3.7 References ................................................................................. 125

4 Dual-centre validation of using magnetic resonance imaging radiomics to
predict stereotactic radiosurgery outcomes 130
4.1 Introduction ......................................................... 130

4.2 Methods ......................................................... 133
  4.2.1 Centre A study sample ....................................... 133
  4.2.2 Centre B study sample ....................................... 135
  4.2.3 Clinical features ............................................. 136
  4.2.4 Radiomic features ........................................... 136
  4.2.5 Machine learning methodology ............................. 136
  4.2.6 Model external validation ................................... 137
  4.2.7 Methodology external validation ............................ 137
  4.2.8 Pooled data validation ....................................... 139
  4.2.9 Feature importance analysis ................................. 139
  4.2.10 Feature harmonization and consolidation .................. 139

4.3 Results .......................................................... 140
  4.3.1 Model external validation ................................... 140
  4.3.2 Methodology external validation ............................ 142
  4.3.3 Pooled data validation ....................................... 142
  4.3.4 Feature importance analysis ................................. 144
  4.3.5 Feature harmonization and consolidation .................. 144

4.4 Discussion ....................................................... 147
  4.4.1 Model external validation ................................... 147
  4.4.2 Methodology external validation ............................ 148
  4.4.3 Pooled data validation ....................................... 148
  4.4.4 Feature importance analysis ................................. 148
  4.4.5 Feature harmonization and consolidation .................. 149
  4.4.6 Limitations and future work ................................. 150

4.5 Conclusions ...................................................... 151

4.6 Data and code availability ...................................... 152
5 Conclusions and future work

5.1 Advances in knowledge and technology

5.1.1 Characterization of machine learning model sensitivity to clinical factors

5.1.2 Method to remove predictive model brain metastasis volume bias developed for and applied to stereotactic radiosurgery

5.1.3 Comparison of clinician-based qualitative analysis versus radiomics-based quantitative analysis for outcome prediction

5.1.4 Method to interpret radiomics-based machine learning models using brain metastasis qualitative appearance developed for and applied to stereotactic radiosurgery

5.1.5 External validation of extending radiomics-based machine learning outcome prediction to linear accelerator stereotactic radiosurgery

5.1.6 External validation of methodology to produce per-centre machine learning models

5.1.7 Characterization of the effect of magnetic resonance scanner model variability on radiomics-based machine learning model performance

5.2 Suggestions for future work

5.2.1 Further external validation studies and prospective data collection

5.2.2 Development and evaluation of enhanced magnetic resonance imaging normalization techniques

5.2.3 Interpretation of radiomic signatures through advanced imaging and pathology

5.2.4 Combination and application of predictive models for stereotactic radiosurgery tumour control and normal tissue toxicity

5.3 References
A Supplementary material related to Chapter 2: “Performance sensitivity analysis of brain metastasis stereotactic radiosurgery outcome prediction using magnetic resonance imaging radiomics” 174

B Supplementary material related to Chapter 3: “Assessment of brain metastasis qualitative appearance interobserver variability and comparison to magnetic resonance imaging radiomics for stereotactic radiosurgery outcome prediction” 183

C Supplementary material related to Chapter 4: “Dual-centre validation of using magnetic resonance imaging radiomics to predict stereotactic radiosurgery outcomes” 197

D Machine learning software framework 206

E Clinical trial data collection and management overview 225

F Clinical trial data collection forms 228

G Clinical trial database schema 248

H Clinical trial database software framework 255

I Clinical trial research ethics board approval 263

Copyright permissions 301

Curriculum vitae 303
List of figures

1.1 Common pathways to a BM diagnosis. 4
1.2 Different diagnostic “donut” detectors. 6
1.3 Examples of a slice from different medical imaging modalities used for BM imaging. 8
1.4 Schematic representation of how photon RT results in cell damage. 18
1.5 Reoxygenation mechanism in fractionated RT. 19
1.6 Representative planned dose distributions for different types of BM RT (a–d). 20
1.7 Schematic representation of five generalizations of how a BM can respond to SRS (a–e) and how the BM could present in follow-up MRI. 24
1.8 A representative comparison of TCP and NTCP curves showing the tradeoff between maximizing tumour control and minimizing toxicity. 31
1.9 Representative examples of BM qualitative appearances in T1w-CE MRI as labelled according to the five-way Rodrigues et al. scheme. 33
1.10 Schematic representation of ML (a) dataset splitting, (b) model training, and (c) model testing. 34
1.11 Representation of underfit, well-fit, and overfit classifier models for the same training and testing datasets. 36
1.12 Example of a NBC differentiating two classes using a single feature. 37
2.1 Comparison of average ROC curves across bootstrapped repetitions when clinical, radiomic, or clinical and radiomic features were used. 78
2.2 Volume dependency of error metrics. 80
2.3 Error metrics across MR scanner models. ..................................................... 83

3.1 Models used for risk stratification of BMs for progression post-SRS. ........... 99

3.2 KM analysis for risk of a BM progressing post-SRS for the six primary risk stratification models evaluated (a–f). .......................................................... 110

3.3 Error metrics from the radiomic consensus appearance experiments. ........... 112

3.4 Comparison of the average predicted probability of progression for each BM from the radiomic progression experiment when grouped by qualitative appearance, as labelled by expert consensus. .............................. 114

3.5 ALE plots comparing the effect of individual features on model predicted probabilities between the radiomic progression experiment and radiomic consensus appearance experiments. .............................. 117

4.1 Visual representation of the three types of validation performed. ................. 138

4.2 ROC curves and operating points for validation experiments. ...................... 141

4.3 Per-centre radiomic feature importance analysis and associated feature consolidation. ............................................................. 145

4.4 Comparison of applying radiomic feature harmonization and consolidation against baseline, and ROC curves when using the optimal top \( n \) consolidated features. ................................................................. 146

5.1 Schematic of using pre-treatment patient and BM data with radiomics-based ML models to choose an optimal SRS/SRT prescription. ......................... 167

A.1 Schematic diagram of the ML experimental method technique used in the study. ................................................................. 177

A.2 Comparison of normalized feature importance scores for each feature (y-axis) and for each feature volume-correlation coefficient threshold value (lower x-axis). ................................................................. 180
B.1 Schematic diagram of the ML experimental method technique used in the study. .......................................................... 187
B.2 Schematic of the ML experiments and analysis associated with (a) the radiomic appearance experiments (as needed for figure 3.1b), (c) the radiomic progression experiment (as needed for figure 3.1c), (d) the radiomic and clinical progression experiment (as needed for figure 3.1d), and (b) the analysis performed for ML model interpretation using the results from (a) and (c). .......................................................... 190
B.3 KM analysis plots for progressive disease for each individual observer for comparison against Expert 1 (see figure 3.2a). ......................... 193
B.4 Error metrics from the radiomic Expert 1 appearance experiments using the observer labels from the original study. ...................... 194
B.5 ALE plots for the highly important features from table 3.3 that were not included in figure 3.5, as these features were not also highly important for any of the radiomic appearance label experiments. ..................... 195
C.1 Schematic of the experimental design for (a) model external validation and (b) methodology external validation or pooled data validation. ...... 201
C.2 ROC curves and operating points for additional validation experiments. . . . . 204
D.1 Symbol and property name legends for UML class diagrams. .................. 207
D.2 UML class diagram providing an overview of the ML software framework. . 208
D.3 UML class diagram of the “GeometricalImagingObject” portion of the ML software framework. .................................................. 209
D.4 UML class diagram of the “ImageVolume” portion of the ML software framework. .................................................. 210
D.5 UML class diagram of the “RegionsOfInterest” portion of the ML software framework. .................................................. 211
D.6 UML class diagram of the “ImagingObjectTransform” portion of the ML software framework. .............................. 212
D.7 UML class diagram of the “ImageVolumeViewer” portion of the ML software framework. ............................. 213
D.8 UML class diagram of the “FeatureExtractionImageVolumeHandler” portion of the ML software framework. .................. 214
D.9 UML class diagram of the “Feature” portion of the ML software framework. 215
D.10 UML class diagram of the “FeatureExtractionRecord” portion of the ML software framework. ............................. 216
D.11 UML class diagram of the “FeatureValues” portion of the ML software framework. ................................. 218
D.12 UML class diagram of the “Classifier” portion of the ML software framework. 219
D.13 UML class diagram of the “FeatureSelector” portion of the ML software framework. ............................. 220
D.14 UML class diagram of the “HyperParameterOptimizer” portion of the ML software framework. .................. 221
D.15 UML class diagram of the “ErrorMetric” and “ObjectiveFunction” portions of the ML software framework. .................. 222
D.16 Class methods and properties of the “Experiment” class in the ML software framework. ............................. 223
D.17 UML class diagram of the “Experiment” portion of the ML software framework. ................................. 224
E.1 Schematic representation of the data sources and data management systems used to collect and process the LHSC SRS dataset. .......................... 227
F.1 “Patient Information” REDCap data collection form. ......................... 229
F.2 “Aria Disease Diagnosis” REDCap data collection form. ......................... 229
F.3 “Primary Cancer Histopathology” REDCap data collection form (two pages). 231
F.4 “Brain Metastasis Histopathology” REDCap data collection form (two pages). 233
F.5 “Systemic Therapy” REDCap data collection form. 234
F.6 “Salvage Therapy” REDCap data collection form. 235
F.7 “Radionecrosis Treatment” REDCap data collection form. 236
F.8 “Brain Radiation Course” REDCap data collection form (two pages). 238
F.9 “Lung Radiation Course” REDCap data collection form (two pages). 240
F.10 “Brain Radiology Pre-Radiation” REDCap data collection form (three pages). 243
F.11 “Brain Radiology Follow-up” REDCap data collection form (two pages). 245
F.12 “Brain Radiology Pseudo-Progression Conclusion” REDCap data collection form (two pages). 247

G.1 Legend for database schema figures. 249
G.2 Database schema representing the storage of data collected from the Aria RVS. 249
G.3 Database schema representing the storage of BM SRS treatment data. 250
G.4 Database schema representing the storage of imaging metadata. 251
G.5 Database schema representing the storage of systemic, RN, and SRS salvage treatment data. 252
G.6 Database schema representing the storage of histopathology data. 253
G.7 Database schema representing the storage of BM radiologic assessment data. 254

H.1 UML class diagram for the objects accessing patient and BM data. 256
H.2 UML class diagram for the objects accessing data collected from the Aria RVS. 257
H.3 UML class diagram for the objects accessing BM SRS treatment data. 258
H.4 UML class diagram for the objects accessing imaging data. 259
H.5 UML class diagram for the objects accessing systemic, RN, and SRS salvage treatment data. 260

H.6 UML class diagram for the objects accessing histopathology data. 261

H.7 UML class diagram for the objects accessing radiologic assessment data. 262

I.1 Study approval letter from the Western University Health Sciences Research Ethics Board 264

I.2 Study application to the Western University Health Sciences Research Ethics Board (30 pages). 294

I.3 Research plan submitted to the Western University Health Sciences Research Ethics Board (6 pages). 300
List of tables

1.1 Approximate relative incidence and risk of developing BMs per primary cancer site. .......................................................... 3
1.2 Summarization of BM presenting symptoms according to brain region. . . . 5
1.3 Summarization and comparison of the most commonly employed treatment techniques for BMs. .......................................................... 13
1.4 Summarization of the HyTEC guidelines for BM SRS/SRT control rates (a) and RN toxicity risk (b), grouped by number of fractions. ............... 25
1.5 Summarization of recommendations for BM treatment selection, based on the guidelines from ASCO-SNO-ASTRO. ............................... 28
1.6 Summary and comparison of the work presented in this thesis to previous studies investigating the prediction of BM response to SRS/SRT using radiomics. .......................................................... 43

2.1 Clinical feature distributions for number of BMs, BMs progressing post-SRS, and patients (where applicable) for the study sample. ................. 72
2.2 Number of patients and BMs scanned by each of the MR scanner model and acquisition parameters configurations. ........................................... 73
2.3 Error metrics derived by analysis per BM primary cancer site. ............... 79

3.1 Study observers that provided qualitative appearance labels for all BMs. . . . 104
3.2 Qualitative appearance labelling agreement rates between all pairs of observers. .......................................................... 111
3.3 Feature importance ranks, scores, and ALE correlation comparison between the radiomic progression experiment and radiomic consensus appearance experiments. ..................................................... 116

4.1 Univariate analysis of clinical and radiomic features for Centres A and B to separate BMs that progressed post-SRS from those that did not, and separation of Centre A and Centre B BMs. ......................................................... 134

4.2 Error metrics values across the validation experiments performed. ........ 143

A.1 Complete catalogue of the 119 features included within the study. ........ 175

A.2 Hyperparameters for the RDF model used. ............................................. 179

A.3 Comparison of the normalized feature importance score and rank for all features with an importance score $\geq 0.75$ for either correlation threshold $= 1$ (no features removed) or threshold $= 0.25$ (volume-correlated features removed). ................................................................. 182

B.1 Clinical feature distributions for number of BMs, BMs progressing post-SRS, and patients (where applicable) for the study sample. .......... 184

B.2 Number of patients and BMs scanned by each of the MR scanner models and acquisition parameter configurations. ................................. 185

B.3 Complete catalogue of the 107 radiomic features included within the study. 186

B.4 Hyperparameters for the RDF model used. ............................................. 189

B.5 Confusion matrices across appearance labels (A–E), for each pairwise comparison between observers. ......................................................... 192

C.1 Number of BMs imaged by each imaging configuration dependent on scanner model, acquisition orientation, and voxel size. ................. 198
C.2 Post-SRS progression, no progression and pseudo-progression calls for Centre B’s dataset, broken down according to patterns of post-SRS BM size over time. .................................................. 199
C.3 Complete catalogue of the 107 radiomic features included within the study. 200
C.4 Hyperparameters for the RDF model used. .............................................. 203
# List of abbreviations, symbols and nomenclature

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>three-dimensional</td>
</tr>
<tr>
<td>ADC</td>
<td>apparent diffusion coefficient</td>
</tr>
<tr>
<td>AEDs</td>
<td>antiepileptic drugs</td>
</tr>
<tr>
<td>ALE</td>
<td>accumulated local effects</td>
</tr>
<tr>
<td>ALK</td>
<td>anaplastic lymphoma kinase</td>
</tr>
<tr>
<td>API</td>
<td>application programming interface</td>
</tr>
<tr>
<td>ARE</td>
<td>adverse radiation effect</td>
</tr>
<tr>
<td>ASCO</td>
<td>American Society of Clinical Oncology</td>
</tr>
<tr>
<td>ASTRO</td>
<td>American Society for Radiation Oncology</td>
</tr>
<tr>
<td>AUC</td>
<td>area under the curve</td>
</tr>
<tr>
<td>AUMC</td>
<td>Amsterdam University Medical Centre</td>
</tr>
<tr>
<td>BM</td>
<td>brain metastasis</td>
</tr>
<tr>
<td>BRAF</td>
<td>B-RaF protein</td>
</tr>
<tr>
<td>CBV</td>
<td>cerebral blood volume</td>
</tr>
<tr>
<td>CE</td>
<td>contrast enhanced</td>
</tr>
<tr>
<td>CI</td>
<td>confidence interval</td>
</tr>
<tr>
<td>CNNs</td>
<td>convolutional neural networks</td>
</tr>
<tr>
<td>Co-60</td>
<td>cobalt-60</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>ComBat</td>
<td>combating batch effects when combining batches</td>
</tr>
<tr>
<td>CSF</td>
<td>cerebral spinal fluid</td>
</tr>
<tr>
<td>CT</td>
<td>computed tomography</td>
</tr>
<tr>
<td>CTV</td>
<td>clinical treatment volume</td>
</tr>
<tr>
<td>DICOM</td>
<td>Digital Imaging and Communications in Medicine</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DS-GPA</td>
<td>diagnosis-specific graded prognostic assessment</td>
</tr>
<tr>
<td>DWI</td>
<td>diffusion weighted imaging</td>
</tr>
<tr>
<td>EANO</td>
<td>European Association of Neuro-Oncology</td>
</tr>
<tr>
<td>ECOG</td>
<td>Eastern Cooperative Oncology Group</td>
</tr>
<tr>
<td>EGFR</td>
<td>epidermal growth factor receptor</td>
</tr>
<tr>
<td>EHR</td>
<td>electronic health record</td>
</tr>
<tr>
<td>ESMO</td>
<td>European Society for Medical Oncology</td>
</tr>
<tr>
<td>FDG</td>
<td>$^{18}$F-2-fluoro-2-deoxy-D-glucose</td>
</tr>
<tr>
<td>FLAIR</td>
<td>fluid attenuated inversion recovery</td>
</tr>
<tr>
<td>FNR</td>
<td>false negative rate</td>
</tr>
<tr>
<td>FPR</td>
<td>false positive rate</td>
</tr>
<tr>
<td>fx</td>
<td>fractions</td>
</tr>
<tr>
<td>GI</td>
<td>gastrointestinal</td>
</tr>
<tr>
<td>GL</td>
<td>gray-level</td>
</tr>
<tr>
<td>GLCM</td>
<td>gray-level co-occurrence matrix</td>
</tr>
<tr>
<td>GLDM</td>
<td>gray-level dependence matrix</td>
</tr>
<tr>
<td>GLRLM</td>
<td>gray-level run length matrix</td>
</tr>
<tr>
<td>GLSZM</td>
<td>gray-level size zone matrix</td>
</tr>
<tr>
<td>GPA</td>
<td>graded prognostic assessment</td>
</tr>
<tr>
<td>GTV</td>
<td>gross tumour volume</td>
</tr>
<tr>
<td>Gy</td>
<td>Gray (J/kg)</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>HA-WBRT</td>
<td>hippocampal avoidance whole brain radiation therapy</td>
</tr>
<tr>
<td>HER2</td>
<td>human epidermal growth factor receptor 2</td>
</tr>
<tr>
<td>HyTEC</td>
<td>Hy Dose per Fraction, Hypofractionated Treatment Effects in the Clinic</td>
</tr>
<tr>
<td>IMRT</td>
<td>intensity-modulated radiation therapy</td>
</tr>
<tr>
<td>KM</td>
<td>Kaplan-Meier</td>
</tr>
<tr>
<td>KNN</td>
<td>k nearest neighbours</td>
</tr>
<tr>
<td>LHSC</td>
<td>London Health Sciences Centre</td>
</tr>
<tr>
<td>linac</td>
<td>linear accelerator</td>
</tr>
<tr>
<td>LITT</td>
<td>laser interstitial thermal therapy</td>
</tr>
<tr>
<td>MAB</td>
<td>monoclonal antibody</td>
</tr>
<tr>
<td>MCR</td>
<td>misclassification rate</td>
</tr>
<tr>
<td>MDPI</td>
<td>Multidisciplinary Digital Publishing Institute</td>
</tr>
<tr>
<td>ML</td>
<td>machine learning</td>
</tr>
<tr>
<td>MLP</td>
<td>multi-layer perceptron</td>
</tr>
<tr>
<td>MPRAGE</td>
<td>magnetization-prepared rapid gradient-echo</td>
</tr>
<tr>
<td>MRI</td>
<td>magnetic resonance imaging</td>
</tr>
<tr>
<td>MRS</td>
<td>magnetic resonance spectroscopy</td>
</tr>
<tr>
<td>NBC</td>
<td>naïve Bayes classifier</td>
</tr>
<tr>
<td>NGTDM</td>
<td>neighbouring gray tone difference matrix</td>
</tr>
<tr>
<td>NSCLC</td>
<td>non-small-cell lung cancer</td>
</tr>
<tr>
<td>NTCP</td>
<td>normal tissue complication probability</td>
</tr>
<tr>
<td>OAR</td>
<td>organ-at-risk</td>
</tr>
<tr>
<td>OOP</td>
<td>object-oriented programming</td>
</tr>
<tr>
<td>OS</td>
<td>overall survival</td>
</tr>
<tr>
<td>PACS</td>
<td>picture archiving and communication system</td>
</tr>
<tr>
<td>PCI</td>
<td>prophylactic cranial irradiation</td>
</tr>
<tr>
<td>PET</td>
<td>positron emission tomography</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>PTV</td>
<td>planning treatment volume</td>
</tr>
<tr>
<td>QOL</td>
<td>quality of life</td>
</tr>
<tr>
<td>QUANTEC</td>
<td>Quantitative Analysis of Normal Tissue Effects in the Clinic</td>
</tr>
<tr>
<td>RANO-BM</td>
<td>Response Assessment in Neuro-Oncology Brain Metastases</td>
</tr>
<tr>
<td>RDF</td>
<td>random decision forest</td>
</tr>
<tr>
<td>REDCap</td>
<td>Research Electronic Data Capture</td>
</tr>
<tr>
<td>RN</td>
<td>radionecrosis</td>
</tr>
<tr>
<td>ROC</td>
<td>receiver operating characteristic</td>
</tr>
<tr>
<td>ROI</td>
<td>region-of-interest</td>
</tr>
<tr>
<td>RPA</td>
<td>recursive partitioning analysis</td>
</tr>
<tr>
<td>RT</td>
<td>radiation therapy</td>
</tr>
<tr>
<td>RTOG</td>
<td>Radiation Therapy Oncology Group</td>
</tr>
<tr>
<td>RVS</td>
<td>record and verify system</td>
</tr>
<tr>
<td>SCLC</td>
<td>small-cell lung cancer</td>
</tr>
<tr>
<td>SIB</td>
<td>simultaneous in-field boost</td>
</tr>
<tr>
<td>SNO</td>
<td>Society for Neuro-Oncology</td>
</tr>
<tr>
<td>SRS</td>
<td>stereotactic radiosurgery</td>
</tr>
<tr>
<td>SRT</td>
<td>stereotactic radiation therapy</td>
</tr>
<tr>
<td>SVM</td>
<td>support vector machine</td>
</tr>
<tr>
<td>T1w</td>
<td>T1-weighted</td>
</tr>
<tr>
<td>T2w</td>
<td>T2-weighted</td>
</tr>
<tr>
<td>TCP</td>
<td>tumour control probability</td>
</tr>
<tr>
<td>TKI</td>
<td>tyrosine kinase inhibitor</td>
</tr>
<tr>
<td>TPS</td>
<td>treatment planning systems</td>
</tr>
<tr>
<td>UML</td>
<td>Unified Modelling Language</td>
</tr>
<tr>
<td>VEGF</td>
<td>vascular endothelial growth factor</td>
</tr>
<tr>
<td>VMAT</td>
<td>volumetric modular arc therapy</td>
</tr>
</tbody>
</table>
voxel volume element
WBRT whole brain radiation therapy
WHO World Health Organization
Chapter 1

Introduction

1.1 Brain metastases

Cancerous malignancies form when genetic mutations allow for the uncontrolled multiplication of cells within the body. These mutations allow cancerous cells to develop various capabilities to replicate faster, resist cell death, evade the body’s immune system, and recruit more resources from the body to enable further growth [1]. The incidence of cancer in the Canadian and global population is dependent on various factors including changing demographics, cancer prevention strategies, exposure to carcinogens, and genetic risk factors. Globally, there was an estimated 19.3 million new cases of cancer in 2020 [2], while in Canada, there was an estimated 223,900 new cases of cancer in 2022, with approximately 2 out of 5 Canadians expected to develop cancer in their lifetime [3]. As many of us know through personal experience or the experiences of our friends or family, cancer can exact a heavy toll on patients through cancer symptoms, treatment side-effects, and potentially death.

Cancer can become particularly lethal and difficult to treat when it spreads from the primary site where it developed to elsewhere in the body. An additional key aspect of cancer cells is their ability to leave the primary site where the cancer formed and survive within one
of the body’s transportation system, such as the lymphatic or circulatory system [1]. The cancer cells can then travel to elsewhere in the body where they can arrest to form a new cancerous lesion, known as a metastasis. Metastases have been found to form preferentially in certain organs of the body such as the lungs, brain, liver, and bones, with this effect also dependent on the primary cancer site involved [4, 5].

A variety of primary cancers can form within the brain, but the brain is also a common target for metastatic spread from elsewhere in the body. Primary cancers that form from brain cells primarily form in the non-neuron glial cells to form gliomas. Gliomas differ in their grade and type of glial cell they form from, with high-grade glioblastoma being the most common form of primary brain cancer [6]. As these primary brain cancers develop from brain cells, they form a distinctly different disease than brain metastases (BMs) that form due to metastatic spread of cancer from elsewhere in the body. BMs are seeded by cancer cells that depart the primary tumour and spread through the bloodstream. These cells travel into the brain and typically stop in areas where blood vessels narrow, such as at gray and white matter junctions and areas far from major cerebral arteries [7]. The circulatory system in the brain features a blood-brain barrier made of tightly joined endothelial cells that work with other glial cells to prevent foreign cells from entering the brain [8]. Metastatic cancer cells are able to disrupt the blood-brain barrier, however, allowing them to then grow into a cancerous lesion within the brain parenchyma [4, 8]. As the BM grows, it can remain a solid mass or become a fluid-filled cystic lesion, depending on the cancer cell physiology [9, 10]. The BM also induces angiogenesis to support its growth, but if the BM grows large enough, it may outstrip this blood supply, causing a hypoxic core within the BM that leads to cell necrosis [11].

While the majority of BMs form in the brain parenchyma as described above, it is also possible for them to form in the meningeal membrane layers surrounding the brain. Seeding into the meningeal layers can occur naturally during the progression of cancer, or as a consequence of surgical resection of a BM providing a vector for the transportation of
BM cells to the meninges [12]. Leptomeningeal disease forms when metastasis occurs in the subarachnoid space, arachnoid mater, or pia mater between the brain parenchyma and dura mater next to the skull, while pachymeningeal disease occurs in the dura mater. Due to the location of leptomeningeal and pachymeningeal disease, it spreads along the surface of the brain parenchyma, making it difficult to treat [12–14].

BM incidence is growing due to enhanced medical imaging surveillance and longer survival of cancer patients due to improved treatments of both primary cancer and oligometastatic disease [17]. The onset of metastatic brain cancer is still associated with the highest stage of cancer, with the median overall survival (OS) of BM patients being relatively short at a reported 8-16 months depending on the primary cancer site [20]. While BM patients do have a short OS, the presence of BMs is not necessarily the cause of the death. Some BM patients do succumb due to neurological complications (about one third), but more
commonly BMs are associated with an advanced primary disease and widespread metastasis throughout the body, with death caused by this extracranial disease [21]. Therefore successful diagnosis and prompt treatment of BMs can achieve meaningful increases to patients’ length and quality of life (QOL).

## 1.2 Diagnosis

BMIs can be diagnosed through a number of pathways. As shown in figure 1.1a, a patient may already have a primary cancer diagnosis when either a symptomatic or asymptomatic BM is discovered (approximately 70% [22]). Alternatively, a BM may be discovered before
<table>
<thead>
<tr>
<th>Brain Region</th>
<th>Associated Symptoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>• Headache</td>
</tr>
<tr>
<td></td>
<td>• Nausea and/or vomiting</td>
</tr>
<tr>
<td></td>
<td>• Seizures</td>
</tr>
<tr>
<td></td>
<td>• Fatigue</td>
</tr>
<tr>
<td>Frontal lobe</td>
<td>• Motor function losses</td>
</tr>
<tr>
<td></td>
<td>• Expressive aphasia</td>
</tr>
<tr>
<td></td>
<td>• Attention and concentration issues</td>
</tr>
<tr>
<td>Parietal lobe</td>
<td>• Sensory processing difficulties</td>
</tr>
<tr>
<td></td>
<td>• Left versus right side confusion or neglect</td>
</tr>
<tr>
<td></td>
<td>• Speaking or writing issues</td>
</tr>
<tr>
<td>Temporal lobe</td>
<td>• Memory loss</td>
</tr>
<tr>
<td></td>
<td>• Receptive aphasia</td>
</tr>
<tr>
<td>Occipital lobe</td>
<td>• Vision loss or changes</td>
</tr>
<tr>
<td>Cerebellum</td>
<td>• Balance or gait disorder</td>
</tr>
<tr>
<td></td>
<td>• Loss of coordination</td>
</tr>
<tr>
<td>Brainstem</td>
<td>BMs rarely observed in this brain region</td>
</tr>
</tbody>
</table>

Table 1.2: Summarization of BM presenting symptoms according to brain region. This is not an exhaustive list of all possible BM symptoms, but represents those that are most commonly observed.

the primary cancer for some patients, which then leads to the diagnosis of the associated primary cancer (figure 1.1b). When using these pathways to make a BM diagnosis, clinicians may utilize the patient’s presenting neurological symptoms, medical imaging, and pathology, as described in the following sections.

1.2.1 Presenting neurological symptoms

As BMs form, they can cause significant deterioration and deformation of the brain tissue. In addition to the physical lesion of cancerous cells, which typically has limited or no microscopic infiltration into the surround brain tissue [9, 23, 24], vasogenic edema commonly forms as fluid leaks across the disrupted blood-brain barrier [25, 26]. These effects can then in turn cause deformation of the brain tissue through mass effect near the BM, ventricular swelling, and a general increase of intracranial pressure [25]. These physical changes to the brain may then result in variety of neurological symptoms that are general or specific to the brain region in which BMs have formed (see table 1.2) [25, 27].
1.2.2 Medical imaging

If a patient presents with BM symptoms, or if a primary cancer that is likely to metastasize to the brain is diagnosed, a variety of medical imaging modalities may be used to detect BMs. The modalities most typically used are x-ray computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET).

**X-ray computed tomography**

A x-ray CT scan is performed by acquiring a series of two-dimensional x-ray images of the patient that are then used to reconstruct a single three-dimensional (3D) image. During a CT scan, a kilovoltage x-ray tube produces an x-ray beam that passes through the patient (see figure 1.2a). X-ray photons in the beam are absorbed and scattered as they pass through the patient’s body, with the probability of interaction described primarily by the
photoelectric effect that is dependent on the atomic number (Z) of tissues. The photons that are transmitted through the body are recorded by an electronic detector and digitized. The source and detector are then rotated and the process is repeated. By using the digitized transmission data, reconstruction algorithms can be used to compute a 3D matrix of “volume elements” (voxels) of the scanned area. Each voxel is assigned a value proportional to its ability to interact with x-rays, which can then be assembled and viewed as a series of image slices in grayscale with high Z voxels represented as white and low Z voxels represented as black, as shown in figure 1.3a.

X-ray CT scans unfortunately provide limited contrast between soft tissues, and so are limited in use for BM detection without the scan being contrast enhanced (CE). As most tissues within the body primarily have a similar effective Z number close to that of water, x-ray CT provides strongest contrast between soft tissues of different densities, air, and bony anatomy. BMs are of similar density and Z compared to brain tissue, and therefore do not have high contrast in x-ray CT [11, 13, 28]. A CE CT can be performed, however, by injecting a high Z iodine contrast agent intravenously before the CT scan, which will preferentially cross the blood-brain barrier where it is disrupted by a BM, making CE CT useful for BM detection [13, 29].

**Magnetic resonance imaging**

While CE CT allows for adequate BM detection, MRI scans provide superior imaging of BMs by exploiting an alternative physical mechanism based on magnetism to produce medical images. To acquire an MR image, the MR scanner produces a high primary magnetic field that the patient is placed within (see figure 1.2b). Each hydrogen atom’s nucleus within the patient’s body has a magnetic moment, which in aggregate produce a net magnetization that aligns with the scanner’s primary magnetic field. The scanner then excites the hydrogen nuclei by transmitting energy from transmit coils in the form of very high frequency radio waves (∼100 MHz), which causes the nuclei’s magnetic moments (and
Figure 1.3: Examples of a slice from different medical imaging modalities used for BM imaging. (a) shows x-ray CT imaging where the BM (indicated by the arrow) is difficult to visualize due to the low soft tissue contrast of the modality. The surrounding edema and mass effect distortion of the ventricle on the left are visible, however. This x-ray CT is reflective of typical clinical practice in that it is not CE, as the patient also received CE MRI. (b) shows a T2w MRI, where the fluid present in the CSF and BM edema appear hyperintense, but the extent of the BM margins are not distinct. (c) and (d) demonstrate T1w MRI pre and post-CE (respectively). While the T1w pre-CE MRI clearly differentiates the gray and white matter of the brain and shows the BM edema and mass effect, the extent of the BM is only revealed under CE. The BM T1w-CE MRI also demonstrates a heterogeneous enhancement pattern, likely due to poor vascularization within the centre of the BM leading to a lack of contrast wash-in. (e) and (f) show the FA and ADC quantitative maps produced by DWI, with the BM and edema best visualized by comparing the indicated location to the contralateral side.

Therefore the net magnetization as well) to precess about the primary magnetic field axis. This precession in turn induces a current within the receive coils of the MR scanner, with this signal being measured over time and digitized.

After the initial excitation, the signal received can decay due to two processes: 1) the hydrogen nuclei release energy to realign with the primary magnetic field (known as T1...
decay), and 2) the hydrogen nuclei precession rates are not identical and so fall out of phase, leading to signal decay (known as T2 decay). An image can then be formed based on either of these decay processes, known as T1 or T2 weighted (T1w or T2w) images. In order to produce a T1w or T2w image value per voxel within the final 3D image, spatial encoding techniques such as slice selection, frequency encoding, and phase encoding can be used to transform a received MR signal into values for voxels at specific locations.

In MRI, neuroanatomy, BMs, and edema have specific appearances if the MR scan is T1w or T2w. In a T1w scan, BMs may appear slightly hypointense or isointense compared to brain tissue, while edema may appear slightly hypointense [13, 27] (figure 1.3c). In a T2w scan, BMs appear slightly hyperintense, while edema is clearly visible as a hyperintense region, as fluid provides a clear T2w signal [13, 27] (figure 1.3b). A type of T2w scan known as a fluid attenuated inversion recovery (FLAIR) scan can be specifically programmed to nullify the signal from cerebral spinal fluid (CSF), and so in these scans edema will be more clearly differentiated from CSF [27].

Similarly to CE CT, contrast agents can also be used with MRI to provide improved BM detection. In T1w-CE MRI, gadolinium contrast agents are administered intravenously before the scan and circulate through the bloodstream. Gadolinium is a paramagnetic atom, giving it a strong magnetic moment that in turn allows nearby hydrogen atoms to have a very short T1 decay time. This property causes areas with gadolinium present to appear strongly hyperintense in T1w MRI. As with iodine contrast in CE CT, gadolinium contrast agents will appear where the blood-brain barrier is disrupted by BMs, allowing for their detection [11, 13, 28]. While small BMs will uniformly enhance in T1w-CE MRI, as BMs grow larger, some BM areas will not enhance if they are cut-off from the blood supply, indicating hypoxia or necrosis [11] (figure 1.3d).

Diffusion weighted imaging (DWI) MRI allows for quantification of the rate and direction of molecular diffusion within the brain. To acquire DWI MRI data, multiple scans are acquired, each with a smaller secondary magnetic field added to the primary magnetic field...
in a different direction. After acquiring the data for secondary magnetic fields in a number of evenly distributed 3D directions, the data can be processed to produce quantitative maps of diffusion metrics. The apparent diffusion coefficient (ADC) and fractional anisotropy (FA) are the two most commonly used diffusion metrics, which quantify the overall amount and preferential directionality of diffusion, respectively (figure 1.3e,f). These scan types are being used more frequently in BM diagnosis, especially in providing differential diagnoses between BMs and primary brain cancers or benign lesions [13, 28].

**Positron emission tomography**

The third imaging modality typically used for BM imaging is PET, which captures the location of positron particles within the body. Positrons, the anti-matter counterparts of electrons, interact with electrons through annihilation of both particles. During such an annihilation event, two $\gamma$-ray photons of equal energy (0.511 MeV each) are released in nearly opposite directions. A PET scanner consists of a ring of $\gamma$-ray detectors around the patient that record the time and location of $\gamma$-rays exiting the patient, which then can be used to reconstruct where positron-electron annihilation events occurred (figure 1.2c). A PET image therefore contains no anatomical information, and so an accompanying CT or MRI scan is used to match areas containing positrons with the patient’s anatomy and provide corrections for the attenuation of $\gamma$-rays by the patient’s tissue.

Positrons do not naturally exist within the body, and so radioactive positron-emitting tracers are injected intravenously and their preferential homing to BMs is imaged and interpreted. The most common PET tracer is $^{18}$F-2-fluoro-2-deoxy-D-glucose (FDG), which mimics glucose while also including a positron-emitting fluorine-18 atom within the molecule. FDG shows areas of the body with high levels of glycolysis, which is especially useful as increased metabolism is often a hallmark of cancerous cells. In the brain, however, FDG PET is of limited use, as neural cells naturally have a high metabolism rate as well, leading to low specificity for BM detection using FDG PET [30, 31]. There are also PET tracers
based on amino acids that are more specific for BM detection due to the increased concentration of amino acids specifically associated with BMs, but these tracers are not widely in use, leaving T1w-CE MRI as the standard modality for BM diagnosis [28, 30].

1.2.3 Pathology

While medical imaging of a brain lesion and an extracranial primary cancer diagnosis can provide a presumed BM diagnosis, pathology is the only technique to provide a definitive diagnosis. Pathological assessment of tissues is performed by obtaining and staining a tissue sample and then examining the sample microscopically to determine if cancer is present and if it displays any specific markers. As shown in figure 1.1, if a patient is diagnosed with extracranial cancer either before or directly after medical imaging confirms lesions in the brain, the lesions will likely be presumed to be BMs, as the probability of concurrently developing a primary brain cancer or benign brain lesion alongside an extracranial primary cancer is low, though not impossible (possibly up to a 10% chance) [10, 32, 33]. Furthermore, primary brain cancers very rarely metastasize extracranially, and so the possibility of extracranial cancer being metastatic spread from a primary brain cancer is extremely low [34]. Brain lesion pathology is therefore most useful in cases where brain lesions are detected but no extracranial cancer is known of, as pathology can quickly determine if the lesions are BMs and if so, where in the body they originated from [32]. There is also evidence that BMs can have distinct genetic mutations compared to the primary cancer they metastasized from, with some of these mutations being specifically targetable for treatment, and so BM specific pathology may be useful even when the primary cancer is well characterized [35].

Once a tissue sample is collected, it is processed and assessed to provide information on the cancerous malignancy present. Tissue samples are typically fixed, mounted, thin sectioned, and then stained to provide cell morphology and immunohistochemistry analysis. Hematoxylin and eosin stain is most commonly used to provide visual contrast for cell mor-
phology, while immunohistochemistry stains provide visual contrast based on whether cells express specific antibody receptors. Expression of these receptors can determine the cancer’s primary site and possibly inform the use of targeted therapies [32]. “Next-generation sequencing” now allows for partial or full sequencing of BM deoxyribonucleic acid (DNA), but these techniques are not yet in widespread clinical use [19, 36].

Brain lesion samples can be acquired either from post-resection surgical specimens or from specifically ordered core biopsies. As will be explained in section 1.3.3, surgical resection of brain lesions may be performed, in which case tissue from the resected lesion is available for analysis. Not all patients receive surgical resection for a variety of reasons (see section 1.3.3), and so a core biopsy would be required to retrieve a tissue sample. Core biopsies of brain lesions are gathered using image-guided stereotactic techniques to minimize the invasiveness of the procedure [37]. Core biopsies do come with inherent risks to the patient, especially for lesions that are seated deep within the brain, and so biopsies are not routinely performed on all patients [37].

1.3 Treatment techniques

After a BM diagnosis is made, a variety of treatment options are available to patients to both manage symptoms and to halt or reverse BM growth. As summarized in table 1.3, the primary treatment techniques used clinically are acute symptom management (in particular, corticosteroids), systemic therapies, surgical resection, whole brain radiation therapy (WBRT), and stereotactic radiosurgery (SRS)/stereotactic radiation therapy (SRT). Choosing which treatments to use for a given patient or BM is a challenging task, but recent guideline documents to aid clinicians and their patients have been provided by the European Association of Neuro-Oncology and European Society for Medical Oncology (EANO-ESMO) in 2021 [32] and the American Society of Clinical Oncology, Society for Neuro-Oncology, and American Society for Radiation Oncology (ASCO-SNO-ASTRO)
in 2022 [38], with additional radiation therapy (RT) specific guidance provided by ASTRO [39]. These treatment guidelines synthesize the available evidence on treatments’ effectiveness and toxicities to provide generalized recommendations on which treatments to employ in different clinical scenarios (further details in section 1.3.7).

Table 1.3: Summarization and comparison of the most commonly employed treatment techniques for BMs.

<table>
<thead>
<tr>
<th>Treatment Aspect</th>
<th>Acute symptom management (corticosteroids)</th>
<th>Surgical resection</th>
<th>WBRT</th>
<th>SRS</th>
<th>SRT</th>
<th>Systemic therapy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rapid relief of acute symptoms/mass effect</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Effective control of targeted metastases</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td></td>
<td>variable effectiveness</td>
</tr>
<tr>
<td>Effective control of new metastases</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td></td>
<td>✗</td>
<td>variable effectiveness</td>
</tr>
<tr>
<td>Viable for extensive metastases</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓ (see “Other advantages”)</td>
</tr>
<tr>
<td>Viable for large metastases</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Non-Invasive</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Short-term Toxicities/Complications</td>
<td>weight gain, GI irritation, infection</td>
<td>partial resection, infection, cognitive losses</td>
<td>fatigue, cerebral inflammation/swelling</td>
<td>fatigue, cerebral inflammation/swelling</td>
<td>varied in type and severity</td>
<td></td>
</tr>
<tr>
<td>Long-term Toxicities/Complications</td>
<td>cannot be used long-term</td>
<td>leptomeningeal disease</td>
<td>cognitive decline (likely), radionecrosis</td>
<td>cognitive decline (less likely), radionecrosis</td>
<td>varied in type and severity</td>
<td></td>
</tr>
<tr>
<td>Time course</td>
<td>weeks</td>
<td>hours (surgery), days (recovery)</td>
<td>weeks (1-2 hours/day)</td>
<td>1-5 days (1-2 hours/day)</td>
<td>weeks</td>
<td></td>
</tr>
<tr>
<td>Patient inclusion</td>
<td>all</td>
<td>favourable prognosis, metastasis in accessible location, no surgical contraindications</td>
<td>all</td>
<td>favourable prognosis</td>
<td>primary cancer site and molecular markers are amendable to a systemic agent</td>
<td></td>
</tr>
<tr>
<td>Other advantages</td>
<td>no visits to clinic required</td>
<td>pathology tissue sample can be acquired</td>
<td>relatively simple technique, memantine/hippocampal avoidance can reduce cognitive decline risk</td>
<td>more viable for extensive metastases when using single iso-centre approach</td>
<td>primary cancer and extracranial metastases may also see effect</td>
<td></td>
</tr>
<tr>
<td>Other disadvantages</td>
<td>none</td>
<td>adjuvant radiation required</td>
<td>none</td>
<td>more complex technique</td>
<td>effectiveness depends on cancer biomarkers</td>
<td></td>
</tr>
</tbody>
</table>

Table 1.3: Summarization and comparison of the most commonly employed treatment techniques for BMs.
1.3.1 Acute symptom management

If a BM patient presents with acute neurological symptoms, it is critical to first manage these symptoms to restore the patient’s QOL. As described previously in section 1.2.1, the mass effect of BMs and edema on the brain tissue lead to a variety of symptoms, some of which are quite acute, and so they must be treated as quickly as possible. Severe headaches are common, with different strengths of painkillers used to combat them, though prolonged use of these painkillers is not recommended [26, 40]. Seizures may also occur, in which case antiepileptic drugs (AEDs) may be prescribed. Levetiracetam is the most commonly employed AED, but comes with possible psychiatric side-effects to the patient [25, 26, 41].

Corticosteroids

Symptoms caused by edema mass effect can be counter-acted with corticosteroids [42]. These anti-inflammatory drugs limit the amount of liquid passing through the blood-brain barrier, therefore lowering edema mass effect and associated symptoms [26]. The corticosteroid dexamethasone is commonly prescribed for BM patients, though its use may lead to GI irritation, weight gain, and a higher susceptibility to pneumocystis pneumonia [25, 26, 42]. The risk of these adverse reactions increases with the length of use, and so a corticosteroid course is typically limited to two weeks, with the patient being gradually tapered off [42, 43]. While highly effective in the short-term, the inability of corticosteroids to be administered for prolonged periods requires patients to have their symptomatic BMs directly managed by eliminating the cancerous cells.

1.3.2 Systemic therapy

Systemic therapy is one method for eliminating cancer cells. The first systemic agents for cancer were cytotoxic chemotherapy agents, though they were seen as largely ineffective against BMs given their limited ability to cross the blood-brain barrier and poor therapeu-
tic action against BMs [44]. More recently, the development of targeted molecular and immunotherapy systemic agents for primary cancers have found a role in the treatment of BMs, with their use now being recommended as a possible first-line treatment [32, 38]. Systemic agents have the distinct advantage over surgical and RT techniques of providing treatment effect concurrently to primary cancers, extracranial metastases, and BMs, but this is a double-edged sword, as treatment toxicities are similarly far reaching.

Cytotoxic chemotherapy agents negatively effect a cell’s ability to divide and replicate, and are therefore preferentially toxic to all cells that divide rapidly, including cancer cells. Chemotherapy’s effectiveness against BMs is mixed, with poor blood-brain barrier penetration, chemotherapy resistance, and intolerable toxicity all playing a part. Currently, few recommendations for the specific usage of chemotherapy exist. The EANO-ESMO guidelines recommend the usage of capecitabine, eribulin or carboplatin for breast cancer patients negative for the HER2 biomarker, and platinum analogs such as cisplatin for SCLC [32], while the ASCO-SNO-ASTRO guidelines only recommend (with weak evidence) the use of capecitabine for HER2-negative breast cancer [38]. Temozolomide is a chemotherapy agent used commonly for primary brain cancers [45], but studies for its use with BMs have found mixed results [38, 46].

Targeted molecular agents such as tyrosine kinase inhibitors (TKIs) and monoclonal antibodies (MABs) were originally developed for primary cancers to disrupt cancer-specific biological mechanisms. TKIs and MABs developed for specific targets related to a given primary cancer are generally found to be effective against the BMs from the same primary cancer [32]. Current EANO-ESMO and ASCO-SNO-ASTRO guidelines include the usage of the MAB bevacizumab and various TKIs or other inhibitors with usage dictated by primary cancer site (breast, melanoma, or NSCLC) and specific molecular markers (epidermal growth factor receptor [EGFR], anaplastic lymphoma kinase [ALK], B-RaF protein [BRAF], and HER2) [32, 38].

Lastly, immunotherapy agents activate the immune system against cancer cells across
the entire body, including in the brain against BMs. Immunotherapy agents operate by either decreasing the effectiveness of cancer cells’ ability to evade the immune system (such as checkpoint inhibitors), or by modifying immune cells to better detect cancer cells (for example, chimeric antigen receptor T cell therapy). MAB immunotherapy agents developed for breast cancer (trastuzumab), melanoma (ipilimumab and nivolumab), and NSCLC (pembrolizumab) have recently been found to be effective against BMs, and have been recommended for use as a possible first-line treatments [32, 38].

1.3.3 Surgical resection

BM surgical resection seeks to remove the targeted BM and alleviate the surrounding edema to provide BM management and acute symptom relief. Access to the brain is typically achieved through a craniotomy, in which a portion of the patient’s skull is removed and then replaced after surgery [47]. Minimally invasive craniotomies are becoming more common, in which the size of the craniotomy is minimized to a 2–5 cm hole through which “keyhole surgery” is performed [48–50]. Surgical resection occurs microscopically, with operating microscopes being the traditional mode of imaging, but endoscopes (camera inserted into the body) and exoscopes (telephoto camera aimed from outside the body) are used for minimally invasive techniques [23]. Before surgery, neuronavigation plans are developed using CT or MR images to plot the best trajectories for surgical instrument insertion through the brain matter [23, 48, 51]. During surgery, stereotactic systems allow for precise 3D localization of surgical instruments to follow the neuronavigation plan, using either framed (instruments and patient attached to a common rigid frame) or frameless (instruments tracked in 3D space) systems [23]. Interoperative CT or MR imaging has also begun to see some use, as neurological anatomy can alter during surgery [48].

Once access to the BM is achieved, it is then resected. BMs are typically non or minimally (< 1 mm) infiltrative into the brain tissue and so can thought to be “pseudo-encapsulated” [23, 24]. This is not guaranteed, however, and so an expansion margin of up
to 5 mm may be employed to ensure complete resection [48]. Ultrasound and fluorescence imaging, as well as rapid pathology using tissue freezing, all aid in ensuring total resection [23, 48]. Small BMs can be removed intact through en bloc resection, but large BMs may require piecemeal resection [23, 38, 52].

Surgical resection is typically a well-tolerated treatment option for BMs with a high success rate and relatively low complication rate. Total resection is usually possible, though sometimes only a partial resection can be performed, especially near critical brain structures. Adjuvant RT to the surgical cavity can aid in the case of a partial resection, but RT is also used to generally to treat microscopic disease and possible leptomeningeal spread (further details in section 1.3.6). A reported 15–25% of resected BMs may recur within the resection cavity [53–56], and 5–10% of resections can result in subsequent leptomeningeal disease along the route of the surgical instruments [55, 57]. General BM surgical resection complication rates are between 4–15%, with up to an 8% rate of major complications requiring prolonged hospitalization, up to 10% incidence of neurological complications (largely dependent on BM location), 1% risk of infection, and potentially a 3% risk of mortality within a month of surgery [52, 58–60].

1.3.4 Radiation Therapy

The primary alternative to surgically resecting BMs is RT, which uses high energy radiation to halt or reverse the growth of cancerous tumours. The radiation typically consists of photon beams produced by a radioactive source or linear accelerator (linac). Radioactive sources, such as cobalt-60 (Co-60), are produced in nuclear reactors and naturally produce high energy photons through the $\gamma$ radioactive decay process. Linacs instead accelerate electrons to near the speed of light at which point they bombard a metal target to produce x-rays. In both cases, the photons are delivered from many positions around the patient and are precisely collimated to produce a desired distribution of radiation dose within the patient that conformally targets the tumour. To facilitate planning the dose distribution
Figure 1.4: Schematic representation of how photon RT results in cell damage. The high energy photon provides a vehicle for the delivery of energy deep within a patient’s tissue, but as can be seen, it is indirectly ionizing radiation, requiring a transfer of energy to a charged particle (e.g. an electron) to cause ionization. The interaction between the photon and matter within the patient causes some or all of the photon’s energy to be transferred to the electron as kinetic energy, which the electron gradually deposits as dose along its traveled path. This deposited dose can then damage DNA through direct (physical) or indirect (chemical) action that possibly causes cell death, if the damage is extensive enough that it cannot be repaired by the cell.

for BM RT, x-ray CT and MRI scans are performed pre-treatment. The MRI is used for defining the BM targets to be treated, while the data in the x-ray CT precisely represents the patient’s position that can be replicated during treatment and allows for modeling of how and where the photon beams will deposit dose. By registering the MRI to the CT, the BM targets can be transferred to the CT and an effective radiation therapy plan developed.

The photons delivered during RT are of high enough energy to be classified as ionizing radiation, implying their ability to damage DNA. When the high-energy photons enter the patient’s tissue, the photons interact with the tissue through of variety of processes that both scatter the photons and transfer energy to produce electrons with high kinetic energy, as shown in figure 1.4. The amount of energy deposited within a given mass of tissue is measured in units of Gray (Gy), with 1 Gy = 1 J/kg. The deposited dose damages cells’ DNA molecules directly through physical interactions or indirectly through the production of highly reactive oxygen chemical species. Some damage to the DNA can be repaired through cellular processes, though in RT the goal is to cause extensive DNA damage within the cancerous cells such that cell death occurs.

In RT, the bulk of the patient’s healthy tissue can avoid toxicity by simply collimating
Radiation Fraction

Reoxygenation

Radiation Fraction

Vascularized Outer Region
(More Radiosensitive)

Necrotic/Hypoxic Core
(Less Radiosensitive)

Figure 1.5: Reoxygenation mechanism in fractionated RT. The outer regions of the tumour are well-vascularized, and therefore well-oxygenated and radiosensitized. The first radiation fraction preferentially damages this outer region, resulting in its removal. The remainder of the tumour is now free to be reoxygenated, sensitizing it to the next radiation fraction.

the radiation source to not deposit dose where it is unneeded. Healthy tissue closer to a targeted tumour cannot be spared in this manner, and so biological differences between cancerous and healthy cells are exploited. By “fractionating” the delivery of radiation over multiple sessions (typically ≥ 10 fractions), cancer cells can be preferentially damaged over healthy tissue, when compared to using a single fraction. One specific biological mechanism exploited by fractionation is reoxygenation. As explained previously, large tumours can outstrip their blood supply and form inner hypoxic regions (section 1.1). Oxygen is a powerful radiosensitizer due to its chemical role in the indirect damage of DNA, and therefore these hypoxic regions are less radiosensitive [61]. By fractionating the radiation delivery, the well-oxygenated areas of the tumour can be destroyed by an initial fraction, allowing the hypoxic regions to be reoxygenated and become sensitized to the next fraction [62] (see figure 1.5).

Whole brain radiation therapy

The simplest RT technique for BMs is WBRT, which was first used in 1931 [63]. As the name suggests, this technique involves targeting the entire organ of the brain with a relatively uniform radiation dose, regardless of BM or healthy tissue location (see figure 1.6a). A wide range of WBRT dose and fractionation schemes are used, but 20 Gy delivered in 5
fractions of 4 Gy each, and 30 Gy in 10 fractions are the most common, with little variation in treatment outcomes seen between schemes [64]. While WBRT is typically used after a BM diagnosis, prophylactic cranial irradiation (PCI) has been explored for SCLC, which is highly prone to BM formation [65, 66]. PCI does prevent BM formation, but the associated risks and side-effects of the radiation leaves PCI receiving mixed recommendations for routine clinical use [32, 38].

WBRT is delivered via a general purpose linac. Using radiation beams collimated to conform to the shape of the brain is the most basic technique for delivering WBRT, but more recent techniques such as intensity-modulated RT (IMRT) and volumetric modular arc therapy (VMAT) feature computerized optimized beam collimation over multiple beam angles to make the dose distribution more homogeneous [68, 69]. IMRT and VMAT techniques therefore also allow organs-at-risk (OARs) for radiation damage to have their received radiation dose minimized. Hippocampal avoidance WBRT (HA-WBRT) is commonly used to the spare the neurocognitively critical hippocampal region from dose (as long as no BMs are present), and has been found to do so while maintaining BM control when compared to traditional WBRT [70, 71] (see figure 1.6b for an example HA-WBRT dose distribution). Avoidance of the similarly sensitive brainstem, salivary glands, optic nerve, and auditory
structures is also crucial to increase patient QOL beyond only neurocognition [72–75].

As the entire brain is targeted, WBRT provides control of both visible and microscopic BMs, but also leads to short and long-term side-effects and toxicities. Common short-term side-effects include fatigue and cerebral inflammation and swelling, though these side-effects are often transitive and well-controlled by short courses of corticosteroids [26]. Long-term side-effects due to neural and glial cell death include necrosis and neurocognitive decline [26, 76, 77], with these long-term side-effects of important concern as patient OS has lengthened. To counter these effects, memantine, a pharmacological agent typically prescribed for Alzheimer’s disease, has now been recommended to be given alongside WBRT to prevent neurocognitive decline [38, 39, 78].

**Stereotactic radiosurgery and radiation therapy**

An increasingly popular alternative RT technique to WBRT is SRS, in which only the BMs are targeted with radiation and the healthy brain tissue is spared as much as possible (as shown in figure 1.6c). SRS was pioneered in 1967 with the development of the Gamma Knife device by neurosurgeon Dr. Leskell, who saw precisely targeted radiation doses as a possible technique to avoid invasive surgery of brain lesions, both cancerous and benign [79]. The term “stereotactic” within SRS refers to the ability of the radiation beams to be precisely aimed with respect to a rigid, well-defined, patient-based coordinate system, similarly to how the term was used previously to describe surgical techniques (section 1.3.3). SRS adopted the same patient-affixed stereotactic frames from neurosurgery at first, though modern x-ray image-guidance has allowed for frameless SRS with removable thermoplastic masks to be used for patient fixation without a loss in accuracy [80]. The original and current Gamma Knife models contain multiple Co-60 radioactive sources coincident on a single isocentre to achieve dose distributions that are highly conformal to BMs. More recently, linacs (both general purpose models and specialty robotic arm CyberKnife units) have also been used to deliver SRS. While dose distributions between treatment modali-
ties can vary, all offer comparable BM conformality and treatment outcomes [81, 82]. The large beam field size offered by general purpose linacs combined with IMRT and VMAT techniques has allowed for multiple BMs to possibly be treated simultaneously with a single isocentre, whereas specialty modalities require one isocentre per BM, leading to longer treatment times of multiple BMs [82].

The dose conformality to targeted BMs is one hallmark of SRS, with the other being the delivery of high radiation doses in one to three fractions (though the precise number of fractions used to define SRS differs in the medical community). This results in cancerous cells receiving ablative doses, in which the previously discussed principles of radiation’s effect on cells still hold, but cell kill may also be driven by damage to tumour vasculature and post-treatment immune response [83–85]. SRS’s few number of fractions and possible induction of vasculature damage also cause the previously discussed reoxygenation effect to be non-existent or limited, and so the presence of hypoxia within a treated BM may decrease SRS effectiveness [84, 86, 87]. The use of ablative doses does present the possibility of greater toxicity to healthy tissue, and so the high dose conformality of SRS is critical for toxicity control, as the previous benefits of fractionation in WBRT are no longer present. To this end, when SRS is prescribed, the target treatment volume is only the gross tumour volume (GTV), which is defined using the enhancement present in the T1w-CE MRI. No clinical treatment volume (CTV) expansion to account for microscopic disease is used, as the BM is assumed to be non or minimally invasive and the dose immediately surrounding the GTV is still quite high due to gradual dose fall-off [24, 88]. A minimal planning treatment volume (PTV) expansion of typically 0–2 mm is applied to the GTV to account for uncertainties in patient positioning and treatment delivery [89].

While SRS attempts to minimize overall risk of toxicities to brain tissue, the healthy tissue directly surrounding the targeted BMs may still receive a large dose of radiation that can cause tissue injury. Similarly to WBRT, acute edema may form shortly after SRS that may require management with corticosteroids [26]. In the longer term, early-delayed and
late-delayed adverse radiation effect (ARE) may occur [90, 91]. Early-delayed ARE occurs typically around three months after treatment with presenting symptoms consisting of fatigue and drowsiness [26, 90]. Early-delayed ARE is transient, however, resolving spontaneously without treatment after a few months [91]. Late-delayed ARE, as the name suggests, often occurs later than early-delayed ARE and consists typically of radionecrosis (RN), necrotic lesions that form due to radiation damage [90]. RN effects are not transient and may present with the same symptoms as a BM, requiring management with corticosteroids, vascular endothelial growth factor (VEGF) inhibitors such as bevacizumab, hyperbaric oxygen, anticoagulants, or possibly surgical intervention via resection or laser interstitial thermal therapy (LITT) [26].

ARE and RN can negatively affect a BM patient’s QOL after receiving SRS, but these radiation injuries also complicate response assessment of BMs to SRS via traditional MRI. As shown in figure 1.7, there are many possible outcomes to any SRS treatment for a given BM, some of which appear near identical in MRI. First, a BM may have all viable cancerous cells destroyed, and the BM will shrink or disappear completely (figure 1.7a) or possibly appear to grow and then shrink in serial MRI without further treatment due to transient early-delayed ARE (figure 1.7b). Figure 1.7c shows how some cancerous cells may remain viable within a BM post-SRS, resulting in BM recurrence and regrowth, and therefore treatment failure. Figure 1.7d also shows regrowth, but due to the formation of a RN lesion instead of cancerous cells. As can be seen, figure 1.7b–d all show a similar BM size trajectory at the first follow-up MRI, and figure 1.7c,d continue this similarity to future follow-ups. This mimicking of true cancerous progression by ARE and RN is known as “pseudo-progression”, and it vastly complicates decision-making, as the best approach for each treatment outcome varies considerably (ARE requires only observation, while RN and cancerous progression treatments have been separately described previously) [90]. Traditional MRI cannot readily distinguish true from pseudo-progression, with more advanced MRI sequences or PET imaging offering enhanced, but not definitive, insights.
Figure 1.7: Schematic representation of five generalizations of how a BM can respond to SRS (a–e) and how the BM could present in follow-up MRI. This schematic broadly simplifies the possible outcomes, and the appearance and depiction of the different tissues within each lesion are intended to be for illustrative purposes only. Furthermore, the timing of lesion growth and shrinkage is conveniently synchronized in this figure, again to illustrate the similarity in lesion size across scenarios b–e.

[28, 30, 92]. To complicate matters further, multiple treatment outcomes may be present within a single BM (figure 1.7e), and so not even stereotactic biopsy of the BM may be entirely conclusive [93].

SRS dose and fractionation schemes are based on balancing BM control and resulting toxicities (most commonly RN, but damage to the optic nerves and chiasm, brainstem,
Table 1.4: Summarization of the HyTEC guidelines for BM SRS/SRT control rates (a) and RN toxicity risk (b), grouped by number of fractions. \( V_x \text{Gy} \) is defined as the volume of tissue (either within the BM or healthy brain) that receives \( \geq x \text{Gy} \).

<table>
<thead>
<tr>
<th>Number of Fractions</th>
<th>Dose (Gy)</th>
<th>BM Diameter (cm)</th>
<th>Control Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>18–24</td>
<td>\leq 2</td>
<td>85–90</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>&gt; 2</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>24–30</td>
<td>2–3</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>21–27</td>
<td>&gt; 3</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>30–35</td>
<td>2–3</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>25–30</td>
<td>&gt; 3</td>
<td>75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Fractions</th>
<th>Dose-Volume Limit</th>
<th>RN Risk (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5 cc</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>15 cc</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>10 cc</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>30 cc</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>20 cc</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>30 cc</td>
<td>20</td>
</tr>
</tbody>
</table>

and auditory structures may also be considered if the BM is nearby). SRS dose and fractionation prescriptions have been historically informed by the landmark Radiation Therapy Oncology Group (RTOG) 90-05 study, which described prescriptions dependent on BM volume [94]. More recently, the comprehensive “Quantitative Analysis of Normal Tissue Effects in the Clinic” (QUANTEC) [72–74, 95] and more specific “Hy Dose per Fraction, Hypofractionated Treatment Effects in the Clinic” (HyTEC) [89, 96, 97] guidelines provide both detailed BM control and RN toxicity probabilities for SRS dose and fractionation schemes based on compilation of numerous clinical studies’ data. As shown in table 1.4a, BM 1-year and 2-year control rates from one or three fraction SRS, and also five-fraction SRS (defined here as SRT), are dependent on the BM size and prescribed dose. The RN toxicity risk profile also relies on the number of fractions used, along with the volume of tissue (either normal or cancerous) irradiated to 12, 20, or 24 Gy (see table 1.4b). Since larger BMs inherently have a large volume of tissue receiving a high dose, but also feature dose distributions with shallower falloff and less conformality, the risk of RN necessitates a less aggressive prescription for larger BMs by prescribing more fractions or a lower dose.

### 1.3.5 Non-standard therapies

In addition to the standard systemic, surgical, and RT treatments available to BM patients, there are some other treatment modalities that are not in widespread use. First, LITT can
also be used as a first-line treatment of BMs in addition to RN lesions [98]. Second, brachytherapy features placing a radioactive source within the BM to deliver a radiation dose to the BM, instead of from external beams [23]. In both treatments, similar surgical techniques used for resection are employed to gain physical access to the BM, and then either the laser or radioactive source is inserted and the treatment delivered.

### 1.3.6 Combinations of treatments

Surgery, WBRT, and SRS/SRT comprise the three primary techniques for treating BMs. These treatments can be combined together as adjuvant treatments, or possibly as salvage treatments if there is a treatment failure or formation of new BMs after treatment [32]. Surgical resection can be used independently, but as previously described (section 1.3.3), cancerous cells can remain in the surgical cavity, and so post-resection RT is used to lower rates of BM recurrence. WBRT has been used after surgery, but using SRS/SRT as an adjuvant is now recommended given its toxicity sparing [39]. Pre-resection SRS/SRT has also been explored since it allows for more simplified delineation of the target, as it is an intact BM instead of surgical cavity [99, 100]. SRS/SRT can also be combined with WBRT, with SRS/SRT being delivered either before or after WBRT. Another option is to combine SRS/SRT and WBRT during treatment delivery as a simultaneous in-field boost (SIB) using IMRT/VMAT treatment planning (see figure 1.6d) [101].

Systemic therapies are used extensively for various primary cancers, including now for BMs, and so their interaction and combination with other BM treatments must be considered. Many chemotherapy and immunotherapy agents are radiosensitizers, and so it is important to consider their effects on healthy tissue outside of the blood-brain barrier that may be exposed during RT, such the skin or mucosal linings [46]. Immunotherapy is possibly synergistic with SRS techniques, with the high ablative doses hypothesized to result in cancer cell material being released and subsequently used by the activated immune system to target cancer cells [102]. This effect must be balanced, however, against the
generally immune suppressive effects of RT and the radiosensitization of healthy tissues by immunotherapy [46].

1.3.7 Comparison and selection of treatments

Given the treatment options for BM patients, optimal treatment choice is critical to balance symptom management, tumour control, and patient QOL. To this end, treatment guidelines (such as those from EANO-ESMO and ASCO-SNO-ASTRO), prognostic models, multi-disciplinary consultation of clinicians, and importantly patient empowerment and involvement, all aid in making these complex clinical decisions.

One treatment question that has been thoroughly investigated is the use of WBRT in modern BM management. SRS provides similar or slightly enhanced control of visible/targeted BMs (in-field control) compared to WBRT [77, 103]. WBRT does provide enhanced control of new BM formation (out-of-field control) compared to SRS, given its treatment of the entire brain, but this comes at the cost of significantly higher risk of neuropsychological decline and no benefit to patient OS [76, 77, 104, 105]. Therefore, WBRT is generally preferred for use with patients with a poor prognosis, low performance status, or an extensive number of BMs (> 4–10 BMs, though > 10 BM SRS is possible at some centres) [32, 38, 39] (see table 1.5). For patients that are palliative (< 3 months OS), WBRT with supportive care has been shown to offer no benefit over supportive care alone [106], and so WBRT is only recommended if required for symptom management [39].

Another key question in BM management is whether to use surgical resection or SRS/SRT. Surgical resection uniquely rapidly relieves symptoms caused by BM mass effect as well as can provide pathology samples for BM patients with no primary cancer diagnosis [32], while SRS/SRT remains the only recommended treatment option for patients with a good prognosis that are not strong candidates for surgery or have BMs in surgically inaccessible locations. Beyond these clear use cases for each treatment, SRS/SRT is generally recommended for 1–4 BMs, and conditionally recommended for up to 10 BMs [32, 38, 39] (see
Table 1.5: Summarization of recommendations for BM treatment selection, based on the guidelines from ASCO-SNO-ASTRO [38, 39]. This summary simplifies the treatment recommendations, with exhaustive decision flow-charts given in the ASTRO guidelines [39]. This table shows the general uses of each treatment, with surgery used for treating BMs with mass effect, SRS/SRT preferred for patients with a good prognosis, and WBRT recommended for patients with poor prognosis (though SRS/SRT could still be considered).

table 1.5). These recommendations do have caveats, especially for the size of BMs, as for BMs above 4 cm in diameter [39] or total volume of BMs > 15 cc [32], the usage of surgery or WBRT may offer improved outcomes for a given patient.

Lastly, the guidance around using targeted molecular and immunotherapy agents is rapidly evolving. For symptomatic BMs, it is still recommended to treat BMs immediately with surgical resection or RT, but for asymptomatic BMs detected in surveillance imaging, systematic therapies have been recommended as a first-line option [32, 38]. These systemic options are currently limited to specific primary cancer types with certain targetable mutations, but in these cases systemic therapy can be used with the BMs closely monitored for progression.
1.4 Prognostic and predictive models for brain metastasis patients

The variety of BM treatments and their associated outcomes and toxicities necessitates nuanced clinician and patient decision-making to maximize patient satisfaction and QOL. In order to inform decision-making, prognostic and predictive models are crucial in forecasting likely future disease and treatment outcomes for a given patient. Prognostic models offer a generalized outlook for a given patient regardless of treatment received, while predictive models provide the likely outcome of a specific treatment [107]. In both cases, these models are provided one or more input variables for a specific patient, that then interact through a mathematical function of varying complexity to provide a prognostic or predictive output, such as OS or probability of tumour progression post-treatment.

1.4.1 Current prognostic models

For BM patients, prognostic models have been largely focused on patient OS, with such models finding clinical use. OS prognosis is helpful for BM patient treatment decision-making, as indicated by the references to using “poor” or “good” patient prognosis in the previous section on treatment selection (section 1.3.7). OS prognostic models aid the patient and their clinical team in deciding if the short or long-term toxicity risks are appropriate given the benefit offered by a specific treatment and the patient’s expected survival.

One of the first widely used BM patient OS prognostic models was produced in 1997 by Gaspar et al., consisting of a recursive partitioning analysis (RPA) model based on over 1000 BM patients treated across three RTOG clinical trials [108]. This model used patient performance status, age, presence of extracranial metastases, and if the patient’s primary cancer was controlled to produce patient prognoses. Over the next decade, multiple additional prognostic models were developed using various patient databases [58, 109–112],
eventually leading to the Graded Prognostic Assessment (GPA) prognostic model [113]. The GPA was soon replaced by the Diagnosis-Specific Graded Prognostic Assessment (DS-GPA) model [114], that featured prognostic models specific to a patient’s primary cancer type. Over the next decade, the DS-GPA model was incrementally improved into its currently form that is based on nearly 7000 patients and features individual prognostic models for the most common primary cancer types and the use of molecular markers [20].

1.4.2 Current predictive models

OS models are an important tool for clinical decision-making, but their use ultimately remains somewhat limited due to their prognostic nature. Predictive models would allow for specific outcomes and toxicities to be predicted for a given treatment, allowing for a treatment’s use to be specifically tailored for a unique patient or even a specific BM.

BM SRS/SRT has especially seen a variety of predictive models developed to predict effective control of BMs and the development of toxicities, such as RN. The previously described QUANTEC and HyTEC guidelines leverage predictive models that aid in choosing SRS dose and fractionation [89, 95, 97] (table 1.4). These guidelines are based on relatively simple mathematical logistic functions of the tumour control probability (TCP) and normal tissue complication probability (NTCP) that are fit to empirical data [89, 97] (example shown in figure 1.8). These models then predict SRS outcome (both BM control and toxicity) as a function of prescription dose or volume of tissue receiving a given a dose. While such models currently inform SRS prescription and planning, the predictions they provide are generalized and not patient or BM specific.

Having more personalized treatment decision-making for individual patients and BMs would be largely beneficial for SRS. First, while the EANO-ESMO and ASCO-SNO-ASTRO guidelines provide general guidance in treatment decision-making, they mainly take only the patient’s prognosis, number of BMs, and BM size into account directly [32, 38]. There is therefore an opportunity to potentially identify patients or BMs that
Figure 1.8: A representative comparison of TCP and NTCP curves showing the tradeoff between maximizing tumour control and minimizing toxicity. As shown, dose $D_1$ provides a small risk of toxicity (5%), but only a 50% chance of successfully treatment of the cancer. $D_2$ provides much more effective tumour control (95%), but at the cost of a higher toxicity risk (25%).

are at higher risk for SRS treatment failure or toxicity that may benefit from alternative treatments. Second, if SRS is the chosen treatment technique, individualized predictive modeling could aid during SRS prescription selection. Prediction of SRS failure and toxicity per BM could allow for a more precise choice of the SRS prescription compared to using the generalized HyTEC guidelines. Third, the SRS planning process could be guided using predictions of toxicities and the need for re-treatment, which could indicate the acceptable level of dose OARs should receive.

In order for predictive models to provide more personalized and precise SRS treatment predictions, the models must incorporate input data from the patient and each BM being treated. Patient-specific data is usually readily available, such as primary cancer site and pathology data. Previous studies have investigated patient specific data, finding that primary cancer site may impact SRS outcomes [97, 115], but the data is not conclusive and has not produced a predictive model.

BM-specific data is more difficult to collect and incorporate into predictive models. Stereotactic biopsy of BMs would provide rich pathology data, but as described previously, such biopsies are invasive, not feasible for all patients or BMs, and sample only a portion of a BM’s tissue (section 1.2.3). While primary cancer pathology is more readily accessible,
it is not fully representative of a BM’s pathology since the cancer can further mutate during BM formation and because the presence of necrosis, hypoxia, or cystic components specific to each BM would be unknown.

Medical imaging provides an alternative method to gather BM-specific data that is non-invasive, readily available, and comprehensive of the entire BM. Non-CE CT is typically used in modern SRS for dose calculation purposes, but given non-CE CTs limited tissue contrast, little BM information would be expected from these images. PET can offer biological functional data for BMs [30, 116], but PET is not routinely accessible for BM patients. MRI is therefore the optimal imaging modality to provide BM-specific data, as it is routinely performed before SRS for BM delineation and also offers a rich dataset through multiple MRI sequences and CE scans.

Previous studies have examined connections between the qualitative appearance of BMs in pre-treatment T1w-CE MRI and post-SRS progression [117–120], with some studies using pre-treatment CE-CT [121] or investigating post-WBRT progression [122]. These studies qualitatively labelled the appearance of BM enhancement in T1w-CE MRI or CE-CT as “homogeneous”, “heterogeneous”, or “ring-enhancing”, and then examined if SRS or WBRT BM control rates differed across the three appearances. In general, the studies found that “homogeneous” BMs had the lowest risk of progression post-SRS, followed by “heterogeneous” BMs, and then “ring-enhancing” BMs. It was hypothesized that since homogeneous enhancement of BMs indicates uniform vascularization and therefore uniform oxygenation, the BM would be less likely to progress post-SRS, as oxygen is a known radiosensitizer. Conversely, heterogeneous BMs would be less uniformly oxygenated and therefore less radiosensitive, while ring-enhancing BMs may have a necrotic/hypoxic core, and therefore would be the least radiosensitive and most likely to progress post-SRS [118, 119]. While these studies discovered and described the connection between BM appearance and SRS or WBRT outcome, they did not propose predictive models based on their results.
A further study by Rodrigues et al. [123] did integrate BM qualitative appearance into a predictive model of post-SRS progression. This study examined a more comprehensive and biologically-relevant five-way appearance labelling of “homogeneous”, “heterogeneous”, “cystic (simple)”, “cystic (complex)”, and “necrotic” BMs, as shown in figure 1.9. These labels were then combined with a diverse set of non-imaging clinical predictors that were analyzed using RPA to produce a post-SRS progression risk stratification model. The RPA found that BM appearance and the SRS prescription were the two most relevant features to use in the model, with heterogeneous or necrotic BMs receiving a less aggressive SRS prescription being at the highest risk of post-SRS progression.

### 1.4.3 Machine learning techniques

An alternative approach to qualitative analysis of BM MRI to predict SRS outcomes is to apply computerized quantitative analysis of digital medical images. The prognostic and predictive models discussed thus far have employed statistical techniques utilizing relatively few variables that interact in simple mathematical relationships to produce predictions. Such techniques are incapable of utilizing the large amount of quantitative data within medical images, and so machine learning (ML) techniques are required.

ML is the process of using large amounts of data to “teach” a computer to perform a task. For applications in predictive or prognostic modeling, the subset of ML called “supervised ML” is most commonly used. In supervised ML, a large dataset is first collected.
Figure 1.10: Schematic representation of ML (a) dataset splitting, (b) model training, and (c) model testing. The creation of the validation dataset in (a) and the optimization of the features and hyperparameters in (b) are not necessary, but typically are performed to increase model accuracy. If the validation dataset is not required, the “Training Dataset” in (b) would be the larger version from the first dataset split in (a).

consisting of individual samples, each with the same set of predictive variables or “features” and known outcome that is being predicted (see figure 1.10a). A ML model is then “trained” by having an algorithm optimize the ML model’s parameters to define relationships between the dataset’s features and outcome, using a portion of the dataset known as the “training dataset” (figure 1.10b).

ML models are highly complex, with hundreds to millions of parameters to be opti-
imized, and the ability to combine these parameters through complex mathematical relationships. While this model complexity allows for possibly more accurate predictions, it also allows for possible overfitting to the training dataset. It is therefore critical to “test” any ML model with a portion of the dataset not used during model training, known as the “testing dataset”. In model testing, only the predictive features from the testing dataset's samples are provided as model inputs, and then the trained model produces a predicted outcome for each sample. These predicted outcomes are then compared to the ground truth outcomes from the testing dataset to compute model error metrics (see figure 1.10c).

As shown in figure 1.11, overfitting leads to poor model accuracy on the testing dataset. Of equal concern, is underfitting, in which the model is not complex enough to properly fit to the training dataset. Balancing underfitting and overfitting (also called the “bias-variance tradeoff”) must be done carefully to produce an optimal, well-fit model.

To optimize model performance, the training dataset may also be used to optimize the set of features used or the design details of the model, known as hyperparameters (see figure 1.10b). To accomplish this optimization, a portion of the training dataset may be used as a “validation dataset” to ensure the optimization of features or hyperparameters is not causing overfitting. After the optimization is complete, the unseen testing dataset is still used to evaluate the model.

While the testing dataset provides an adequate description of model performance for the ML dataset’s population sample, it is vital that the model is also evaluated on an external population sample, known as an external validation or testing dataset. In medical applications, the training and testing dataset is typically from one medical centre, while the external testing dataset is from a second medical centre to ensure the model’s performance is not specific to the centre’s dataset on which it was trained.
Figure 1.11: Representation of underfit, well-fit, and overfit classifier models for the same training and testing datasets. The classification model is shown as the decision boundary between the positive and negative samples plotted for the two features in the dataset. The underfit model (high bias, low variance) lacks the number of parameters required to properly fit to the training dataset, leading to five misclassifications on the testing dataset. The overfit model (low bias, high variance) has too many parameters, allowing the training dataset to overly influence the decision boundary location, leading to three misclassifications on the testing dataset. The well-fit model shows the appropriate trade-off between the two previous models to achieve optimal accuracy on the testing dataset (low bias and low variance).

1.4.4 Machine learning models and radiomics

How supervised ML models make predictions can be illustrated using the naïve Bayes classifier (NBC). When applied to a binary classification task of whether a not a sample is predicted to be positive or negative for some outcome, the NBC fits a Gaussian distribution (or other probabilistic distribution) to the training dataset’s positive samples’ feature val-
Figure 1.12: Example of a NBC differentiating two classes using a single feature. The training dataset consists of 50 samples (25 per class), plotted as circles. Gaussian probability density functions are then fit to these samples and plotted. The decision boundary for the classifier is shown, with samples having feature values left of the boundary being classified as “negative”, and samples to the right being classified as “positive”. As the Gaussians intersect above 0, the positive and negative samples are non-separable when using only the single feature shown, and so classification error is unavoidable. Since each class had the same number of samples, the prior probabilities are assumed to be equal and so were not applied for simplicity.

values, and then another Gaussian to the negative samples, as shown in figure 1.12. To then classify a new sample, if the “positive” Gaussian distribution is greater for the sample’s feature value than the “negative” distribution, then sample is predicted to be positive, and vice versa for a sample predicted to be negative. The operation of the NBC demonstrates the ultimate task of supervised ML: to find the probabilistic distributions of positive and negative samples for a given set of features and to use these probabilistic distributions to make predictions. In many applications, simple parametric probability distributions do not adequately describe the true distributions, and so more advanced models are required.

The random decision forest (RDF) model offers a more advanced method to find the samples’ probabilistic distributions. The RDF consists, as its name suggests, of many individual decision trees. For each tree, a random selection or “bag” of samples is taken from the training dataset, and from this bag, a random set of the predictor features are also selected for use. Using binary classification once more as an example, a decision tree is then constructed to optimally classify the random bag’s samples as positive or negative. This is done by first finding the feature and value of that feature that best separates the positive
and negative samples, providing the first split in the decision tree. For each branch of this first split, the process is repeated using only the samples in the branch to create further splits. This is continued until all the positive and negative samples are fully separated at the bottom “leaves” of the decision tree.

After a set of $n$ unique decision trees are created in the forest, a new sample can be classified as positive or negative by performing each decision tree’s classification for the new sample, which results in $n$ votes of either “positive” or “negative”. The model then outputs the proportion of trees that classified the sample as positive as the predicted probability the sample is positive. RDFs provide high predictive model complexity, natively support non-numerical categorical variables, and provide relative importance scores for the features used. The training dataset samples not selected to be in each decision tree’s random bag (called the “out-of-bag samples”) also conveniently provide a validation dataset for hyperparameter optimization.

The NBC and RDF both represent “classical” ML models which are capable of utilizing hundreds of features simultaneously to make predictions. Integrating digital medical images into these models poses a problem, however, as even a modestly-sized 3D medical image may contain a million individual voxels, each effectively representing a model input/feature. Classical ML models cannot be trained with such high-dimensional inputs, and so “radiomics” is used to overcome this challenge. “Radiomics” consists of extracting quantitative features from radiologic images [124–126]. In particular, radiomic features are designed each with a unique equation specifying how the feature’s value is calculated for a given grayscale image and region-of-interest (ROI) [127]. Some of these features are simple “first-order” statistical measures of the voxel values within the ROI, such as the mean or standard deviation. Other features capture the shape and size of the ROI using volume, surface area, or elongation, for example. Lastly, “second-order” texture features capture the spatial relationships between voxels through first calculating matrices dependent on the grayscale values of adjacent voxels, and then calculating features from these matrices.
Second-order gray-level (GL) matrices include the GL co-occurrence matrix (GLCM), GL Run Length Matrix (GLRLM), GL Dependence Matrix (GLDM), GL Size Zone Matrix (GLSZM), and the Neighbouring Gray Tone Difference Matrix (NGTDM). Once these radiomic features are extracted for a dataset of images, these are provided to ML models as a set of predictive features.

A newer ML technique that has been rapidly adopted in many fields is neural networks (NNs). NN models consist of multiple layers of artificial neurons, with each neuron processing multiple inputs and then producing a single output. The outputs of one neuron layer are provided as the input to another neuron layer to construct a network. In particular, NNs with many layers and millions of parameters have revolutionized the field of ML and artificial intelligence through “deep learning”. A variant of NNs called convolutional NNs (CNNs) have also been developed to directly use imaging data as input. These CNNs can be trained to effectively design their own radiomic features that are optimal for the task being performed. The use of deep learning and CNNs goes hand-in-hand with the concept of “big data”, in which datasets of millions of samples are required to train the most complex NNs. As large amounts of medical treatment outcome data is typically difficult to obtain, the use of NNs and deep learning compared to classical ML in medical applications is mixed.

### 1.4.5 Current machine learning models

ML models have been used in multiple applications for BM patients. While the DS-GPA prognostic model has widespread clinical adoption, newer OS models have also been developed using ML [128–130], including incorporating radiomic analysis [131], but they require further validation. BM radiomic analysis using classical ML and CNNs has also been applied for automated BM detection and segmentation [132–135], distinguishing BMs and primary brain cancers [136, 137], labelling BMs’ primary cancer or mutational status [138, 139], distinguishing pseudo-progression or RN from cancerous progression after
predicting treatment outcomes after WBRT [144], and predicting the formation of new BMs after SRS [145].

Predicting if a BM will respond to SRS/SRT or progress post-treatment has also been explored using radiomics and ML. As can be seen in table 1.6, using ML for this task was first reported in 2018, although this study by Cha et al. [146] had a weak ML design and used non-CE CT, and so the mechanism for outcome prediction is unclear. In late 2019 and early 2020, the first studies using more robust ML designs and MRI were reported [147–149], followed by a larger number of studies in the last two years [150–164]. The predictive performance reported by these studies is promising, but wide-ranging (area under the receiver operating characteristic [ROC] curve [AUC] 0.73–0.95).

The studies in table 1.6 all predicted BM response to SRS or SRT, but with a variety of techniques and datasets. First, the patient populations studied, while all retrospective, vary widely on exclusion and inclusion criteria, if well reported. The difference in included patient primary cancer sites is included in table 1.6, but other exclusion criteria such as performance status, previous/adjuvant treatments, and BM size were also varied. Second, the data used for radiomic analysis was quite diverse. Every study included analysis of T1w-CE MRI for the BM GTV ROI, but some studies included further MRI sequences, peri-tumoural ROIs, and dose maps, to mixed effect. Some studies also included non-imaging clinical features, but which clinical features were used and found to be predictive varied. Third, almost all studies used radiomic feature extraction and classical ML models, except for Jalalifar et al. that used CNNs pre-trained on a large BM detection dataset in their 2022 studies [157, 159]. All previous studies predicting SRS outcomes were consistently based on patients treated with Gamma Knife or CyberKnife, while the SRT studies have been based on a growing cohort of linac-treated patients at the same institution.
<table>
<thead>
<tr>
<th>Author (Year)</th>
<th>Treatment Modality</th>
<th># Patients (# BMs)</th>
<th>Patient Population</th>
<th>Scanners</th>
<th>Radiomic Features (ROIs Used)</th>
<th>Clinical Features</th>
<th>Analysis Technique</th>
<th>Key Results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Thesis Chapter 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeVries et al.</td>
<td>SRS (linac)</td>
<td>99 (123)</td>
<td>General</td>
<td>1× 1 T</td>
<td>T1wCE (BM ROI)</td>
<td>12×</td>
<td>RDF; models trained to predict progression and qualitative appearance</td>
<td>p = 0.0003 risk stratification</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4× 1.5 T</td>
<td></td>
<td></td>
<td>High qualitative appearance interobserver variability</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T1wCE (BM ROI)</td>
<td>12×</td>
<td>Radiomic models superior to 5-way qualitative appearance RPA model</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Progression radiomic signature linked to heterogeneous and necrotic BMs</td>
<td></td>
</tr>
<tr>
<td><strong>Thesis Chapter 4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeVries et al. [163] (May 2023)</td>
<td>SRS (linac)</td>
<td>Centre 1: 99 (123)</td>
<td>General</td>
<td>1× 1 T</td>
<td>T1wCE (BM ROI)</td>
<td>8×</td>
<td>RDF</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Centre 2: 62 (117)</td>
<td></td>
<td>9× 1.5 T</td>
<td></td>
<td></td>
<td>AUC 0.77, AUC 0.80 (retrained per centre)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T1wCE, T2w, DWI (BM and edema ROIs)</td>
<td>9×</td>
<td>AUC 0.70 (external test set)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AUC 0.78 (pooled data)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Harmonization of centres' features</td>
<td></td>
</tr>
<tr>
<td>Du et al. [162] (Mar. 2023)</td>
<td>SRS (Gamma Knife)</td>
<td>Centre 1: 307 (504)</td>
<td>General</td>
<td>1× 1.5 T</td>
<td>T1wCE, T2w, FLAIR</td>
<td>8×</td>
<td>Cox model</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Centre 2: 30 (55)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AUC 0.95 (AUC 0.95 independent test set)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AUC 0.93 (external test set)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Clinical and radiomic features optimal</td>
<td></td>
</tr>
<tr>
<td>Carloni et al. [161] (Jan. 2023)</td>
<td>SRS (CyberKnife)</td>
<td>148 (276)</td>
<td>NSCLC</td>
<td>&gt;2× (and field strength unknown)</td>
<td>8×</td>
<td>GNB, KNN, RDF, SVM, MLP, Adaptive boosting</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T1wCE (BM ROI)</td>
<td></td>
<td>Differences found between 2× radiomic platforms</td>
<td></td>
</tr>
<tr>
<td><strong>Thesis Chapter 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeVries et al. [160] (Dec. 2022)</td>
<td>SRS (linac)</td>
<td>99 (123)</td>
<td>General</td>
<td>1× 1 T</td>
<td>T1wCE (BM ROI)</td>
<td>12×</td>
<td>RDF</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4× 1.5 T</td>
<td></td>
<td></td>
<td>AUC 0.77</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T1wCE, T2w, T2w FLAIR</td>
<td>None</td>
<td>Clinical and radiomic features optimal</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Primary cancer and MR scanner variability impact accuracy</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>BM volume impacts accuracy and can be accounted for</td>
<td></td>
</tr>
<tr>
<td>Gutsche et al. [159] (Nov. 2022)</td>
<td>SRS (CyberKnife)</td>
<td>150 (308)</td>
<td>General</td>
<td>1× 1.5 T</td>
<td>T1wCE, T2w, T2w FLAIR</td>
<td>None</td>
<td>RDF</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1× 3 T</td>
<td></td>
<td></td>
<td>AUC 0.75 (AUC 0.74 independent test set)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T1wCE, T2w FLAIR None</td>
<td></td>
<td>Low qualitative appearance interobserver variability</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>direct image input; 3D volume</td>
<td></td>
<td>3-way qualitative appearance scoring inferior to radomics</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Radiomic features significantly different between qualitative appearances</td>
<td></td>
</tr>
<tr>
<td>Jalalifar et al. [158] (Nov. 2022)</td>
<td>SRT (linac)</td>
<td>124 (156)</td>
<td>General</td>
<td>2× 1.5 T</td>
<td>T1wCE, T2w FLAIR None</td>
<td>CNN with transfer learning from brain MR dataset</td>
<td>AUC 0.93 (0.91 independent test set)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(direct image input; 3D volume)</td>
<td></td>
<td>CNN attention methods compared</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CNN’s attention on BM margin</td>
<td></td>
</tr>
<tr>
<td>Jalalifar et al. [157] (Oct. 2022)</td>
<td>SRT (linac)</td>
<td>124 (156)</td>
<td>General</td>
<td>2× 1.5 T</td>
<td>T1wCE, T2w FLAIR None</td>
<td>SVM</td>
<td>AUC 0.81</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(BM ROI)</td>
<td></td>
<td>Clinician ROI vs. automated ROI had limited effect on accuracy</td>
<td></td>
</tr>
<tr>
<td>Author et al.</td>
<td>Treatment (Year)</td>
<td># Patients (BM)</td>
<td>Treatment Modality</td>
<td>Patient Population</td>
<td>Scanners</td>
<td>Radiomic Features (ROIs Used)</td>
<td>Clinical Features</td>
<td>Analysis Technique</td>
</tr>
<tr>
<td>---------------</td>
<td>------------------</td>
<td>----------------</td>
<td>-------------------</td>
<td>--------------------</td>
<td>----------</td>
<td>-------------------------------</td>
<td>------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Jalaliifar et al. [156] (Jul. 2022)</td>
<td>SRT (linac)</td>
<td>124 (156)</td>
<td>General</td>
<td>2x 1.5 T</td>
<td>T1wCE, T2w FLAIR 10× (direct image input; 3D volume)</td>
<td>BMs</td>
<td>CNN with transfer learning from brain MR dataset</td>
<td>• AUC 0.88 (0.86 independent test set) • Clinical features increase accuracy • CNN’s attention on BM margin</td>
</tr>
<tr>
<td>Jiang et al. [155] (Jan. 2022)</td>
<td>SRS (Gamma Knife)</td>
<td>137 (213)</td>
<td>NSCLC</td>
<td>1x 3 T</td>
<td>T1wCE, T1w, T2w, MPRAGE, T2w, T2w FLAIR, ADC, CBV (BM and edema ROIs)</td>
<td>BMs</td>
<td>RDF</td>
<td>• AUC 0.93 (0.85 independent test set) • BM and edema ROIs both importance • T2w, T2w FLAIR, and CBV features important</td>
</tr>
<tr>
<td>Jaberipour et al. [154] (Nov. 2021)</td>
<td>SRT (linac)</td>
<td>120 (171)</td>
<td>General</td>
<td>1x 1.5 T</td>
<td>T1wCE, T2w FLAIR 8× (BM, edema, peri-BM ROIs)</td>
<td>BMs</td>
<td>KNN model</td>
<td>• AUC 0.87 (independent test set) • Adding clinical to radiomic features has no accuracy increase • BM and edema ROIs both importance</td>
</tr>
<tr>
<td>Zheng et al. [153] (Sep. 2021)</td>
<td>SRS (Gamma Knife)</td>
<td>44 (81)</td>
<td>Breast</td>
<td>1x 1.5 T</td>
<td>T1wCE, T2w, ADC (BM ROI)</td>
<td>BMs</td>
<td>Non-ML: univariate/ Cox model, RPA, and nomogram</td>
<td>• One radiomic feature (T1wCE kurtosis) significant • Age and T1wCE kurtosis selected for nomogram</td>
</tr>
<tr>
<td>Mulford et al. [152] (May 2021)</td>
<td>SRS (Gamma Knife)</td>
<td>67 (71)</td>
<td>General; post-resection only</td>
<td>1x 1.5 T</td>
<td>T1wCE (BM ROI)</td>
<td>BMs</td>
<td>Gradient boosted trees</td>
<td>• AUC 0.82 (0.78 independent test set) • Combining T1wCE and dose map features optimal</td>
</tr>
<tr>
<td>Wang et al. [151] (Jul. 2021)</td>
<td>SRS (Gamma Knife)</td>
<td>28 (179)</td>
<td>General</td>
<td>1x 1.5 T</td>
<td>T1wCE, SRS planned dose map (BM ROI)</td>
<td>BMs</td>
<td>Logistic regression</td>
<td>• AUC 0.95 (independent test set) • Combining clinical and radiomic features optimal</td>
</tr>
<tr>
<td>Liao et al. [150] (Aug. 2021)</td>
<td>SRS (Gamma Knife)</td>
<td>256 (976)</td>
<td>NSCLC</td>
<td>Unknown</td>
<td>T1wCE, T1w, T2w (BM ROI)</td>
<td>BMs</td>
<td>SVM</td>
<td>• AUC 0.82 (0.78 independent test set) • Combining clinical and radiomic features optimal</td>
</tr>
<tr>
<td>Kawahara et al. [149] (Jan. 2021)</td>
<td>SRS (Gamma Knife)</td>
<td>52 (157)</td>
<td>Melanoma</td>
<td>1x 1.5 T</td>
<td>T1wCE (BM ROI)</td>
<td>BMs</td>
<td>NN</td>
<td>• AUC 0.79 • Adding radiomic to clinical features increases accuracy • BM and peri-BM ROIs both important</td>
</tr>
<tr>
<td>Huang et al. [148] (Feb. 2020)</td>
<td>SRS (Gamma Knife)</td>
<td>161 (576)</td>
<td>NSCLC</td>
<td>1x 1.5 T</td>
<td>T1wCE (BM ROI)</td>
<td>BMs</td>
<td>Non-ML: univariate/ Cox model analysis</td>
<td>• One radiomic feature (GLSZM zone percentage) found to be significant</td>
</tr>
<tr>
<td>Mouraviev et al. [147] (Jan. 2020)</td>
<td>SRS (Gamma Knife)</td>
<td>87 (408)</td>
<td>General</td>
<td>1x 1.5 T</td>
<td>T1wCE, T2w FLAIR 6× (BM and peri-BM ROIs)</td>
<td>BMs</td>
<td>RDF</td>
<td>• AUC 0.82 (0.87 independent test set) • 3-way qualitative appearance scoring inferior to radiomics</td>
</tr>
<tr>
<td>Karami et al. [146] (Dec. 2019)</td>
<td>SRT (linac)</td>
<td>100 (133)</td>
<td>General</td>
<td>1x 1.5 T</td>
<td>T1wCE, T2w FLAIR None (BM, edema, peri-BM ROIs)</td>
<td>BMs</td>
<td>SVM</td>
<td>• AUC 0.82 • Edema and peri-BM ROIs most important</td>
</tr>
<tr>
<td>Cha et al. [145] (Sep. 2018)</td>
<td>SRS &amp; SRT (CyberKnife)</td>
<td>89 (110)</td>
<td>General</td>
<td>1x CT</td>
<td>CT (direct image input; central slice)</td>
<td>BMs</td>
<td>CNN</td>
<td>• AUC 0.86 • Ensemble of CNNs optimal</td>
</tr>
</tbody>
</table>

42
Table 1.6: Summary and comparison of the work presented in this thesis to previous studies investigating the prediction of BM response to SRS/SRT using radiomics. Articles are presented in chronological order by date of publication. For article results, AUC values are given by default for experiments using a single dataset via cross-validation or bootstrapped resampling. If independent internal or external testing datasets were used, separate AUC values will be indicated. The three research projects contained in this thesis as chapters 2–4 are indicated in bolded font. Abbreviations: k nearest neighbours (KNN), support vector machine (SVM), multi-layer perceptron (MLP), magnetization-prepared rapid gradient-echo (MPRAGE), cerebral blood volume (CBV)
1.4.6 Knowledge gaps and unmet needs in predicting brain metastasis response to stereotactic radiosurgery using machine learning

From this recent and growing body of work, it is clear that BM response to SRS/SRT can be predicted per BM using radiomics-based ML. These results indicate that steps towards clinical translation should be intentionally pursued, but multiple knowledge gaps and unmet needs currently present barriers to translation.

First, previous studies leave ML model sensitivity to critical clinical factors unexplored. Specifically model sensitivity to primary cancer site and BM volume are important to consider given the results of previous non-ML studies that show the importance of primary cancer specific prognostics (e.g. the DS-GPA model) [20, 114] and the impact of BM volume in SRS planning [89, 97]. Furthermore, while some studies have had two MR scanners represented in their dataset, none have specifically examined the impact of MR scanner variability on model performance. Studies in other fields have found MR scanner variability critical to consider [165, 166], and so it would be beneficial to examine this effect in BM SRS outcome prediction as well. This is especially important considering that centres treating BM patients may have a mixture of MR scanner models, may upgrade MR scanners over time, and may occasionally use MR images acquired externally. It has also not been explored if the observed results hold for linac-based SRS. While Gamma Knife and CyberKnife-based SRS has historically been the default treatment modality for large centres with high BM caseloads, linac-based SRS is becoming more common as smaller centres begin to deliver SRS and as all centres consider using linacs to treat high numbers of BMs concurrently per patient.

Second, previous studies have not compared their results to existing post-SRS progression predictive models. The studies by Kawahara et al. [150] and Gutsche et al. [160] both compared their ML models to using only the “homogeneous”, “heterogeneous”, or “ring-enhancing” qualitative appearance of BMs to predict SRS outcomes, showing the ML
models to be far superior. This result is not surprising, as the three qualitative appearances studied were not developed into a robust predictive model, but rather were reported as a single predictive variable [118, 119]. Furthermore, the three qualitative appearances examined are not as biologically descriptive nor as precise as the five-way appearance labelling performed by Rodrigues et al., which were also used to develop a predictive model [123].

Another critical aspect to consider when comparing qualitative and quantitative analysis is the interobserver variability of the qualitative analysis technique. The interobserver variability of five-way appearance labelling performed by Rodrigues et al. [123] has not yet been characterized.

Third, ML models are notorious for their operation being opaque and difficult to interpret. Current studies have offered extremely limited interpretation of the ML models produced. Studies have shown which radiomic features, MRI sequences, or ROIs are the most important to the ML models on an individual level, but these interpretations are often conflicting, lack biological relevance, and do not encompass the operation of the entire, highly complex ML model.

Fourth, external validation of current ML models and techniques across multiple centres is required. No external validation of a radiomics-based ML model predicting BM SRS response had been reported until Du et al.’s recent study in March 2023 [163]. While encouraging results were reported, Du et al.’s external validation dataset had a limited number of patients (n = 30) and both centres used the same, single MR scanner model. The validation of a locked method to train ML models per centre or the pooling of data across centres was also unexplored.

1.5 Thesis outline and research objectives

This research thesis aims to address the identified knowledge gaps and unmet needs in the field of predicting BM response to SRS using radiomics-based ML models, as demon-
strated in table 1.6. To do so, the following research objectives were accomplished in the indicated chapters.

1. To validate that linac-based SRS outcomes can be predicted using radiomics-based ML models and to determine which combination of clinical and/or radiomic features is optimal for this application (chapter 2).

2. To examine the sensitivity of radiomics-based ML models to primary cancer site, BM volume, and MR scanner model (chapter 2).

3. To characterize the interobserver variability of BM qualitative appearance labelling and its impact on risk stratification for post-SRS progression (chapter 3).

4. To compare using qualitative appearance-based analysis versus quantitative radiomics-based analysis of BMs in T1w-CE MRI to predict post-SRS progression (chapter 3).

5. To interpret radiomics-based ML models by relating their operation to biologically-relevant BM qualitative appearances (chapter 3).

6. To externally validate radiomics-based ML models and the methodology used to train them with datasets from two independent centres (chapter 4).

1.6 References


[40] International Headache Society. Headache Classification Committee of the International Headache Society (IHS) The International Classification of Headache Disor-


Chapter 2

Performance sensitivity analysis of brain metastasis stereotactic radiosurgery outcome prediction using magnetic resonance imaging radiomics

The following research chapter was completed to meet thesis research objectives 1 and 2 outlined in section 1.5:

1. To validate that linac-based SRS outcomes can be predicted using radiomics-based ML models and to determine which combination of clinical and/or radiomic features is optimal for this application.

2. To examine the sensitivity of radiomics-based ML models to primary cancer site, BM volume, and MR scanner model.

This chapter provides the foundation of the further research conducted in chapters 3 and 4 by validating that radiomics-based ML models can be applied to linac-based SRS and then characterizing model sensitivity to relevant clinical factors.
2.1 Introduction

Primary cancers can spread to form BMs. Improved treatments have lengthened patient survival, leading to BM incidence in 20–40% of patients [1]. Since patients typically have advanced cancer extracranially when BMs develop, BMs are less likely to cause death, but do cause serious symptoms affecting QOL [2]. Prognosis is poor due to extracranial factors, with a median survival of seven months [3]. It is therefore critical to choose the optimal treatment for BMs at the outset, as there is limited time to pivot to an alternative treatment approach.

Common treatments include surgical resection and RT, along with newly developed systemic targeted and immunotherapy agents. Surgery is an effective option, but is contraindicated by inoperable BMs near eloquent locations and patient comorbidities. Radiation-based treatments offer a non-invasive, but effective alternative [4]. WBRT irradiates the entire brain over many fractions with the intent to spare healthy tissue. SRS uses highly conformal radiation doses delivered in 1–3 fractions to spare healthy tissue and cognitive function [4]. SRS may be delivered on a linac, or on more specialized machines, such as Gamma Knife or CyberKnife. SRS can also be hypofractionated to five fractions which will be defined as SRT.

Despite the advantages of SRS, it is not always successful; SRS fails for up to 30% of BMs [5]. In these cases, patients may continue to experience decreased QOL from the BM. SRS also has side-effects [4] including possible radiation necrosis which may increase patient steroid dependency.

It would be advantageous to have a prognostic model predicting whether a BM would progress after SRS. Such a model could aid in justifying SRS dose escalation or deciding between surgery and SRS. While prognostic models for BM patient OS have been developed, these do not inform if individual BMs will respond to SRS [6, 7]. Previously reported prognostic factors for BM progression post-SRS include BM volume, dose and fractiona-
tion, and BM appearance in pre-treatment T1w-CE MRI [2, 7, 8].

Previous studies linking BM MRI appearance to SRS outcomes distilled all the MRI information into a single, qualitative appearance score, such as homogeneous, heterogeneous, or rim-like enhancement. ML models and radiomic analysis of MRI have been incorporated into many health fields to provide prognostics [9–11]. Radiomics provides quantitative image analysis by extracting computational features from voxel data. These features can be combined using ML techniques to predict a desired outcome.

Previous studies have reported on prediction of SRS outcomes for BMs treated with Gamma Knife or CyberKnife. They investigated the use of radiomic features from multiple MRI sequences, the planned SRS dose distribution, and the peri-tumoural region beyond the treated BM [12–16]. There have also been similar studies for BM SRT and post-surgical resection SRS using Gamma Knife [17–19].

While these studies have offered preliminary insights into BM SRS prognostics, there remains an unmet need for a study that provides adequate evidence supporting the pursuit of external, multi-center validation of a model that predicts the outcome of BM SRS. To date, ML studies in this area have not examined the effects of primary cancer type, BM volume, and multiple MR scanner models on predictive accuracy. Investigating the effect of primary cancer site is important to consider as current OS prognostic models have been specifically developed for individual cancer sites [6]. Furthermore, some current ML studies use patient samples with mixed primary cancer sites, while some ML studies use single cancer site samples, with possible differences between these approaches remaining unexplored [14, 15]. Exploring the BM volume and MR scanner effects is also important as both ROI volume and MR scanner model have widely been found to have effects on the performance of radiomics systems [20–23]. There has also not yet been a robust comparison of the accuracy of clinical and radiomic features for this task. Our study addresses this unmet need and provides more evidence for the feasibility of external validation of a radiomics-based prognostic model. Our study addresses these pertinent research questions:
1. What performance increase does adding a set of radiomic features offer to a prognostic model, compared to only using clinical features?

2. Does a prognostic radiomics-based model offer equal predictive accuracy for BMs from different primary cancer types?

3. Do prognostic radiomics-based models predict outcomes of BMs with different volumes with varied accuracy?

4. Does the removal of volume-correlated features produce prognostic models with more balanced accuracy across BMs of different volumes?

5. Which clinical and radiomic features are most important for a prognostic model both before and after the removal of volume-correlated features?

6. What effect does the inclusion of multiple MR scanner models have on the accuracy of a radiomics-based model?

2.2 Methods

2.2.1 Study sample

We analyzed 99 patients from the cohort studied by Rodrigues et al. [7] for whom MRI and BM contouring data were available. These 99 patients were randomly selected from the initial cohort to minimize selection bias. For the initial cohort, there were implicit inclusion and exclusion criteria reflective of clinical practice due to the retrospective design of the study by Rodrigues et al. [7]. Specifically, patients with highly symptomatic BMs or poor prognosis would be excluded from receiving SRS. For the retrospective study itself, the inclusion criteria consisted of patients with up to three newly diagnosed, radiologically confirmed BMs that were treated using SRS between 2003 and 2011. Patients with previous surgical resection, prior RT to the brain, or recurrent BMs were excluded. Due to the data
requirements for MRI analysis and a radiologically confirmed endpoint, patients were also excluded if they did not have a pre-treatment or follow-up MRI.

The SRS treatments were performed using either the Novalis or Novalis TX linac with a Gill-Thomas-Cosman frame or BrainLAB frameless mask system for immobilization (BrainLAB, Feldkirchen, Germany). BMs < 7.5 cc generally received the most aggressive prescription (21 Gy in one fraction), while larger BMs would receive less aggressive prescriptions. This dose prescription guideline was developed by the centre at which the data were collected, as unified prescription protocols were not yet widely available. 127 BMs in total were treated with SRS. Four of these BMs were excluded as they consisted of a single voxel due to partial volume effects during application of the BM contouring data to the MRI voxel grid, and many radiomic features (e.g. all second-order texture features) would not be computable as they require more than one voxel. Outcome prediction and analysis was performed per BM instead of per patient, giving a total of \( n = 123 \) BMs that were individually analyzed. Table 2.1 shows the clinical features of the patients and BMs in this study.

Each BM’s ROI was defined as the GTV manually contoured during SRS treatment planning by an experienced radiation oncologist, which was based on the outer border of the enhancing region in high-resolution T1w-CE MRI. Non-enhancing regions within the border were included in the contour. T1w-CE MRI was acquired with different MR scanner models, voxel sizes, and acquisition orientations. Five scanner models from Siemens and General Electric (Erlangen, Germany; Chicago, USA) and eight scan configurations were used (table 2.2).

2.2.2 Clinical features and study endpoint

For each patient, a set of 12 clinical features that were readily available for collection were used (table 2.1), representing data that would be typically available in clinical practice before SRS.
<table>
<thead>
<tr>
<th>Clinical Features</th>
<th># Patients</th>
<th># BMs (% Progression)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex (p = 0.831)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>55</td>
<td>66 (21.2%)</td>
</tr>
<tr>
<td>Male</td>
<td>44</td>
<td>57 (22.8%)</td>
</tr>
<tr>
<td>Age (p = 0.925)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (Range)</td>
<td>58.0 (38.4-86.0) years</td>
<td></td>
</tr>
<tr>
<td>Primary Cancer Active (p = 0.002)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>44</td>
<td>55 (9.1%)</td>
</tr>
<tr>
<td>No</td>
<td>55</td>
<td>68 (32.4%)</td>
</tr>
<tr>
<td>Primary Cancer Site (p = 0.003)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung</td>
<td>59</td>
<td>70 (12.9%)</td>
</tr>
<tr>
<td>Breast</td>
<td>10</td>
<td>14 (35.7%)</td>
</tr>
<tr>
<td>Renal</td>
<td>10</td>
<td>15 (13.3%)</td>
</tr>
<tr>
<td>Colorectal</td>
<td>8</td>
<td>10 (40.0%)</td>
</tr>
<tr>
<td>Skin</td>
<td>8</td>
<td>9 (66.7%)</td>
</tr>
<tr>
<td>Other</td>
<td>4</td>
<td>5 (20.0%)</td>
</tr>
<tr>
<td>Primary Cancer Histology (p = 0.001)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>49</td>
<td>65 (20.0%)</td>
</tr>
<tr>
<td>NSCLC</td>
<td>31</td>
<td>36 (11.1%)</td>
</tr>
<tr>
<td>Melanoma</td>
<td>8</td>
<td>9 (66.7%)</td>
</tr>
<tr>
<td>Squamous Carcinoma</td>
<td>7</td>
<td>8 (50.0%)</td>
</tr>
<tr>
<td>Other</td>
<td>4</td>
<td>5 (0.0%)</td>
</tr>
<tr>
<td>Extracranial Systemic Metastases (p = 0.665)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>39</td>
<td>50 (20.0%)</td>
</tr>
<tr>
<td>No</td>
<td>60</td>
<td>73 (23.3%)</td>
</tr>
<tr>
<td>Systemic Therapy Status (p = 0.033)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Radical</td>
<td>51</td>
<td>10 (20.0%)</td>
</tr>
<tr>
<td>Palliative</td>
<td>41</td>
<td>60 (31.7%)</td>
</tr>
<tr>
<td>None</td>
<td>7</td>
<td>53 (11.3%)</td>
</tr>
<tr>
<td>Neurological Symptoms Steroid Response (p = 0.426)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fully resolved</td>
<td>48</td>
<td>31 (16.1%)</td>
</tr>
<tr>
<td>Improvement</td>
<td>7</td>
<td>4 (50.0%)</td>
</tr>
<tr>
<td>Limited improvement</td>
<td>4</td>
<td>56 (25.0%)</td>
</tr>
<tr>
<td>No improvement</td>
<td>26</td>
<td>11 (9.1%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>14</td>
<td>21 (23.8%)</td>
</tr>
<tr>
<td>ECOG/WHO Performance Score (p = 0.580)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>31</td>
<td>39 (17.9%)</td>
</tr>
<tr>
<td>1</td>
<td>60</td>
<td>73 (21.9%)</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>39 (33.3%)</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2 (50.0%)</td>
</tr>
<tr>
<td>GTV Volume (p &lt; 0.001)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (Range)</td>
<td>3.07 (0.02-30.23) cc</td>
<td></td>
</tr>
<tr>
<td>&lt; 7.5 cc</td>
<td>-</td>
<td>94 (17.0%)</td>
</tr>
<tr>
<td>&gt; 7.5 cc</td>
<td>-</td>
<td>29 (37.9%)</td>
</tr>
<tr>
<td>BM Location (p = 0.626)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supratentorial</td>
<td>-</td>
<td>96 (22.9%)</td>
</tr>
<tr>
<td>Infratentorial</td>
<td>-</td>
<td>27 (18.5%)</td>
</tr>
<tr>
<td>SRS Prescription (p = 0.003)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 Gy in 1 fraction</td>
<td>-</td>
<td>5 (0.0%)</td>
</tr>
<tr>
<td>18 Gy in 1 fraction</td>
<td>-</td>
<td>36 (30.6%)</td>
</tr>
<tr>
<td>21 Gy in 1 fraction</td>
<td>-</td>
<td>72 (13.9%)</td>
</tr>
<tr>
<td>24 Gy in 3 fractions</td>
<td>-</td>
<td>10 (60.0%)</td>
</tr>
</tbody>
</table>

Table 2.1: Clinical feature distributions for number of BMs, BMs progressing post-SRS, and patients (where applicable) for the study sample. The “Neurological Symptoms Steroid Response” feature qualitatively scores the improvement of neurological symptoms after the administration of steroids, based on the previously reported methodology of Lagerwaard et al. [24]. p-values provided for statistical comparisons between BMs that progressed and did not progress post-SRS. The Wilcoxon rank sum test was used for continuous features (age and GTV volume). The Chi-squared test was used for the remaining categorical features. A histology scoring of “NSCLC” represents NSCLC BMs which were not scored as adenocarcinoma. Abbreviations: Eastern Cooperative Oncology Group (ECOG), World Health Organization (WHO)
<table>
<thead>
<tr>
<th>Scanner Model and Field Strength</th>
<th>Acquisition Orientation</th>
<th>Voxel Size (mm$^3$)</th>
<th># Patients</th>
<th># BMs (% Progression)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Siemens Magnetom Vision (1.5 T)</td>
<td>Sagittal</td>
<td>1x1x1.5</td>
<td>35</td>
<td>39 (28.2%)</td>
</tr>
<tr>
<td>Siemens Avanto (1.5 T)</td>
<td>Sagittal</td>
<td>0.5x0.5x1</td>
<td>30</td>
<td>37 (13.5%)</td>
</tr>
<tr>
<td></td>
<td>Axial</td>
<td>0.5x0.5x2</td>
<td>5</td>
<td>8 (0.0%)</td>
</tr>
<tr>
<td>Siemens Magnetom Expert (1.0 T)</td>
<td>Sagittal</td>
<td>1x1x1.5</td>
<td>21</td>
<td>29 (31.0%)</td>
</tr>
<tr>
<td>Siemens Sonata (1.5 T)</td>
<td>Axial</td>
<td>1x1x1.5</td>
<td>1</td>
<td>1 (0.0%)</td>
</tr>
<tr>
<td></td>
<td>Axial</td>
<td>1x1x2</td>
<td>1</td>
<td>3 (0.0%)</td>
</tr>
<tr>
<td>General Electric Signa HDxt (1.5 T)</td>
<td>Sagittal</td>
<td>1x1x1.5</td>
<td>1</td>
<td>1 (0.0%)</td>
</tr>
</tbody>
</table>

Table 2.2: Number of patients and BMs scanned by each of the MR scanner model and acquisition parameters configurations. Chi-squared test for progression yielded $p = 0.226$ across all scanner models and $p = 0.069$ across further investigated Vision, Avanto, and Expert scanners.

The study endpoint was progression of each BM post-SRS, which was assessed on follow-up T1w-CE MRI acquired approximately every three months. Across the study sample, there was a median follow-up time of nine months, with an interquartile range of six months, minimum of four months, and the two longest followed patients were last imaged at 16 and 24 months. BMs that had not progressed when lost to follow-up were scored as non-progression. For each BM, the maximum diameter was measured in three perpendicular directions (superior-inferior, mediolateral, posterior-anterior) by a single radiation oncologist. The product of the three maximum diameters was taken to provide a correlate to BM volume given the approximate spherical appearance of most BMs. If this quantity increased by $\geq 25\%$ post-treatment, then the BM was recorded as having progressed. This endpoint was used to define a binary classification label for each BM, with BM progression defined as “positive”, and non-progression defined as “negative”. Our study contained 22.0% BMs that progressed, versus 78.0% that did not.

BM can appear to progress in T1w-CE post-SRS but not be associated with cancerous progression. This is known as pseudo-progression [25]. We differentiated between true and pseudo-progression based on expert clinician judgement using post-SRS serial MRI and medical records. Pseudo-progression was scored as non-progression.
2.2.3 Radiomic features

The T1w-CE MRI scans were pre-processed to account for voxel size and intensity scaling differences between MR scanner models. The mean and standard deviation of all voxels within the brain were computed to apply a Z-score normalization to the entire scan at three standard deviations [22, 26]. The scan was then linearly interpolated to a voxel size of $0.5 \times 0.5 \times 0.5$ mm$^3$.

Within each BM, radiomic features were extracted from the pre-treatment T1w-CE MRI. 107 unique radiomic features were extracted using the open-source PyRadiomics library v3.0.1 in Python v3.6.13 [27] (list provided in supplementary table A.1). Where required for feature extraction, 64 intensity value bins were used.

2.2.4 Machine learning experiment design

Our ML experiments were performed using a study design template built in Matlab 2019b v9.7.0.1190202 (The Mathworks Inc., Natick, USA), consisting of a bootstrapped resampling design in which the dataset was partitioned into a training and testing dataset through random sampling with replacement. This was done at a per-patient level to avoid data leakage, by disallowing a patient’s BMs from being in the training and testing datasets.

The training dataset was used for inter-feature correlation filtering, hyper-parameter optimization, and training of a RDF model. This trained model was then evaluated on its predicted probabilities of BM progression in the testing dataset. After 250 iterations of bootstrapped resampling, we calculated the mean model predictive accuracy and associated confidence intervals (CIs). For reproducibility, a detailed explanation of the process is provided in supplementary figure A.1 and table A.2.
2.2.5 Machine learning experiment error metric calculation

We computed a set of error metrics to describe the predictive accuracy of each technique. We used the predicted probabilities of progression and ground truth SRS outcomes, aggregated across all the bootstrapped repetitions’ testing datasets, to compute an average ROC curve, AUC, and associated 95% CIs. It is well known that the average AUC from a bootstrapped resampling experiment is an underestimate of performance on unseen data [28]. To have AUC values that were more readily comparable to other studies, we also computed the commonly employed AUC$_{0.632+}$ that was developed to correct for this underestimation [28].

To calculate the average misclassification rate (MCR), false negative rate (FNR), and false positive rate (FPR), an operating point on the average ROC curve had to be chosen. The predicted probabilities from the out-of-bag datasets for each tree within each RDF were aggregated across all bootstrapped repetitions. From these aggregated probabilities, another average ROC curve was constructed and the optimal upper-left operating point found. The out-of-bag dataset probabilities were strictly informed by each repetition’s training dataset, and so provided an optimal operating point while avoiding overfitting to the testing dataset ROC curve. The optimal operating point was then transferred to the testing dataset ROC curve to compute the average MCR, FNR, and FPR. Supplementary figure A.1 provides an illustration of this approach.

2.2.6 Machine learning experiments addressing the research questions

To answer the first research question, on the relative predictive accuracy of clinical and radiomic features, three experiments were performed, each identical except for the features available to the model. The first experiment used only clinical features, the second used only radiomic features, and the third used all features.

The second research question, surrounding the effect of primary cancer type on predic-
tive accuracy, was addressed through stratified calculation of error metrics by primary cancer type. The testing dataset and out-of-bag predicted progression probabilities from the previous experiment using both clinical and radiomic features were re-aggregated across bootstrap repetitions, with probabilities grouped according to the primary cancer site for each BM. These grouped probabilities were used to calculate the error metrics for each primary cancer site. This re-aggregation and grouping were required as some primary cancer types were associated with very few BMs, and so training models specifically for each primary cancer type was not feasible.

To investigate the effects of BM volume described in the third research question, similar techniques to that described above for stratified primary cancer type analysis were used, except based on BM volume. The BMs were stratified into one group of small BMs (<7.5 cc, approximately 2.4 cm in diameter if BM is assumed to be spherical), and a second group of large BMs (>7.5 cc), to reflect the clinical use of the 7.5 cc volume threshold for dose prescription. Error metrics were then individually calculated for each BM volume group.

We answered the fourth research question on the effect of removing volume-correlated features by identifying features for removal using statistical tests for dependence on and correlation to BM volume. For categorical clinical features, the dependence of a feature on BM volume was determined using the Wilcoxon rank sum test or Kruskal-Wallis test, depending on whether the feature was binary or not. If a feature was found to have a p-value less than a Bonferroni-corrected alpha value of $\alpha = 0.05/118$, it was removed. Pearson correlation coefficient values were computed between continuous features and both BM volume and its cubic root (analogous to BM diameter). If the p-values associated with the correlation coefficients of a feature was significant at $\alpha = 0.05/118$, then the feature was available for removal if either correlation coefficient was greater than a given correlation coefficient threshold. Correlation coefficient thresholds between 0.1 and 0.85 at a spacing of 0.15 were tested, along with a threshold of 0. This methodology was designed to explore
gradually blinding the model to BM volume to result in more balanced predictive accuracy between the two volume groups.

The fifth research question addressing the relative importance of clinical and radiomic features in the prognostic models was answered using feature importance rankings inherent to RDF models. To aggregate the feature importance scores across bootstrapped repetitions, the raw feature importance scores for each repetition were normalized to be between 0 and 1, and any features removed by the inter-feature correlation filter (as opposed to the volume correlation filter) were given a feature importance score of 0. The scores were then averaged across all bootstrapped repetitions and then renormalized to be between 0 and 1. A feature was deemed “highly important” if it had a final feature importance score above 0.75.

Lastly, to address the final research question around the effects of MR scanner variability, repeat experiments were performed with datasets stratified by MR scanner model. For the study sample, most patients were imaged using either the Magnetom Vision, Magnetom Expert, or Avanto scanner models (see table 2.2). Three repeat experiments using all radiomic and clinical features were then performed with a subset of the study sample that was imaged only on two of these three scanner models. A relatively high predictive accuracy from one of these experiments would suggest the two scanners being in a similar data domain, while a low predictive accuracy would suggest the opposite.

2.2.7 Ethics declaration

The collection and analysis of the retrospective patient data used in this study was approved by the Amsterdam University Medical Centre (AUMC) Medical Ethical Review Committee and was conducted within the approved guidelines. As the study was retrospective on a cohort of deceased patients, written consent from study participants was waived by the AUMC Medical Ethical Review Committee.
Figure 2.1: Comparison of average ROC curves across bootstrapped repetitions when clinical, radiomic, or clinical and radiomic features were used. The AUC for each curve is provided, along with the corrected AUC\textsuperscript{0.632+} value. The optimal upper-left operating point determined from the out-of-bag dataset is plotted as “+”, for which the MCR, FNR, and FPR were found. A false negative represents a BM that was predicted to respond to SRS, but instead progressed after SRS. 95% CIs are provided for the curves (shaded bands) and error metrics (in parentheses).

### 2.3 Results

#### 2.3.1 Clinical and radiomic features predictive accuracy

Figure 2.1 shows the average ROC curves comparing the performance of using clinical features alone, radiomic features alone, and their combination. Comparison of the AUCs reveals that clinical and radiomic features offer the superior prognostic model. While the AUC differences between the scenarios are statistically significant, they are small (0.01 to 0.03). Given these results, we used clinical and radiomic features for the remainder of the
<table>
<thead>
<tr>
<th>Primary Site</th>
<th>AUC</th>
<th>MCR %</th>
<th>FNR %</th>
<th>FPR %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colorectal</td>
<td>0.90 (0.04)</td>
<td>19.8 (3.6)</td>
<td>15.8 (3.3)</td>
<td>20.9 (3.7)</td>
</tr>
<tr>
<td>Skin</td>
<td>0.84 (0.05)</td>
<td>27.4 (4.5)</td>
<td>17.8 (3.2)</td>
<td>30.1 (4.9)</td>
</tr>
<tr>
<td>Renal</td>
<td>0.72 (0.03)</td>
<td>39.1 (3.2)</td>
<td>29.1 (4.9)</td>
<td>41.9 (2.8)</td>
</tr>
<tr>
<td>Lung</td>
<td>0.64 (0.02)</td>
<td>35.5 (2.0)</td>
<td>45.2 (3.3)</td>
<td>32.7 (1.7)</td>
</tr>
<tr>
<td>Breast</td>
<td>0.49 (0.04)</td>
<td>41.5 (3.1)</td>
<td>48.9 (4.0)</td>
<td>39.4 (2.9)</td>
</tr>
</tbody>
</table>

Table 2.3: Error metrics derived by analysis per BM primary cancer site. 95% CIs provided in parentheses.

The reported AUC values are likely pessimistically biased values due to the nature of the bootstrapped resampling technique. In figure 2.1, the $\text{AUC}_{0.632+}$ consistently shows a lift in AUC, up to $\text{AUC}_{0.632+} = 0.77$ when clinical and radiomic features are used. The small 95% CIs demonstrate the consistent performance of our technique across repetitions.

Figure 2.1 also contains the MCR, FNR, and FPR values for each ROC curve. They show that FPR remains stable between feature types, with higher AUC values correlating with lower FNRs.

### 2.3.2 Primary cancer site dependency of predictive accuracy

The predictive accuracy of the model varied widely between primary sites when using clinical and radiomic features (table 2.3). Progression is predicted most accurately for colorectal and skin, while renal and lung were predicted with AUCs closer to the overall accuracy across all sites. For breast cancer primaries, predictive accuracy was at chance.

The 95% CIs of the error metrics were found to vary across primary sites. Lung cancer was the most common primary site, yielding more data points to determine the average ROC curve, reducing the associated CIs.

### 2.3.3 Volume dependency of predictive accuracy

Our technique had varying predictive accuracy depending on BM volume when using clinical and radiomic features. For small BMs ($< 7.5 \text{ cc}$) we found that response to SRS was
Figure 2.2: Volume dependency of error metrics. (a) Shows the comparison of error metrics between all BMs, BMs with volume < 7.5 cc, and BMs with volume > 7.5 cc before the removal of any volume-correlated features. (b–d) show the AUC, FNR, and FPR for each BM volume group as a function of the feature volume-correlation coefficient threshold value used to remove continuous features, with correlation threshold = 1 corresponding to the removal of no features (baseline). At all correlation thresholds < 1, categorical clinical features dependent on BM volume are also removed. For example, values at correlation threshold = 0.7 represent the error metric values when the experiment was repeated with all features with volume-correlation coefficient values greater than 0.7 removed. For all values, a 95% CI is provided by error bars or shaded bands.

predicted with accuracy comparable to using the overall dataset (figure 2.2a). Performance for large BMs (> 7.5 cc) fell considerably with statistical significance. The lower AUC value for large BMs mostly translated into a higher FPR, nearing 50%. Large BMs had the widest 95% CIs, likely due to their relatively small number.

The only clinical feature found to be dependent on BM volume, and therefore removed, was the received dose and fractionation. In total, 72 features were removed at the corre-
lation threshold of 0.25, leaving 47 features for use by the model. As the threshold was reduced past 0.25, no further features were candidates for removal due to their level of statistical significance. As volume-correlated features were removed down to a correlation threshold of 0.25, we found the difference in AUC, FNR and FPR for small and large BMs was reduced to within the associated 95% CIs. Figure 2.2b–d demonstrates that while FNR remains well-matched across volume groups, FPR fell and AUC rose significantly for large BMs, while the metrics for small BMs and the overall dataset remained stable. As no further features were removed past a threshold of 0.25, model performance remained constant past the 0.25 threshold as well.

2.3.4 Feature importance analysis

The relative importance of all the clinical and radiomic features at each correlation coefficient threshold are provided within supplementary figure A.2. Supplementary table A.3 provides a detailed comparison between the highly important features at the correlation threshold of 1 and 0.25, but the key results are provided here.

Primary cancer site and histology were highly important before and after volume-correlated features were removed. Before volume-correlated features were removed, primary cancer site and histology were the second and fourth most important features, respectively, and were the third and fourth most important features when the correlation threshold was 0.25. The only other highly important clinical feature was GTV volume, but only before it was removed. Univariate analysis showed that the “primary cancer active” and “systemic therapy status” features predicting post-SRS progression with $p < 0.05$ (table 2.1). These two features were not found to be important by the model however, suggesting that alternative features provided superior predictive value.

When the correlation threshold was set to 0.25, eight radiomic features were highly important, and they gained importance as more volume-correlated features were removed. Four of the eight radiomic features, GLCM cluster shade, GLCM cluster prominence,
GLCM contrast, and interquartile range, were found to transition to become highly important features as the correlation threshold was reduced to 0.25. The remaining four features (kurtosis, GLCM inverse variance, neighbouring gray tone difference matrix contrast, and 10th percentile value) consistently remained highly important as features were removed. As volume-correlated features were removed to reach the 0.25 threshold, 11 radiomic features that were highly important before feature removal were found to be correlated to BM volume or diameter.

2.3.5 Effects of multiple magnetic resonance scanner models

The pairing of the Magnetom Vision and Expert scanners yielded increased predictive accuracy; AUC increased (from a baseline of 0.69), translating mostly into a decrease of FNR (figure 2.3). When the Avanto scanner was paired with either the Magnetom Vision or Expert scanner, the performance of the model was found to decline when compared to the baseline.

2.4 Discussion

Our study showed the synergistic relationship of clinical and radiomic features. Previous studies have reported on the relative accuracy of clinical and radiomic features in BM SRS; however, the disagreement of their results may be due to varying experimental designs. Mouraviev et al. observed an 18% AUC increase between using clinical features alone to using radiomic and clinical features, though using radiomic features alone was not investigated [12]. Jiang et al. saw no increase in AUC when using radiomic features alone compared to using radiomic and clinical features, though no baseline of using clinical features alone was established [15]. Studies on post-operative BM SRS or SRT found radiomic features outperformed clinical features and that clinical features decreased accuracy when combined with radiomic features [17, 18].
Table 2.3: Error metrics across MR scanner models. The average ROC curves across bootstrapped repetitions are compared when using clinical and radiomic features only from pairings of the three primary MR scanner models used (Siemens Magnetom Vision, Magnetom Expert, and Avanto). As for figure 2.1, for each curve in this figure, associated error metrics, optimal operating point (+), and 95% CIs (shaded band and values in parentheses) are provided.

<table>
<thead>
<tr>
<th>ROC Models</th>
<th>AUC</th>
<th>AUC_{0.632+}</th>
<th>MCR %</th>
<th>FNR %</th>
<th>FPR %</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Vision &amp; Expert</td>
<td>0.77 (0.01)</td>
<td>0.84</td>
<td>30.4 (1.8)</td>
<td>29.6 (2.1)</td>
<td>30.6 (1.8)</td>
</tr>
<tr>
<td>B Vision &amp; Avanto</td>
<td>0.65 (0.02)</td>
<td>0.73</td>
<td>34.1 (1.8)</td>
<td>43.4 (2.4)</td>
<td>31.5 (1.6)</td>
</tr>
<tr>
<td>C Expert &amp; Avanto</td>
<td>0.58 (0.02)</td>
<td>0.66</td>
<td>35.7 (1.9)</td>
<td>52.6 (2.6)</td>
<td>31.0 (1.7)</td>
</tr>
</tbody>
</table>

2.4.1 Clinical and radiomic features predictive accuracy

Our study presents a more robust comparison of clinical and radiomic features than previously reported. All previous studies used fewer clinical features than investigated here. Some studies selected radiomic features using the entire dataset, leading to possible overfitting [12, 19]. All studies selected radiomic features in the absence of the clinical features, leading to possibly rejecting radiomic features that are only beneficial in conjugation with clinical features. By including more clinical features and performing repeated experiments including clinical and/or radiomic features, our study provides a firmer result that the use
of clinical and radiomic features together does indeed offer increased accuracy, but not as large of a benefit as previously reported.

2.4.2 Primary cancer site dependency of predictive accuracy

Different primary cancer sites have been associated with varying BM SRS success rates [2, 3, 29]. Our study offers a subtly alternative and unexplored insight that the outcomes of BMs from some primary sites are more difficult to predict than others with a model trained with data from multiple primary sites. Due to small sample sizes of some primary sites and training the model on all primary sites, it is difficult to make strong conclusions as to why this occurs. Radiomic features of BMs have been shown to be moderately predictive of primary cancer site, with AUCs up to 0.87 [30]. There could therefore be radiomic features unique to certain primary cancer types that are useful for outcome prediction. Our small primary cancer subsamples and the fact that no other studies have provided similar analysis as presented here make probing this hypothesis difficult. Future study of this question with sufficient sample size is warranted. We also explored controlling for primary cancer site in a similar manner as was used for controlling for the BM volume effect by removing any features found to be dependent on primary cancer site as determined through Kruskal-Wallis and Chi-Squared tests and a Bonferroni-corrected $\alpha = 0.05/118$. The primary cancer active, primary cancer histology, and systemic metastases clinical features were all dependent on primary cancer site and were removed along with primary cancer site. The same experiments and analysis were then conducted and a negligible impact on the error metrics per primary cancer site was found, leaving the disparity between primary cancer sites intact.

Our results indicate that externally validating a model on a sample with differing primary cancer ratios than the model’s training sample could lead to difficulty interpreting the results. For example, our technique would perform quite poorly if validated with a sample containing a high incidence of breast cancer. Furthermore, such imbalanced performance could lead models to provide little benefit to patients with certain primary cancers, despite
promising overall performance. Performance evaluation per primary cancer site should therefore become standard practice.

This performance imbalance could be addressed by constructing distinct models for each primary site. Radiomic models have been developed for SRS treating BMs from lung [15] and skin [14] primary cancers, and they yield higher accuracy metrics than every radiomic study on mixed primary cancers. Alternatively, our techniques appear to offer a very useful model for colorectal or skin cancer patients as is, and so training a model using data with mixed primary cancers may still offer high value to some patient sub-populations. Performing a similar evaluation based on the primary cancer histology and developing histology specific models could also be a critical avenue for future exploration. As primary cancer histology is typically known for a BM prior to SRS treatment, histology specific models would likely be as clinically translatable as models that are primary site specific.

### 2.4.3 Volume dependency of predictive accuracy

SRS is well known to be less likely to fail for smaller BMs [2]. This is reflected in our study sample with GTV volume being the only clinical variable providing a univariate prediction of SRS outcome with \( p < 0.001 \). Therefore, a model that predicts the response of small BMs well, but not large BMs, lacks clinical utility.

The correlation of many radiomic features with ROI size [20, 21] (BM volume or diameter in our study) also demonstrates a challenge for predicting BM SRS outcomes. When comparing the 74% of radiomic features that were volume-correlated in this study to the important features identified in other BM SRS studies, it was found that 60–73% of important features reported elsewhere were likely volume-correlated [12–14, 16]. The exception was the study of Jiang et al., but since their methodology rejected BMs < 10 mm in diameter and did not find volume to be a univariate predictor of SRS response, it is not surprising that no volume-correlated radiomic features were found to be important [15]. Gutsche et al.
removed volume-correlated radiomic features, but only at a correlation coefficient > 0.9, and so retained many features that remained strongly correlated with volume [16].

Our study found that by effectively blinding the model to BM volume through the removal of volume-correlated features, balanced accuracy for small and large BMs could be achieved. This result may initially appear counterintuitive. One would wonder: if BM volume is such a strong univariate predictor of SRS outcomes, why does removing it and other volume-correlated features not significantly decrease accuracy? The reason for more balanced performance between small and large BMs is similarly unclear. We speculate that the reason for this is that small BMs receive more aggressive prescriptions and are less hypoxic/sufficiently vascularized only in general. Therefore if the model is blinded to volume, but still has access to direct information about hypoxia/vascularization through the remaining radiomic features, overall accuracy should not fall. Furthermore, by only having these features more specific linked to SRS response, we speculate that the model would tend towards making more nuanced predictions for large BMs.

We identified an optimal correlation threshold of 0.25; with a less strict threshold (> 0.40), accuracy remained imbalanced between the volume groups, and with a stricter threshold (< 0.1), no further features were identified for removal. The volume-correlation threshold of 0.25 needs to be externally validated in future work. It should also be noted that our results were presented only for a single, but clinically relevant, volume threshold of 7.5 cc. While this volume threshold was relevant at the centre at which the study data were collected, external validation of this threshold, or thresholds more relevant at an external centre, is also required.

2.4.4 Feature importance analysis

In our study, novel and previously reported radiomic features were found to predict BM SRS outcomes. Of the eight highly important radiomic features not correlated to volume, two were also found to be highly important in other studies: GLCM cluster shade [14] and
kurtosis [13, 16]. This is the first report of multiple radiomics features being reproducibly identified as important across more than one centre for predicting BM SRS response. The remaining six highly important radiomic features were newly identified; their importance may be unique due to our sample, removal of volume-correlated features, or MRI pre-processing.

Interpretation of radiomic features, and especially their complex interactions within a ML model, is notoriously difficult. Previous literature has linked qualitative BM appearance and SRS outcomes [7, 8]. While this study has linked some radiomic features with BM progression post-SRS, future study should more directly link radiomic features with qualitative BM appearance to aid in interpretability.

2.4.5 Effects of multiple magnetic resonance scanner models

The accuracy variability observed across scanner models shows that not all scanner differences could be normalized by our methodology. We speculate that the different native resolution of the Avanto scanner compared to the Expert and Vision scanners explains the decrease in accuracy when using Avanto scanner images. A dataset with more scanner models and resolutions would be required to confirm this. The majority of the Avanto scanner images (37 BMs) were acquired sagittally, but eight BMs were acquired as slightly thicker axial slices (table 2.2). The Expert and Vision scanners exclusively acquired sagittal slices. To ensure this difference was not a factor in the observed accuracy variability, the experiments were repeated omitting the eight BMs acquired axially. These results were consistent with those presented previously, indicating that mixed slice orientations did not cause the accuracy variability.

Our findings on the effects of mixed MR scanner models are relevant to clinical translation. When using only the more closely matched Expert and Vision scanners, we achieved AUC 0.84 despite using half of our study sample (68 BMs). This AUC exceeds or is comparable to the current state of the field. Given the drop in accuracy when including the
Avanto scanner, it is likely that current literature results are optimistic compared to expected results during external validation using a different scanner model. Some studies in this area have used datasets with more than one scanner [13, 16, 17]. Mulford et al. employed a 1.5 T and 3 T scanner, but 96% of the data was from the 1.5 T scanner, likely masking any 3 T scanner effects. Studies by Wang et al. and Gutsche et al. also used a 1.5 T and 3 T scanner, but in both cases the ratio of data from each scanner and scanner model analysis were not provided, making drawing conclusions challenging. The magnitude of the drop in accuracy when including Avanto scanner is also important to consider. AUC was found to decrease up to 13.0% when introducing the Avanto scanner, which exceeds the AUC changes observed when altering the available features. It is therefore critical to characterize and correct for the effect of MR scanner variability and not only the available features, as has been the standard practice to date in the literature.

Our study’s methodology was unable to normalize for differences across all scanner models, but further techniques could be investigated. The instability of radiomic features across MR scanner models has been previously reported [22, 23, 31, 32]. Research into producing more generalizable models across imaging scanners has shown the success of pre-processing normalization, post-feature extraction harmonization, and data augmentation [22, 31–33]. The Z-score transform performed in this study is but one of these techniques, though was not found to operate optimally. A robust comparative study of these techniques is required to produce a model that will more likely maintain performance across multiple centres or can be calibrated for use per MR scanner.

2.4.6 Limitations and future work

It is important to also consider the limitations of this study with respect to the clinical endpoint considered. First, this study examined only the endpoint of per BM post-SRS progression, whereas prediction of progression free survival and OS may also hold clinical utility. While BM MRI radiomics are the most likely to be linked to per BM pro-
gression, radiomics of the entire brain, primary cancer, and extra-cranial metastases may enhance predictions of these additional per patient endpoints. Second, to definitively differentiate pseudo-progression against true progression post-SRS, biopsy would need to be performed on each BM that progressed, as reliable imaging techniques to distinguish pseudo-progression from true progression are not yet available [34]. This study did not use this approach due to its retrospective nature, as biopsies to confirm pseudo-progression are not routinely gathered for all patients in standard clinical practice. While prospective studies of SRS outcome modeling would benefit from more robust pseudo-progression diagnostic techniques as outlined in the Response Assessment in Neuro-Oncology Brain Metastases (RANO-BM) protocol [34], we believe our study’s methodology still provides insight into SRS outcome prediction that could motivate these prospective studies. Third, this study’s endpoint measurements were performed before the publication of the RANO-BM protocol [34]. The volumetric measurement then used in this study does provide a more comprehensive description of the change in BM size compared to the RANO-BM protocol, but its inability to be directly compared to studies utilizing RANO-BM does provide a limitation to this study. Fourth, another limitation of this study’s endpoint is the variable follow-up time for each patient due the short OS for many BM patients. This shortcoming is inherent to all BM studies, as by definition all BM patients are in the most advanced stage of their cancer. Patients lost to follow-up early on may then have had BMs that would have progressed if the patient had survived long enough, but because of discontinued follow-up, the BMs would be scored as non-progression. This limitation therefore affects the dataset’s endpoints for some metastases, and so the results presented here, and in the field more broadly, should be viewed as conservative compared to expected results if the true endpoint was known for each BM.

This study was also limited due to constraints of its retrospective design. First, as with many radiomics ML studies, the small, single centre, retrospective study sample limit the strength of the conclusions. Second, the study sample’s patients were treated using a
linac between 2003 and 2011. While this study presents the first results on SRS outcome prediction performed on a general purpose linac (whereas previous studies examined only Gamma Knife or CyberKnife), the results may not generalize to alternative or more recent SRS treatment modalities. Third, this study only utilized clinical features that were readily available at the institution where the data was collected, and so the results presented are specific to these clinical features and associated scorings. While this study offered the widest variety of clinical features to date in the field, there are further clinical features to examine that may hold promise. In particular, biomarkers specific to certain primary cancer sites could present unique opportunities if used with site specific models. Fourth, while the bootstrapped experimental design used is a robust technique free of data leakage between the training and testing datasets, the use of an independent testing dataset would ensure that overfitting to the dataset was prevented. We are currently conducting a study at a separate centre that will address these limitations. This study will contain a larger and more recent sample of linac-based SRS patients. This study will also allow for external validation of the models produced here and further investigation of MR scanner model variation.

2.5 Conclusions

In conclusion, this study showed that predictive accuracy is highest when using clinical and radiomic features together, and depends highly on primary cancer site, BM volume and the inclusion of multiple MR scanner models. We found that performance was particularly high for colorectal and skin cancer patients, and so our techniques may be of more benefit to these sub-populations. We also showed that accuracy dependency on BM volume can be eliminated without sacrificing overall accuracy by removing volume-correlated features, with radiomic features uncorrelated with volume becoming more important. We believe before such prognostic models can be externally validated or clinical translated, it is imperative that these effects be analyzed, understood, and accounted for in future exploratory
By doing so, we believe that prognostic models suitable for clinical implementation can be produced that can aid clinicians in prescribing SRS to minimize risk to patients and BM progression post-treatment.

2.6 Data and code availability

The ground truth labels, sample stratifying variables, and predicted progression probabilities from each reported experiment’s models to replicate this study’s analysis are available for use at the following URL: https://github.com/baines-imaging-research-laboratory/radiomics-for-srs-performance-sensitivity-data-share.

All the computer code used for the study’s experiments, analysis and figure generation are available online at: https://github.com/baines-imaging-research-laboratory/radiomics-for-srs-performance-sensitivity-code.

2.7 References


Chapter 3

Assessment of brain metastasis qualitative appearance interobserver variability and comparison to magnetic resonance imaging radiomics for stereotactic radiosurgery outcome prediction

The following research chapter was completed to meet thesis research objectives 3–5 outlined in section 1.5:

3. To characterize the interobserver variability of BM qualitative appearance labelling and its impact on risk stratification for post-SRS progression.

4. To compare using qualitative appearance-based analysis versus quantitative radiomics-based analysis of BMs in T1w-CE MRI to predict post-SRS progression.
5. To interpret radiomics-based ML models by relating their operation to biologically-relevant BM qualitative appearances.

This chapter takes the radiomics-based ML models and associated methodologies developed in chapter 2 and compares them to using qualitative appearance labelling to predict post-SRS progression. The qualitative appearance labelling is also used to provide interpretation of the results from chapter 2 with respect to possible underlying biological mechanisms.

3.1 Introduction

BMs form when cancer spreads to the brain, and are a hallmark of advanced disease. As improvements in cancer treatments have increased patients’ life expectancies, their risk of developing BMs at some point during the course of their disease has increased to 10–20% [1]. Given the symptoms and risks associated with BMs, and the short median survival of 8–16 months [2] associated with BMs, it is critical that BM patients receive the most appropriate treatment as soon as possible.

Treatment options for BMs currently include surgical resection, WBRT, or SRS [3, 4]. Surgical resection is recommended for patients with a favourable prognosis who present with few and accessible BMs. WBRT irradiates the entire brain and is non-invasive, avoiding standard surgical risk, but has a higher prevalence of neurocognitive decline post-treatment. In contrast, SRS limits the risk of neurocognitive toxicity by targeting only the BMs with radiation delivered with high doses in 1–3 fractions [5]. SRT is similar to SRS, except that it is delivered in more than three fractions, usually to larger BMs or post-surgical cavities.

SRS is highly effective, but up to 30% of treated BMs can progress post-SRS, constituting a treatment failure [5]. Escalation of SRS dose may reduce this failure rate, but may also increase the risk of toxicity [6, 7]. Therefore, developing predictive models of BM
response to SRS would aid in decision making to balance the risk of treatment failure and toxicity.

Previous studies have used qualitative interpretation of BM appearance in pre-treatment T1w-CE MRI or x-ray CT to predict the outcome of WBRT [8] and SRS [9–13]. These studies labelled BMs as either being “homogeneous”, “heterogeneous”, or “ring-enhancing”, and generally found that “homogeneous” BMs had the lowest risk of progression post-SRS, followed by “heterogeneous”, and then finally “ring-enhancing” BMs. It was hypothesized that the uniform uptake of contrast by a BM’s vasculature shown by “homogeneous” enhancement indicates strong oxygenation, which is linked to enhanced cell kill from radiation exposure.

Rodrigues et al. [14] presented a more advanced BM qualitative appearance interpretation using five (instead of three) qualitative labels: “homogeneous”, “heterogeneous”, “cystic (simple)”, “cystic (complex)”, and “necrotic”. In a RPA on multiple outcome predictors, they produced a model (henceforth referred to as “the RPA model”) that stratified BMs by risk of progression post-SRS using the SRS dose and fractionation prescription, and BM appearance (figure 3.1a). The BMs receiving a less aggressive SRS prescription were at a higher risk of progression, and within this group, BMs labelled as “heterogeneous” or “necrotic” were at the highest risk of progression. In contrast to the other appearance scoring approaches described above, this RPA model was specifically developed for outcome prediction using multiple variables and uses the most descriptive qualitative appearance labels to date, and so represents the most advanced predictive model of progression post-SRS using qualitative appearance scoring. While this RPA model has potential clinical benefit, the interobserver variability of the qualitative appearance labelling has not yet been measured.

More recently, ML and quantitative radiomic analysis of MRI have also been used to successfully predict outcomes of SRS [15–23], post-surgical SRS [24], and SRT [25–28]. Radiomics involves extracting quantitative features from medical images, which are then
Figure 3.1: Models used for risk stratification of BMs for progression post-SRS. Each model stratifies BM into four risk groups, with group “1” intended to be BMs at the lowest risk for progression and each next group being at higher risk for progression. (a) shows the original RPA model. The “Clinician Observer” could be any of the five observers, as well as the expert observer consensus. (b) shows the replacement of the “Clinician Observer” in (a) with the five radiomic appearance label experiments’ results, each using one of the five appearance labels scored by either Expert 1 or the expert consensus. (c) shows the integration of the radiomic progression experiment results into the RPA model structure. While the left and right branches use the same experimental results, they use different ROC operating points to separate “low” versus “high” probabilities of progression. The direct model in (d) is no longer based on the RPA model, but rather only on the radiomic and clinical progression experiment’s results. Three ROC operating points (one at each split in the tree) were used to assign a risk group for a BM’s given probability of progression. Supplementary figure S2 provides further details on the ML experiments used for (b–d) and how ROC operating points were chosen for the probability splits in (c) and (d). Abbreviation: fractions (fx)
typically coupled with complex ML models to predict a value of interest [29, 30]. While these radiomic-based ML models provide objective, quantitative, automated, and highly accurate prediction of SRS outcomes, their complexity and the abstract nature of their underlying radiomic features causes their operation to remain largely uninterpretable. This lack of model interpretation can make clinical translation and future research hypothesis generation more difficult.

Comparison between qualitative and quantitative approaches to BM SRS outcome prediction is also important. Kawahara et al. [18] and Gutsche et al. [16] both produced radiomic ML models predicting post-SRS progression, and compared them to a simple qualitative model using only the three-way “homogeneous”, “heterogeneous”, or “ring-enhancing” appearance labelling. Both found their ML models to be superior to the qualitative models, which were found to have low accuracy (44–62%). Neither study examined the more specific five-way appearance labelling of the RPA model, the more clinically relevant multi-variable RPA model using this labelling, nor performed risk stratification analysis to directly compare their results to the original BM qualitative appearance studies. Radiomic ML model operation could also be interpreted using qualitative appearance labels, allowing for model behaviour to be linked to biological hypotheses, encouraging clinical translation and informing research efforts.

Given the open questions surrounding the use and comparison of qualitative and quantitative MRI analysis for BM SRS outcome prediction, we conducted a study that examines the interobserver variability of qualitative appearance labelling, compares qualitative and quantitative techniques, and uses qualitative appearance labels to interpret complex radiomic systems. Our study therefore addresses the following research questions:

1. What is the interobserver variability of the BM qualitative appearance scoring used by the RPA model?

2. Does the interobserver variability in appearance scoring decrease the RPA model’s
ability to stratify BMs for risk of post-SRS progression?

3. Can radiomics-based ML models replicate the qualitative appearance labelling performed by clinicians, and what impact do these models’ inaccuracies have on risk stratification using the RPA model?

4. Do radiomics-based ML models directly predicting post-SRS progression provide enhanced risk stratification compared to qualitative appearance-based models?

5. Do BMs of different qualitative appearance labels have different probabilities of post-SRS progression predicted by radiomics-based ML models?

6. Are the feature importance scores of radiomics-based ML models correlated with biologically-relevant qualitative appearance labels?

3.2 Methods

3.2.1 Study sample

Our study’s sample consisted of 99 patients randomly selected from the original cohort studied by Rodrigues et al. [14], for which the raw pre-treatment T1w-CE MRI was made available. The original cohort was collected retrospectively at the AUMC (The Netherlands), and so excluded patients with highly symptomatic BMs or poor prognosis. For the retrospective data collection, only patients who received first-line SRS for newly diagnosed, radiologically confirmed BMs were included. Any patients without pre-treatment or follow-up MRI were excluded. All patients were treated between 2003–2011 using linear-accelerator based SRS on a Novalis/Novalis TX unit (BrainLAB, Feldkirchen, Germany). BMs were prescribed 15, 18, or 21 Gy in one fraction, or 24 Gy in three fractions, with smaller BMs receiving the more aggressive prescriptions. Up to three BMs were treated concurrently per patient, and so a total of \( n = 123 \) BMs were individually analyzed.
3.2.2 Clinical features and study endpoint

For each patient and BM, a set of 12 clinical features was available. These features were patient sex, age, primary cancer active status, primary cancer site, primary cancer histology, extracranial metastases status, systemic therapy status, neurological symptoms response to corticosteroids, and ECOG score, along with per BM volume, location (supra versus infratentorial), and SRS prescription. Supplementary table B.1 provides a summary of all the clinical features and their distributions for our study sample. Since Kaplan-Meier (KM) analysis was performed, post-SRS survival or follow-up length per patient was also collected to provide censorship data.

For each BM, post-SRS progression was defined radiographically using longitudinal post-treatment MRI. As the original dataset was collected before the introduction of the standardized RANO-BM protocol [31], a non-standard measurement technique was used. In this technique, each BM’s maximum diameters in three perpendicular directions (superior-inferior, mediolateral, posterior-anterior) were measured by a single, expert radiation oncologist and then their product was taken. If this product increased by ≥ 25% in any of the follow-up MRI post-SRS, the BM would be scored as “progression” (or positive/+ ) and scored as “no progression” (or negative/− ) otherwise, defining a binary progression label.

A confounder in all BM studies that use radiographically-defined endpoints is pseudo-progression, in which a BM can appear to grow in size, but this growth is unrelated to true cancerous progression [6]. All BMs scored as progression were reviewed by an expert clinician, who used longitudinal MRI and accompanying patient medical records to determine if pseudo-progression was present. If pseudo-progression had occurred, the BM was rescored as “no progression” (−).
3.2.3 Imaging data and radiomic features

For each patient, pre-treatment T1w-CE MRI and BM ROI contours were collected. T1w-CE MRI was acquired pre-SRS for radiation treatment planning. Five scanner models were represented in our study sample, across which a total of eight different acquisition orientation and voxel size configurations were used (see supplementary table B.2 for full details). Each BM’s ROI was defined by the GTV planning contour that was manually drawn in three-dimensions by an expert clinician using the outer edge of the MRI contrast enhancement.

The MRI data was pre-processed and then radiomic features were extracted for each BM’s ROI. To account for variability in voxel size and intensity scaling across MR scanner models, the MRI data was pre-processed by first using the mean and standard deviation of all voxel values within the brain to apply a Z-score normalization at three standard deviations [32]. The image was then linearly interpolated to a common voxel size of 0.5 × 0.5 × 0.5 mm³. After pre-processing, 107 radiomic features were extracted from the MRI for each BM’s ROI using PyRadiomics v3.0.1 [29] in Python v3.6.13 with 64 intensity bins used where applicable (full feature list in supplementary table B.3).

3.2.4 Machine learning experimental design

Our ML experiments were all conducted using Matlab 2019b v9.7.0.1190202 (The Mathworks Inc., Natick, USA) based on a common template. This template consisted of a RDF model that was trained and tested using a 250-iteration bootstrapped resampling technique across patients to avoid data leakage. During each iteration, the training dataset would be used to perform inter-feature correlation filtering, hyper-parameter optimization, and training of the RDF, which then would be tested using the iteration’s testing dataset (further details in supplementary figure B.1 and table B.4).

For each ML experiment performed in our study, error metrics were calculated by ag-
Table 3.1: Study observers that provided qualitative appearance labels for all BMs.

<table>
<thead>
<tr>
<th>Observer</th>
<th>Clinical Speciality</th>
<th>Medical Centre</th>
<th>Initials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert 1</td>
<td>Radiation Oncology</td>
<td>VUMC, The Netherlands</td>
<td>J.Z.</td>
</tr>
<tr>
<td>Expert 2</td>
<td>Radiation Oncology</td>
<td>LHSC, Canada</td>
<td>J.L.</td>
</tr>
<tr>
<td>Expert 3</td>
<td>Neuro-radiology</td>
<td>LHSC, Canada</td>
<td>A.L.</td>
</tr>
<tr>
<td>Trainee 1</td>
<td>Radiation Oncology</td>
<td>LHSC, Canada</td>
<td>T.T.</td>
</tr>
<tr>
<td>Trainee 2</td>
<td>Neuro-radiology</td>
<td>LHSC, Canada</td>
<td>A.A.</td>
</tr>
</tbody>
</table>

To investigate the question of interobserver variability in appearance labelling and replicability of the RPA model (research question 1), we recruited four clinicians from the London Health Sciences Centre (LHSC) in Canada (see table 3.1) to repeat the labelling of each BM to compare against Rodrigues et al.’s original observer (henceforth referred to as “Expert 1”) [14]. This new set of clinicians relabelled all BMs as “homogeneous”, “heterogeneous”, “cystic (simple)”, “cystic (complex)”, or “necrotic” using only the same pre-treatment T1w-CE MRI available originally to Expert 1. We developed a custom application in Slicer v4.11.20210226 [34] that was used to sequentially display interactive axial, sagittal, and coronal views of each BM to the observers and record their appear-
ance labelling results. After the appearance labelling, confusion matrices across the five appearance labels were created for each possible pair of observers, agreement rates were calculated (number of BMs with same appearance label from both observers divided by total number of BMs), and Fleiss’ kappa test was performed using SAS v9.4 (SAS Institute Inc., Cary, USA).

To address research question 2, we then applied the RPA model to each BM (figure 3.1a), but with the appearance labels from each observer. KM analysis of the BM risk of progression stratification was then performed per observer and SAS was used to perform the KM log-rank test across all groups. The appearance labels across the three expert observers were also used to generate a set of “expert consensus” qualitative labels, in which BMs received the common label chosen by a majority of the expert observers. Separate KM analysis was also performed using this expert consensus labelling with the RPA model.

3.2.6 Qualitative appearance machine learning models

We explored research question 3 on replicating qualitative appearance labelling with ML by first using the BM qualitative appearance labels from Expert 1 to train ML models. As shown in supplementary figure B.2a, bootstrapped resampling ML experiments based on the previously described template were performed using the radiomics features and one of the appearance labels as the model output. Therefore, an ML experiment was performed to label BMs as “homogeneous” or “not homogeneous”, another for labelling them as “heterogeneous or “not heterogeneous”, and similarly for the remaining appearance labels. The term “radiomic Expert 1 appearance experiments” is defined here to refer to these ML experiments.

The results of the five radiomic Expert 1 appearance experiments were first used to calculate error metrics associated with replicating each appearance label. For each experiment, the testing dataset prediction probabilities could be aggregated across the bootstrapped iterations to provide an average prediction probability per BM. The distance of this average
probability per BM from each experiment’s optimal ROC operating point could then be found, and the maximum value taken across the five appearance label experiments to assign a qualitative appearance label to each BM (see supplementary figure B.2a). The RPA model was then used again, except with the appearance labels assigned by the ML models (as shown in figure 3.1b), and the same KM analysis performed. This entire methodology was then repeated in additional experiments to also explore using the expert consensus appearance labels to train the ML models instead of the appearance labels from Expert 1. These experiments will be referred to as the “radiomic consensus appearance experiments”.

3.2.7 Post-stereotactic radiosurgery progression machine learning models

To address research question 4, we then examined the scenario in which instead of training ML models to provide qualitative appearance labels, the radiomic features would be used to directly predict if a BM would progress post-SRS. A ML model would therefore be unconstrained by trying to replicate clinician-based qualitative appearance labels and be able to find arbitrary radiomic signatures that were the most predictive of BM SRS response. This is the same technique employed by other ML studies, and so allowed our study to compare between the qualitative-based and quantitative-based approaches to SRS outcome prediction.

First, an ML experiment was conducted using only the radiomic features as the model input and BM post-SRS progression as the output using the common experiment template (see supplementary figure B.2c). This experiment will be henceforth referred to as the “radiomic progression experiment”. Error metrics were calculated from the experiment results along with average predicted probabilities of progression per BM across the bootstrapped iterations. As shown in figure 3.1c, the SRS dose and fractionation prescription split from the RPA model was still used to provide the first stratification of BMs, but after this the average predicted probability of progression from the ML experiment replaced the qual-
tative appearance labels to perform the final stratification splits. To provide comparison to the other RPA model results, the same KM analysis was performed on this new BM risk stratification.

Next, a second ML experiment was conducted using both the radiomic and clinical features to predict post-SRS progression, henceforth referred to as the “radiomic and clinical progression experiment” (see supplementary figure B.2d). This approach allowed for the RPA model to be completely unused, with the ML model instead being able to completely optimize the use of clinical and radiomic features to make its post-SRS progression predictions. To provide comparison to the RPA models results, a stratification with four risk groups was still performed, and so a two-layer splitting of BMs based on their average predicted progression probability from the ML experiment was used, as shown in figure 3.1d.

3.2.8 Post-stereotactic radiosurgery progression machine learning model interpretation

To answer research question 5 and gain insight into the radiomic signature from the radiomic progression experiment, we first analyzed if the predicted probabilities from the radiomic progression experiment’s ML models were related to the more interpretable qualitative appearance labels. To do so, we took the average predicted probabilities of post-SRS progression per BM and then grouped them based on the expert consensus appearance labels (see supplementary figure B.2b). The statistical distributions of the predicted probabilities per appearance label were then compared using the Kruskal-Wallis test to compare distribution separation across all appearance labels.

We also analyzed the feature importance scores inherently provided by the RDFs trained within our experiments to investigate research question 6 by examining if any features were important in both the radiomic progression experiment and the radiomic consensus appearance experiments. For a given experiment, the RDF importance scores were normalized
between 0 and 1 for each bootstrapped iteration, with 1 representing the most important feature. Features that were removed before model training by the inter-feature correlation filter received a score of 0. Each feature’s scores were then averaged across all bootstrapped iterations, and then renormalized between 0 and 1. We then selected the highly important features from the radiomic progression experiment (importance score ≥ 0.75) and determined whether these features were similarly important for any of the radiomic consensus appearance experiments.

Lastly, the identified highly important features underwent accumulated local effects (ALE) analysis. ALE analysis provides plots for each feature in which the change in a complex ML model’s predicted probability (of post-SRS progression or an appearance label) is shown on the y-axis as a function of the change in the feature’s value along the x-axis, spanning from the feature’s minimum to maximum value represented in the dataset. ALE plots therefore show if a feature leads to an increase or decrease in a model’s predicted probability (negative or positive ALE plots values), as well as at which values of the feature these changes in predicted probability occur. The calculation details for ALE plots are given in the following references [35, 36], which we applied by calculating a feature’s ALE values for each bootstrapped iteration in an experiment, and then averaging across iterations to get a single ALE plot per feature. This calculation was performed for each highly important feature across the six radiomic progression and consensus appearance experiments. For a single feature, the ALE plots of post-SRS progression and a given qualitative appearance were compared by taking the Pearson correlation coefficient (\(\rho\)) between them. If a strong positive correlation coefficient was found, then the feature predicted both progression and the qualitative appearance for the same values of the feature.

3.2.9 Ethics declaration

The collection and analysis of the retrospective patient data used in this study was approved by the AUMC Medical Ethical Review Committee and was conducted within the approved
guidelines. As the study was retrospective on a cohort of deceased patients, written consent from study participants was waived by the AUMC Medical Ethical Review Committee.

3.3 Results

3.3.1 Qualitative appearance interobserver variability

Using the 123 BM sub-sample from the original Rodrigues et al. study [14], we re-established the baseline performance of the RPA model using the BM appearance labels from Expert 1 as shown in figure 3.2a’s KM analysis for progressive disease. As would be expected, the KM results closely matched those of the original study and demonstrated statistically significant risk stratification, though at a lower level of significance due to the reduced number of samples. This first result therefore acted as a validated baseline of risk stratification performance against which to compare further results.

When observers relabelled the BM qualitative appearances, agreement with Expert 1’s labels ranged between 32.5–50.4% (see table 3.2). The relabelling observers’ results were also compared to each other, with Expert 2, Expert 3, and Trainee 2 showing higher agreement rates between 61.8–65.9% (table 3.2). Across all observers, Fleiss’ kappa test confirmed low agreement at $\kappa = 0.38$, with $\kappa = 0.33$ across only expert observers. The labelling confusion matrices showed that across all observers, 32.5% of disagreements involved “heterogeneous” labels, 23.8% included “necrotic” labels, and the remaining labels were each involved in 14.1–15.0% of disagreements (all data in supplementary table B.5). Similarly, when considering only the expert observers, the disagreement rates were 30.9% “heterogeneous”, 23.6% “necrotic”, and 13.6–16.2% for each of the other labels.

We found the labelling disagreement between observers impacted the RPA model risk stratification, with the additional expert observers (Experts 2 and 3) having KM log-rank test $p$-values an order of magnitude greater than Expert 1 (Expert 2: 0.02, Expert 3: 0.03), and Trainees 1 and 2 achieving more comparable results (Trainee 1: 0.003, Trainee 2: 0.002).
Figure 3.2: KM analysis for risk of a BM progressing post-SRS for the six primary risk stratification models evaluated (a–f). The risk group number for each risk curve is labelled on the right y-axis, and the number of BMs at risk per 3-month follow-up interval is given below each x-axis. The stated $p$-values are from the log-rank test performed over all risk groups.
Table 3.2: Qualitative appearance labelling agreement rates between all pairs of observers.

0.005). KM plots per observer are given in supplementary figure B.3, but in summary, all observers produced a high-risk group with a 45.2–53.3% final risk of progression post-SRS, but the remaining three risk groups highly varied in amount of separation and final levels of progression risk.

The expert consensus appearance labelling showed a similar risk stratification to Expert 1 (figure 3.2b) at a comparable level of statistical significance ($p = 0.009$ versus $p = 0.007$ for Expert 1). Risk groups 2 and 3 displayed slightly decreased separation for the expert consensus labelling, and the highest risk group (group 4) did not reach the same end level of risk at 18 months for the expert consensus (49.8%) compared to Expert 1 (54.4%).

### 3.3.2 Qualitative appearance machine learning models

The radiomic Expert 1 appearance experiments showed varied accuracies when using radiomic features to perform each appearance labelling. The “homogeneous” results showed the highest $\text{AUC}_{0.632+} = 0.92$, while $\text{AUC}_{0.632+}$ values of 0.77, 0.83, 0.88, and 0.72 were achieved for the “heterogeneous”, “cystic (simple)”, “cystic (complex)”, and “necrotic” la-
Figure 3.3: Error metrics from the radiomic consensus appearance experiments. As each appearance label (e.g. “homogeneous”) had specific models trained to make a binary labelling decision (e.g. “homogeneous” or “not homogeneous”), error metrics for each appearance label are presented. The error bars for the non-AUC\textsubscript{0.632}+ error metrics represent the 95% CI of each value determined from the 250 bootstrapped resampling iterations.

As can be seen in figure 3.2c, when these qualitative appearance ML model results were used with the RPA model (as per figure 3.1b), the KM risk curves were negatively impacted considerably ($p = 0.08$).

The radiomic consensus appearance label experiments showed all appearances achieved AUC\textsubscript{0.632}+ $\geq$ 0.84 (figure 3.3). “Cystic (complex)” was labelled with the lowest MCR = 14.1% and highest AUC\textsubscript{0.632}+ = 0.95. “Homogeneous”, “heterogeneous”, and “necrotic” were labelled with similar accuracy (AUC\textsubscript{0.632}+ 0.84–0.85), with the highest MCR reported for “heterogeneous” (26.6%). The observed inaccuracy also negatively impacted the RPA model risk stratification (figure 3.2d), but statistical significance was retained ($p = 0.04$).

### 3.3.3 Post-stereotactic radiosurgery progression machine learning models

The radiomic progression experiment directly predicted progression using only radiomic features with AUC\textsubscript{0.632}+ = 0.74. When using direct progression prediction instead of the
qualitative appearance labels with the RPA model (as per figure 3.1c), the KM analysis revealed enhanced risk stratification (figure 3.2e, \( p = 0.0003 \)). In particular, the analysis produced the risk groups with the lowest and highest risk of post-SRS progression (6.6\% and 58.5\%, respectively). Interestingly, risk group 2 was at higher risk for progression compared to group 3, the inverse of the RPA model.

The radiomic and clinical progression experiment achieved \( \text{AUC}_{0.632} = 0.77 \), and when used to stratify BMs into risk groups (as per figure 3.1d), improved stratification was achieved compared to the RPA model results (figure 3.2f, \( p = 0.0006 \)), with the results comparable to the previous stratification using the radiomic progression experiment (figure 3.2e).

### 3.3.4 Post-stereotactic radiosurgery progression machine learning model interpretation

We found that the average probability of post-SRS progression values per BM from the radiomic progression experiment were significantly different when compared across all appearance labels (figure 3.4a, Kruskal-Wallis \( p = 0.0005 \)). BMs labelled “homogeneous” by expert consensus were associated with the lowest median progression probabilities, while “necrotic” had the highest. “Heterogeneous” demonstrated the largest interquartile range, nearly twice that of any other label. Ad hoc Wilcoxon rank-sum tests were then performed between pairs of labels, with three showing a significant difference (\( \alpha = 0.005 \) after Bonferroni correction for 10 comparisons): “homogeneous” versus “cystic (complex)” or “necrotic”, and “cystic (simple)” versus “necrotic” (figure 3.4a).

The large “heterogeneous” probability interquartile range motivated further analysis between BMs that did progress post-SRS (+) and those that did not (−) for each appearance label. Wilcoxon rank-sum tests between the average predicted progression probabilities of + and − BMs for each appearance label found only “heterogeneous” labelled BMs to be significantly different, both with and without a Bonferroni correction for five comparisons.
Figure 3.4: Comparison of the average predicted probability of progression for each BM from the radiomic progression experiment when grouped by qualitative appearance, as labelled by expert consensus. (a) shows the comparison for each of the five appearance labels, with whiskers indicating the extreme values and outliers classified as being 1.5× the interquartile range from the 25th or 75th percentile. The Kruskal-Wallis test across all groups found $p = 0.0005$, with ad hoc Wilcoxon rank-sum $p$-values shown on the plot for statistically significant comparisons (after Bonferroni correction). (b) shows the splitting of the “Heterogeneous” distribution in (a) based on whether the BMs progressed or did not progress post-SRS, with the Wilcoxon rank-sum test $p$-value shown. Ad hoc Wilcoxon rank-sum tests between the “Heterogeneous (No Progression)” and “Heterogeneous (Progression)” distributions with the other appearance labels found that “Heterogeneous (No Progression)” was significantly different from the “Necrotic” BMs ($p = 0.002$), while “Heterogeneous (Progression)” was significantly different from the “Homogeneous” ($p = 0.0009$) and “Cystic (Simple)” ($p = 0.001$) BMs (significance determined after Bonferroni correction). The size and scale of the y-axes for (a) and (b) are equivalent to allow for simpler comparison.

The “heterogeneous” BMs that did not progress had lower predicted probabilities of progression, while the BMs that did progress had a median predicted probability value greater than that of the “necrotic” BMs (figure 3.4b). Ad hoc Wilcoxon rank-sum tests between the + and − “heterogeneous” BMs and other appearance labels found that after a correction for eight comparisons, the − “heterogeneous” BMs were significantly different from the “necrotic” BMs, while + “heterogeneous” was significantly different from “homogeneous” and “cystic (simple)” (figure 3.4b).

Our feature importance analysis revealed 13 radiomic features that were highly im-
important in the radiomic progression experiment. Ten of these features were second-order texture features, while two were first-order statistical features, and one was a shape and size-based feature, as shown in table 3.3. When compared to each radiomic consensus appearance experiment’s feature importance analysis, “necrotic” was found to have the highest median feature importance score across the same 13 features (0.70), with nine of the features also being in the top 13 most important features for labelling “necrotic” (see table 3.3). “Homogeneous” and “cystic (complex)” both had median importance scores near 0.60, and the remaining appearance labels were lower again near 0.50.

ALE plot analysis showed that the behaviour of many features used for labelling “necrotic” and “cystic (complex)” qualitatively matched that of features predicting progression, as shown in figure 3.5 and supplementary figure B.5. This observation was borne out quantitatively, with the “necrotic” and “cystic (complex)” labels having the highest median ALE $\rho$ values of 0.83 and 0.79, respectively (table 3.3). Only for the “homogeneous” label were the features found to be negatively correlated with predicting progression with median ALE $\rho = -0.39$. “Necrotic” also had the lowest interquartile range of ALE $\rho$ values at 0.20, while “heterogeneous” had the largest at 1.31 (table 3.3).

### 3.4 Discussion

Our study found that interobserver variability of BM qualitative appearance labelling is high, but that an expert consensus produced similar risk stratification results to the original results using Expert 1’s labels and was more readily replicated by radiomics-based ML models ($\text{AUC}_{0.632} = 0.84–0.94$ across appearance labels). We also found that radiomics-based ML models directly predicting progression provided enhanced risk stratification compared to using qualitative appearance labelling, and that their radiomics signature was correlated with those of qualitative appearance labels.
<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Feature Name</th>
<th>Rank</th>
<th>Score</th>
<th>Homogeneous Rank</th>
<th>Score</th>
<th>ALE ρ</th>
<th>Heterogeneous Rank</th>
<th>Score</th>
<th>ALE ρ</th>
<th>Cystic (Simple) Rank</th>
<th>Score</th>
<th>ALE ρ</th>
<th>Cystic (Complex) Rank</th>
<th>Score</th>
<th>ALE ρ</th>
<th>Necrotic Rank</th>
<th>Score</th>
<th>ALE ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLRLM</td>
<td>Gray-Level Non-Uniformity</td>
<td>1</td>
<td>1.00</td>
<td>46</td>
<td>0.33</td>
<td>+0.29</td>
<td>48</td>
<td>0.21</td>
<td>+0.77</td>
<td>11</td>
<td>0.68</td>
<td>−0.55</td>
<td>39</td>
<td>0.39</td>
<td>+0.92</td>
<td>33</td>
<td>0.44</td>
<td>+0.74</td>
</tr>
<tr>
<td>NGTDM</td>
<td>Contrast</td>
<td>2</td>
<td>0.95</td>
<td>24</td>
<td>0.51</td>
<td>−0.09*</td>
<td>16</td>
<td>0.46</td>
<td>−0.31</td>
<td>30</td>
<td>0.50</td>
<td>+0.28</td>
<td>30</td>
<td>0.52</td>
<td>−0.24</td>
<td>26</td>
<td>0.51</td>
<td>+0.78</td>
</tr>
<tr>
<td>GLSZM</td>
<td>Inverse Difference Normalized</td>
<td>3</td>
<td>0.93</td>
<td>40</td>
<td>0.39</td>
<td>−0.49</td>
<td>14</td>
<td>0.50</td>
<td>+0.82</td>
<td>14</td>
<td>0.61</td>
<td>−0.88</td>
<td>50</td>
<td>0.31</td>
<td>+0.82</td>
<td>6</td>
<td>0.72</td>
<td>+0.96</td>
</tr>
<tr>
<td>GLCM</td>
<td>Zone Entropy</td>
<td>4</td>
<td>0.93</td>
<td>25</td>
<td>0.50</td>
<td>−0.42</td>
<td>24</td>
<td>0.40</td>
<td>+0.25</td>
<td>19</td>
<td>0.57</td>
<td>+0.43</td>
<td>17</td>
<td>0.61</td>
<td>+0.99</td>
<td>4</td>
<td>0.74</td>
<td>+0.93</td>
</tr>
<tr>
<td>GLCM</td>
<td>Inverse Variance</td>
<td>5</td>
<td>0.92</td>
<td>7</td>
<td>0.67</td>
<td>+0.98</td>
<td>12</td>
<td>0.52</td>
<td>−0.42</td>
<td>24</td>
<td>0.53</td>
<td>+0.95</td>
<td>15</td>
<td>0.63</td>
<td>+0.79</td>
<td>25</td>
<td>0.51</td>
<td>+0.95</td>
</tr>
<tr>
<td>GLSZM</td>
<td>Zone Percentage</td>
<td>6</td>
<td>0.88</td>
<td>44</td>
<td>0.36</td>
<td>−0.74</td>
<td>57</td>
<td>0.18</td>
<td>+0.41</td>
<td>77</td>
<td>0.07</td>
<td>−0.21</td>
<td>66</td>
<td>0.14</td>
<td>−0.76</td>
<td>32</td>
<td>0.46</td>
<td>+0.80</td>
</tr>
<tr>
<td>GLDM</td>
<td>Dependence Entropy</td>
<td>7</td>
<td>0.86</td>
<td>10</td>
<td>0.63</td>
<td>−0.64</td>
<td>21</td>
<td>0.41</td>
<td>−0.87</td>
<td>32</td>
<td>0.49</td>
<td>+0.83</td>
<td>3</td>
<td>0.80</td>
<td>+0.85</td>
<td>7</td>
<td>0.70</td>
<td>+0.98</td>
</tr>
<tr>
<td>Shape &amp; Size</td>
<td>Surface Volume Ratio</td>
<td>8</td>
<td>0.83</td>
<td>4</td>
<td>0.86</td>
<td>−0.39</td>
<td>31</td>
<td>0.36</td>
<td>+0.32</td>
<td>5</td>
<td>0.85</td>
<td>−0.64</td>
<td>18</td>
<td>0.60</td>
<td>+0.72</td>
<td>12</td>
<td>0.59</td>
<td>−0.29</td>
</tr>
<tr>
<td>GLRLM</td>
<td>Gray-Level Non-Uniformity Normalized</td>
<td>9</td>
<td>0.82</td>
<td>8</td>
<td>0.65</td>
<td>−0.14*</td>
<td>8</td>
<td>0.56</td>
<td>+0.87</td>
<td>23</td>
<td>0.53</td>
<td>+0.72</td>
<td>13</td>
<td>0.64</td>
<td>−0.43</td>
<td>5</td>
<td>0.73</td>
<td>+0.73</td>
</tr>
<tr>
<td>First Order</td>
<td>Kurtosis</td>
<td>10</td>
<td>0.81</td>
<td>13</td>
<td>0.60</td>
<td>−0.67</td>
<td>13</td>
<td>0.51</td>
<td>+0.89</td>
<td>34</td>
<td>0.49</td>
<td>+0.81</td>
<td>12</td>
<td>0.64</td>
<td>+0.93</td>
<td>1</td>
<td>1.00</td>
<td>+0.83</td>
</tr>
<tr>
<td>GLCM</td>
<td>Correlation</td>
<td>11</td>
<td>0.81</td>
<td>56</td>
<td>0.23</td>
<td>−0.53</td>
<td>63</td>
<td>0.14</td>
<td>−0.99</td>
<td>42</td>
<td>0.38</td>
<td>+0.84</td>
<td>64</td>
<td>0.14</td>
<td>+0.90</td>
<td>8</td>
<td>0.70</td>
<td>+0.94</td>
</tr>
<tr>
<td>First Order</td>
<td>Interquartile Percentile</td>
<td>12</td>
<td>0.81</td>
<td>17</td>
<td>0.59</td>
<td>+0.49</td>
<td>1</td>
<td>1.00</td>
<td>−0.34</td>
<td>10</td>
<td>0.77</td>
<td>−0.27</td>
<td>5</td>
<td>0.74</td>
<td>−0.04*</td>
<td>2</td>
<td>0.87</td>
<td>−0.31</td>
</tr>
<tr>
<td>GLCM</td>
<td>Informational Measure of Correlation 2</td>
<td>13</td>
<td>0.78</td>
<td>14</td>
<td>0.59</td>
<td>+0.83</td>
<td>2</td>
<td>0.79</td>
<td>−0.89</td>
<td>26</td>
<td>0.52</td>
<td>−0.31</td>
<td>11</td>
<td>0.67</td>
<td>+0.73</td>
<td>3</td>
<td>0.81</td>
<td>+0.92</td>
</tr>
</tbody>
</table>

Median Values: 7 | 0.86 | 24 | 0.53 | +0.28 | 17 | 0.61 | +0.79 | 7 | 0.70 | +0.83 |

Interquartile Range Values: 6.5 | 0.12 | 31.5 | 0.25 | 0.90 | 24.3 | 0.21 | 1.31 | 19.3 | 0.14 | 1.18 | 30.0 | 0.28 | 1.00 | 21.5 | 0.25 | 0.20 |

Table 3.3: Feature importance ranks, scores, and ALE correlation comparison between the radiomic progression experiment and radiomic consensus appearance experiments. Notes: “ALE ρ” refers to the Pearson correlation coefficient value between a given feature’s ALE plots for the radiomic progression experiment and a radiomic consensus appearance experiment. All ALE ρ values were significant at p < 0.05, except for the three values indicated (*). Underlined “Rank” and “Score” values are for emphasis only to show values within the same range as the radiomic progression experiment’s values (≥ 13 and ≥ 0.75 for importance rank and score, respectively).
Figure 3.5: ALE plots comparing the effect of individual features on model predicted probabilities between the radiomic progression experiment and radiomic consensus appearance experiments. Each row of ALE plots is for a single feature (as shown in the x-axes labels), with each plot comparing the feature’s effect on predicted probability of progression (same thin line across each plot in the same row), to the feature’s effect on labelling BMs with one of the appearance labels (thick lines). Each column of ALE plots therefore shows the effect for multiple features for one appearance label compared to predicting progression. The five features chosen to be displayed are all those that were highly important (importance score $\geq 0.75$) for both the radiomic progression experiment and at least one radiomic appearance label experiment (see table 3.3). ALE plots for the remaining highly important features not shown here can be found in supplementary figure S5. Each ALE plot consists of ALE values for 25 intervals chosen using 25 quantiles of a feature’s values across the entire dataset. For each ALE plot comparison, the Pearson correlation coefficient between the two curves is taken to produce the “ALE $\rho$” values in table 3.3.
3.4.1 Qualitative appearance interobserver variability

The high interobserver variability in appearance labelling highlights the difficulty of using subjective, observer-based models. Across the observers, expertise level did not clearly impact variability, demonstrating the inherent uncertainty in the task. Differentiating between “heterogeneous” or “necrotic” labels was the most difficult, likely due to both appearances presenting with areas of hypointensity in the T1w-CE MRI, and therefore requiring a subtle call of whether necrosis is present. Gutsche et al.’s study using three appearance labels (“homogeneous”, “heterogeneous”, or “ring-enhancing”) employed three observers with much higher interobserver agreement ($\kappa = 0.75$ versus $\kappa = 0.33–0.38$) [16]. This is not entirely unexpected, as the three appearance labels used by Gutsche et al. did not require observers to make the difficult distinction between “heterogeneous” and “necrotic”.

The RPA model was somewhat resistant to interobserver variability, likely due to the first-level split based on the observer-independent SRS dose and fractionation and grouping of two or three appearance labels for each risk group. Despite this, the interobserver variability was high enough that model performance in turn varied widely, demonstrating the difficulty of using a single-observer model in practice. The highest-risk RPA group was the most stable, benefiting from “heterogeneous” and “necrotic” BMs being in the same risk group, which negated the effect of the high interobserver variability observed specifically for this distinction. The expert consensus provided a technique to combat the interobserver variability, producing strong risk stratification, but applying such a technique to clinical practice would be impractical. Currently no standardized protocol for scoring the qualitative appearance of BMs exist, and so developing such a protocol or incorporating hypoxia-specific imaging techniques [37] could possibly reduce interobserver variability, but these methods would need to be developed and validated.
3.4.2 Qualitative appearance machine learning models

Our radiomic Expert 1 and consensus appearance experiments’ results showed that replicating observer-based appearance labelling with radiomics-based ML was not feasible. This resulted in the poorest BM risk stratification due to ML model inaccuracy, which also highlighted the RPA’s model sensitivity to this inaccuracy. Our interobserver variability results showed that BM appearance across labels can be highly similar, and so ML models also appear to struggle to recognize patterns that are specifically indicative of each appearance label. The results from replicating Expert 1 again showed the difficulty in distinguishing “heterogeneous” and “necrotic” BMs, with these labels reporting the lowest ML model accuracy. Using the expert consensus appearance labels did result in more accurate replication by ML models, likely due to more consistent appearance labelling compared to a single observer, but the remaining inaccuracy still severely compromised the risk stratification. Having an objective radiomics model that produces highly interpretable qualitative appearance labels would be ideal for clinical use, but our results demonstrate that such a system is not currently feasible using the techniques we employed.

3.4.3 Post-stereotactic radiosurgery progression machine learning models

Using radiomics-based ML models to directly predict BM progression presents an alternative approach to outcome prediction that our results show is superior for post-SRS progression risk stratification compared to observer-based techniques. While Gutsche et al. [16] and Kawahara et al. [18] showed similar superiority of ML models to the three-way qualitative appearance labelling scheme, our study provides the first comparison of ML models to a more comprehensive five-way appearance labelling scheme and comparison to a more clinically applicable qualitative appearance predictive model built on multi-variate data. These results are also the first to compare radiomics-based ML and qualitative appearance
models using risk stratification KM analysis, instead of the AUC and MCR error metrics typically reported in ML studies. Risk stratification analysis not only allows direct comparison to non-ML studies, but is also highly clinically interpretable.

As we used ML models using only radiomic features to both produce qualitative appearance labels and the probability of post-SRS progression, by comparing the two approaches in the same RPA model (figure 3.1b versus 3.1c), the effect of constraining the ML models to use the five qualitative appearance labels could be seen. Our results clearly show that allowing the ML model to find the optimal radiomic signature to predict progression independently of the five qualitative appearance labels led to the greatest stratification of risk. This effect was so pronounced that the ML model’s risk group 2 was at higher risk for progression post-SRS compared to risk group 3, whereas the original RPA model indicated this would not occur when basing decisions only on the qualitative appearance labels.

The results from the radiomic and clinical progression experiment present the most desirable risk stratification for clinical use (figures 3.1d and 3.2f). The low-risk groups (1 and 2) indicate a population of BMs that would not benefit from SRS treatment modification. Risk groups 3 and 4 represent BMs that may benefit from SRS dose escalation (if feasible with respect to possible adjacent OARs), with group 3 specific to patients with a favourable prognosis, as risk of progression increased after 12 months.

3.4.4 Post-stereotactic radiosurgery progression machine learning model interpretation

The enhanced risk stratification of ML models directly predicting post-SRS progression is encouraging, but they do require interpretation to provide insight into their operation. Our results comparing the radiomic progression experiment’s predicted probability of progression across qualitative appearance labels offers some of these insights (figure 3.4). It appears from this analysis that our ML models trained across bootstrapped iterations, while free to find any radiomic signature predictive of progression, discovered an optimal
radiomic signature related to the qualitative appearance labels. The ML models independently discovered that “necrotic” BMs were at the highest risk of post-SRS progression and “homogeneous” BMs were at the lowest risk, which agrees with previous studies’ hypotheses that implicated hypoxic conditions within the BM in causing reduced SRS effectiveness [11, 12]. The ML models also found radiomic differences in the “heterogeneous” BMs that successfully separated the BMs that progressed post-SRS and those that did not. This indicates that there are additional radiomic differences within the “heterogeneous” BMs that are useful for predicting response, but are not readily visually distinguished. Given the similarity of the predicted probability of progression distributions between the “necrotic” BMs and the “heterogeneous” BMs found to progress, we speculate that there could be a BM population labelled as “heterogeneous” that are hypoxic enough to negatively impact SRS outcomes, but do not yet qualitatively appear necrotic in T1w-CE MRI.

To provide further interpretation of the ML models, we also analyzed if the radiomic progression experiment’s radiomic signature was correlated with those of the radiomic consensus appearance experiments. Our results (table 3.3, figure 3.5) indicated that the progression radiomics signature was most closely correlated with the “necrotic” radiomics signature, both in terms of feature importance scores and correlation between ALE plots. This shows that the radiomics progression experiment not only independently found necrotic BMs to be at highest risk for progression, but also used a similar radiomic signature to do so. The other qualitative appearance radiomic signatures have a lower similarity with the progression radiomic signature in terms of feature importance, but their ALE correlation values generally match the results from figure 3.4 (e.g. “homogeneous” BMs had the lowest predicted probability of progression and the “homogeneous” radiomic signature was also the most anti-correlated with the progression radiomic signature). Furthermore, the “heterogeneous” radiomic signature had the largest interquartile range of ALE correlation values, showing that this radiomic signature was both highly correlated and anti-correlated with the progression radiomic signature. This matches our previous finding of two dis-
tinct subpopulations of “heterogeneous” BMs, which could then lead to a “heterogeneous” radiomic signature with mixed correlation with the progression radiomic signature.

These model interpretation results promisingly indicate a possible connection between the post-SRS progression radiomic signature and necrosis, but the ultimate strength of this connection is limited by a few factors. First, the labelling of BMs as “necrotic” (or any other appearance label), while performed by expert consensus, is unlikely to be exactly reflective of the true underlying biology. Second, the radiomic consensus appearance experiments did not produce perfectly accurate models, and third, the radiomic progression experiment also had model inaccuracy. Therefore, while our results are informative, they are not conclusive.

Previous ML studies interpreting their models’ operation have mostly found texture-based radiomic features to be predictive of SRS response (as opposed to shape/size or first-order statistical features), but further interpretation of these features is difficult [15–26]. Some studies have considered radiomic features from different MRI sequences and ROIs (tumour core, peri-tumoural regions, or surrounding edema), and so some interpretation of these results is possible [16, 17, 19–21, 25, 26]. Studies using CNNs can similarly analyze which data the network was focusing its “attention” when making predictions [27, 28]. Across these studies, however, there is disagreement on the relative value of both different MRI sequences and ROIs. Gutsche et al. found all 10 of their important radiomic features had significantly different values when grouped between “homogeneous”, “heterogeneous”, and “ring-enhancing” BMs [16]. While this analysis technique provides insight into the individual radiomic features used by a ML model, ML models incorporate complex and non-linear relationships between these features. Therefore, the analysis we provide of interpreting the model as a whole through analysis of the model predicted probabilities and ALE plots is critical to provide a more comprehensive model interpretation.
3.4.5 Limitations and future work

Our study’s results need to be considered in the context of our methodology’s limitations. First, our study was retrospective, and so our study sample inherently reflects only the population of BM patients treated at the AUMC during the time of data collection. As our patient sample was treated between 2003–2011, our results need to be confirmed on patients treated more recently due to advancements in imaging technology, SRS techniques, and systemic therapies. Second, due to our small dataset size, we used a bootstrapped resampling ML experimental design. While our methods took care to prevent leakage of the testing dataset into the model training process, an external validation dataset is required to confirm our results. Lastly, our studied endpoint was both non-standard and radiographically defined. As our dataset was acquired before the introduction of the standardized RANO-BM protocol [31], it relies on non-standard BM measurements. While our measurements were more comprehensive than those mandated by RANO-BM [31], they are unfortunately not directly comparable to other studies. Furthermore, our radiographically-defined endpoint is susceptible to pseudo-progression, which cannot be fully controlled for in a retrospective study. Therefore, prospective studies with well-defined protocols for controlling for pseudo-progression are required.

While confirmation of our T1w-CE MRI interpretability results is critical, it is also important to extend our analysis techniques to studies using other MRI sequences and ROIs, especially given their reported predictive value. Such studies would require the development and validation of qualitative appearance labels specific to each MRI sequence and ROI to ensure interpretability is grounded in clinically and biologically relevant explanations. A joint BM MRI radiomic and pathology study is ultimately required to provide definitive explanations of MRI radiomics-based ML models. Such a study is highly feasible given that T1w-CE MRI is routinely collected before most BM surgical resections, and a retrospective study may even be possible if banked or digitized tissue samples are...
available.

3.5 Conclusions

In conclusion, we have compared observer-based qualitative appearance models to ML models using quantitative MRI radiomic features for predicting BM response to SRS. We showed that the interobserver variability of appearance labelling is high and negatively impacts predictive model performance. Trying to replicate appearance labelling with ML models was unsuccessful, motivating the use of ML models that directly predict outcomes, which were found to outperform observer-based models. We also provided interpretation of our radiomics models, showing that their operation was related to qualitative appearance labels, with “necrotic” metastases and a subset of “heterogeneous” metastases predicted to have the highest probability of progression post-treatment. Our results therefore provide a necessary step in model interpretation that is required for the eventual clinical translation of these models that will allow for optimized treatment outcomes for BM patients.

3.6 Data and code availability

The progression labels, qualitative appearance labels, and model prediction probabilities from each reported experiment needed to replicate this study’s analysis are available for use at the following URL: https://github.com/baines-imaging-research-laboratory/radiomics-for-srs-model-interpretability-data-share

All the computer code used for this study’s ML experiments, analysis and figure generation are available online at: https://github.com/baines-imaging-research-laboratory/radiomics-for-srs-model-interpretability-code
3.7 References


Chapter 4

Dual-centre validation of using magnetic resonance imaging radiomics to predict stereotactic radiosurgery outcomes

The following research chapter was completed to meet thesis research objective 6 outlined in section 1.5:

6. To externally validate radiomics-based ML models and the methodology used to train them with datasets from two independent centres.

This chapter externally validates the radiomics-based ML models and associated methodologies developed in chapter 2 by collecting an additional dataset at a second centre for use as an external testing dataset.

4.1 Introduction

Approximately 10–20% of cancer patients develop BMs, the spread of cancer to the brain [1]. With cancer treatment improvements, both BM incidence and BM patient life expectancy has increased. However, the overall prognosis remains poor, with a median over-
all survival between 8 and 16 months [2]. As a result, BM patients must be treated quickly and effectively, while minimizing short and long-term toxicities.

SRS represents an effective BM treatment, in which ablative radiation doses are delivered conformally in one to three fractions to limit normal tissue toxicity [3]. SRT, a variant of SRS, delivers hypofractionated radiation over five fractions to larger BMs or surgical cavities. SRS has important advantages over alternative BM treatment options. Compared to WBRT, where radiation is delivered to the entire brain, SRS limits normal brain tissue irradiation to reduce the risk of long-term cognitive side-effects that negatively impact patients’ QOL [4]. Compared to surgery, SRS is non-invasive and so avoids standard surgical risks.

Despite the ablative nature of SRS, it is associated with treatment failure rates up to 30%, in which the targeted BMs progress post-SRS [4]. Higher prescription doses could decrease failure rates, but would also increase the risk of toxicities [5–7]. Knowing whether SRS is likely to fail at a given dose prescription may help to drive decisions to treat with higher prescription doses, balancing this against the risk of toxicities. Therefore, having predictive models of BM SRS response could aid in treatment selection and optimizing SRS dose prescription.

Previous studies have demonstrated an association between qualitative BM appearance in T1w-CE MRI and SRS outcomes [8, 9]. Specifically, it has been hypothesized that hypointense areas following contrast injection may indicate areas of hypoxia with reduced radiosensitivity.

Recently, quantitative ML techniques have also been investigated to predict SRS outcomes. These techniques rely on quantitative radiomic features extracted from medical images and ML models to make treatment predictions [10, 11]. These studies have produced accurate predictive models incorporating a variety of MRI sequences and clinical datapoints [12–18]. Similar studies have also predicted SRT or post-resection SRS outcomes [19–24]. However, all these studies have used single centre datasets, leaving the
generalizability of these models and techniques an open question.

To motivate clinical translation of ML techniques, multi-centre external validation must be performed to demonstrate generalizability [25]. It is critical to first validate that a model trained at one centre can produce accurate predictions at another centre, as this represents the most straightforward route towards clinical deployment. Alternatively, each centre could train its own model, but this requires external validation of the model training methodology to ensure it reliably creates accurate models at independent centres. We therefore provide the first external validation of ML-based SRS outcome prediction by answering the following research questions:

1. Can a model trained with one centre’s dataset make accurate predictions on another centre’s dataset without retraining?

2. Can the methodology used to produce a model from one centre’s dataset be successfully used to produce a similarly performing model from another centre’s data?

3. Does a model trained on pooled data from two centres offer equivalent accuracy for each centre?

4. When two models are trained using an identical methodology, but at different centres, are the same clinical and radiomic features found to be important?

5. Can the accuracy of models described in research question 1 be improved by either harmonizing feature values across centres, or by utilizing a consolidated set of features that are mutually important across both centres?
4.2 Methods

4.2.1 Centre A study sample

To answer our research questions, we acquired datasets from study samples at two centres. The first study sample was provided by the AUMC in the Netherlands (defined as “Centre A”) and was the identical 99 patient sample used in our previous ML study [17]. The AUMC Medical Ethical Review Committee approved this retrospective data collection, with participant consent waived due to the study being retrospective and on deceased patients.

Centre A’s dataset consisted of imaging and clinical datapoints collected per BM. The retrospective study sample inherently reflects Centre A’s clinical practice when the patients were treated (2003–2011). Our study’s inclusion and exclusion criteria restricted patients to have received first-line SRS and at least one follow-up MRI. The AUMC only used linac-based SRS, with prescriptions ranging from 15–24 Gy in one or three fractions (see table 4.1). \( n = 123 \) BMs across the 99 patients were individually analyzed, with demographics provided in table 4.1. Pre-treatment T1w-CE MRI was collected per BM, along with the GTV contour used for treatment planning as defined by an experienced clinician using the enhancing region’s border. Five MR scanner models were represented in the dataset, with 94% of BMs imaged across three models (see supplementary table C.1).

We defined BM post-SRS progression radiographically using follow-up MRI approximately every three months. The AUMC’s data collection occurred before the introduction of the RANO-BM protocol [26], and so BM size was defined by measuring the maximum diameter in three perpendicular directions (superior-inferior, mediolateral, posterior-anterior). The product of the maximum diameters provided an approximation of BM volume, with an increase of \( \geq 25\% \) indicating progression. Post-SRS progression was therefore defined as a binary label, with a positive (“+”) representing a BM that progressed post-
Table 4.1: Univariate analysis of clinical and radiomic features for Centres A and B to separate BMs that progressed post-SRS from those that did not, and separation of Centre A and Centre B BMs. Notes: p-values were calculated using the Chi-squared test for the categorical clinical features, while the Wilcoxon rank sum test was used for the continuous clinical features and radiomic features. “+ vs. − p-value” represents testing between BMs that progressed post-SRS (“+”) and those that did not (“−”). “A vs. B p-value” represents testing between all BMs from Centre A against all BMs from Centre B. Given the large number of radiomic features, the number of features below α = 0.05 is given for each class of radiomic feature, along with percentage of features from each class this number of features represents. A full table of all features’ p-values are provided in the supplementary material (see section 4.6).

SRS, and negative (“−”) representing no progression. BMs can display pseudo-progression, radiographic growth not linked to cancerous progression [6]. Pseudo-progression was controlled for based on expert clinician judgement using MRI and patient medical records. In total, 22.0% of Centre A’s BMs were scored as true progression.
4.2.2 Centre B study sample

We collected a second, external dataset retrospectively at the London Regional Cancer Program in Canada (“Centre B”). First-line SRS and at least one follow-up MRI were also required. Linac-based SRS was also solely used in one or three fractions with doses of 18–27 Gy (see table 4.1). \( n = 117 \) BMs across 62 patients treated in 2016–2020 were included. GTV contours and T1w-CE MRI were also collected, across which five MR scanner models were represented, with 55% and 30% of BMs imaged across two models (supplementary table C.1). The Western University Research Ethics Board approved this study and waived patient consent, and data collection used the Research Electronic Data Capture (REDCap) platform [27, 28].

To maintain consistency, we used Centre A’s definition of progression for Centre B. However, Centre B’s dataset contained smaller BMs than Centre A’s (Centre A contained no BMs of volume < 100 mm\(^3\)). A percentage-based progression metric was inappropriate for Centre B’s BMs < 100 mm\(^3\), as an insignificant change in absolute BM size would still represent a \( \geq 25\% \) change. This limitation in percentage-based scoring for small BMs was also noted in the RANO-BM protocol for unidimensional measurements, with a 3 mm absolute change in the longest BM diameter recommended by RANO-BM instead [26]. We adapted this technique for 3D measurements, with a \( (3 \text{ mm})^3 = 27 \text{ mm}^3 \) change in size required to constitute progression for BMs in which a 25% change in volume was < 27 mm\(^3\).

We controlled for pseudo-progression using the trajectory of BM size, salvage therapies, and pathology reports (supplementary table C.2). In total, 17.1% of Centre B’s BMs progressed post-SRS. Five BMs were particularly challenging to assess for progression due to immunotherapy or targeted therapies (supplementary table C.2). These BMs could have possibly confounded model accuracy, and so we repeated Centre B’s methodology external validation experiments described below without these BMs. The results changed negligibly, and so we included these BMs.
4.2.3 Clinical features

We also collected pre-treatment clinical features at both centres. As the Centre A and B datasets were retrospective, each centre’s clinical features did not directly overlap, and so we used a common set of eight clinical features (listed in table 4.1). The presence of primary cancer site and histology feature categories at one centre and not the other was accounted for using an “Other” category. We investigated separation of BMs from each centre and of positive and negative BMs using per-clinical-feature univariate statistical tests (table 4.1).

4.2.4 Radiomic features

Before radiomic feature extraction, we pre-processed the pre-treatment T1w-CE MRI to account for variability in voxel size and intensity scaling. Specifically, we computed the mean and standard deviation of the voxels within the brain and not the BMs to apply Z-score normalization to the MR scan at three standard deviations [29]. The voxels were then linearly interpolated to a size of $0.5 \times 0.5 \times 0.5$ mm$^3$. After pre-processing, 107 radiomic features were extracted for each BM’s GTV ROI (feature list in supplementary table C.3) using PyRadiomics library v3.0.1 in Python v3.6.13 [30] with 64 intensity value bins. Univariate statistical tests were also performed on the radiomic features (table 4.1).

4.2.5 Machine learning methodology

For each ML experiment, we used Matlab 2019b v9.7.0.1190202 (The Mathworks Inc., Natick, USA) to train RDF models, with hyperparameter optimization performed before training. We explored three feature combinations: clinical features, radiomic features, and all features (clinical and radiomic). If radiomic features were included, an inter-feature correlation filter removed redundant features with a correlation coefficient $> 0.8$. Some experiments had a single training dataset from one centre and a single testing dataset from the
other centre, while some experiments required using one dataset that was bootstrap resampled into training and testing datasets over 250 repetitions. We developed this methodology in our previous study based only on Centre A’s dataset [17], with further detail provided in supplementary figure C.1 and table C.4.

From each experiment’s testing dataset, we calculated ROC curves and the associated error metrics of AUC, MCR, FNR, and FPR. For non-bootstrap experiments, ROC curves were constructed for both the testing dataset’s samples and the training dataset’s RDF out-of-bag samples. To avoid positive bias from using the testing dataset, we used the training dataset’s ROC to choose the upper-left operating point, minimizing the sum of squares of FNR and FPR. We transferred the operating point to the testing dataset’s ROC to allow for the calculation of MCR, FNR, and FPR, with the AUC calculated from the testing ROC itself. For bootstrap experiments, the process was similar, except that the 250 repetitions allowed for the calculation of average ROC curves and error metrics with associated 95% CIs. The average AUC from bootstrap resampling under-estimates the true expected AUC, and so we separately reported the common AUC$_{0.632+}$ correction [31].

4.2.6 Model external validation

To address research questions 1–3, we conducted three types of validation. The first type, “model external validation”, answered research question 1 by training a model with one centre’s dataset (either A or B), and then testing it with the other centre’s dataset. As shown in figure 4.1a, this validation type is the most rigorous, with a locked model being tested at another centre.

4.2.7 Methodology external validation

To address research question 2, we performed “methodology external validation” by taking the locked methodology to train a model developed using Centre A’s dataset and applying it to Centre B’s dataset via bootstrap resampling to validate training a model per centre using
(a) Model External Validation

Train Dataset: A, Test Dataset: B

![Diagram of Model External Validation]

(b) Methodology External Validation

Train & Test Dataset: A

![Diagram of Methodology External Validation]

Train & Test Dataset: B

![Diagram of Methodology External Validation]

(c) Pooled Data Validation

Train & Test Dataset: A & B

![Diagram of Pooled Data Validation]

Figure 4.1: Visual representation of the three types of validation performed. (a) depicts model external validation, in which a model is trained using Centre A’s dataset, locked, and then tested using Centre B’s dataset. The inverse in which Centre A’s and B’s datasets are exchanged was also performed, as shown. (b) shows methodology external validation, in which a locked methodology is used to train and test models using Centre B’s dataset via bootstrap resampling, as shown under “Train & Test Dataset: B”. “Train & Test Dataset: A” shows the similar technique employed to develop the original model training methodology using only Centre A’s dataset that was then locked (indicated by the “*”) and then used and externally validated in the other experiments. “Train & Test Dataset: A” was also used in this study to revalidate the locked methodology with Centre A’s dataset. Pooled data validation in (c) is identical to (b), except that Centre A and B’s datasets are pooled before bootstrap resampling.
a methodology developed externally. The methodology was “locked” in the sense that all design decisions for how models are trained and have their hyperparameters optimized were held constant. Figure 4.1a,b compares this “methodology external validation” to the previous “model external validation”.

4.2.8 Pooled data validation

Lastly, to answer research question 3, we combined both centres’ datasets to perform “pooled data validation”. Figure 4.1c shows how bootstrapped resampling allows for data from Centre A and B to be represented in distinct training and testing datasets. We calculated overall and per-centre error metrics to evaluate if model accuracy was consistent for both centres’ testing samples.

4.2.9 Feature importance analysis

Based on each centre’s methodology external validation results, we performed feature importance analysis to answer research question 4. RDFs inherently supply feature importance scores, which were normalized between 0 and 1 for each bootstrap iteration. Radiomic features removed by the inter-feature correlation filter were unused by the RDFs, and so were assigned a score of 0. The scores were then averaged across bootstrap repetitions and re-normalized to be between 0 and 1.

4.2.10 Feature harmonization and consolidation

To explore the feature harmonization proposed in research question 5, we used “combating batch effects when combining batches” (ComBat) harmonization [32]. MRI acquisition parameters and scanner model are known to systematically affect radiomic features [29, 33]. While our MRI pre-processing can mitigate these effects, ComBat harmonization provides further corrections after feature extraction. ComBat finds multiplicative and additive terms
that represent the effect of data being collected at a given centre, and then their inverse is applied to minimize inter-centre differences [32, 34]. We therefore repeated the model external validation (figure 4.1a) using only radiomic features after applying ComBat.

We also hypothesized that radiomic feature variation could be mitigated through “consolidation”, by only including features that are important at both centres. The idea behind consolidation is that there may be features that are uniquely important at one centre, but also other features that are important across all centres. For our two centres, the previous feature importance analysis provided lists of radiomic features ranked by importance. From these lists, the top 10 commonly important features were consolidated using a ranking metric of the lowest importance score a feature received across the two centres. We then conducted model external validation experiments using only the top \( n \) consolidated radiomic features with ComBat harmonization also applied, with \( n \) ranging from 1 to 10.

### 4.3 Results

#### 4.3.1 Model external validation

The model trained on Centre A’s clinical features and then tested on Centre B produced an AUC of 0.59 and MCR of 50.4%. The ROC curve from this experiment is provided in figure 4.2a, with the error metrics provided in table 4.2 (first row of section A). The inverse scenario of training a model on Centre B and testing on Centre A also yielded low accuracy (ROC curve in figure 4.2b; error metrics in table 4.2, fourth row of section A). No CIs are associated with these model external validation results, as only a single model was tested per experiment.

Figure 4.2a,b and table 4.2 (section A) show the ROC curves and error metrics from the model external validation using only radiomic features and using all features. AUC values were near those from using clinical features only, except when using all features from Centre A to train a model, which had the highest AUC of 0.70. The ROC operating points
determined from the training datasets were inappropriate, leading to highly imbalanced FNRs and FPRs.
Figure 4.2: ROC curves and operating points for validation experiments. ROC curves for model external validation are provided when using Centre B’s dataset (a) or Centre A’s dataset (b) for testing. ROC curves when using only clinical features (“Clinical”), only radiomic features (“Radiomics”), or both clinical and radiomic features (“All”) are labelled, with each curve’s operating point marked by a “◦”. No CIs are provided in these cases, as a single testing dataset was used. (c) and (d) provide methodology external validation ROC curves for Centre A and B, respectively (shaded bands represent the 95% CIs). Pooled data validation ROC curves are shown in (e), with (f) showing the per-centre sub-analysis when using only radiomic features. For all ROCs, associated error metrics are quantified in table 4.2, with the right-most “ROC Figure” column of table 4.2 allowing for cross-referencing.

4.3.2 Methodology external validation

We revalidated the model training methodology developed with Centre A’s dataset in our previous study [17] due to the adjusted clinical feature categories and reduced number of clinical features from 12 to eight in Centre A’s dataset required to match Centre B. We found negligible differences compared to the previous study (figure 4.2c; table 4.2, section B), validating the methodology and adjusted clinical features, and providing a comparison baseline for Centre B.

Methodology external validation using Centre B’s dataset demonstrated slightly elevated accuracy compared to Centre A, except when using clinical features alone where a drop of 0.09 in $\text{AUC}_{0.632+}$ was seen (figure 4.2d; table 4.2, section B). Using only radiomic features or all features provided equivalent performance, with equal $\text{AUC}_{0.632+}$ values of 0.80 and overlapping MCR, FNR, and FPR 95% CIs.

4.3.3 Pooled data validation

The pooled data validation produced accuracy values between those of Centre A’s and Centre B’s methodology external validation results (figure 4.2e; table 4.2, section C, first three rows). Using only radiomic features and all features offered equivalent accuracy, matching Centre B’s methodology external validation results (compare against figure 4.2d). $\text{AUC}_{0.632+}$ values for all feature combinations fell approximately halfway between Centre
A’s and Centre B’s external methodology validation AUC_{0.632+} values.

Per-centre analysis revealed that a model trained on both centres’ datasets provided balanced performance for each centre (table 4.2, section C for all error metrics; figure 4.2f for radiomic features only ROCs). The AUC differences between the per-centre error metrics were small (0.02–0.04) with non-overlapping CIs.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Features</th>
<th>AUC</th>
<th>AUC_{0.632+}</th>
<th>MCR %</th>
<th>FNR %</th>
<th>FPR %</th>
<th>ROC</th>
<th>Figure</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: Model External Validation</td>
<td>Clinical</td>
<td>0.59</td>
<td>-</td>
<td>50.4</td>
<td>45.0</td>
<td>51.5</td>
<td>Figure 4.2a</td>
<td></td>
</tr>
<tr>
<td>A: Model External Validation</td>
<td>Radiomic</td>
<td>0.61</td>
<td>-</td>
<td>20.5</td>
<td>90.0</td>
<td>6.2</td>
<td>Figure 4.2b</td>
<td></td>
</tr>
<tr>
<td>A: Model External Validation</td>
<td>All</td>
<td>0.70</td>
<td>-</td>
<td>22.2</td>
<td>90.0</td>
<td>8.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B: Methodology External Validation</td>
<td>Clinical</td>
<td>0.67 (0.01)</td>
<td>0.73</td>
<td>35.3 (1.4)</td>
<td>46.3 (1.8)</td>
<td>32.1 (1.3)</td>
<td>Figure 4.2c</td>
<td></td>
</tr>
<tr>
<td>B: Methodology External Validation</td>
<td>Radiomic</td>
<td>0.68 (0.01)</td>
<td>0.74</td>
<td>32.4 (1.5)</td>
<td>41.4 (2.1)</td>
<td>29.9 (1.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B: Methodology External Validation</td>
<td>All</td>
<td>0.69 (0.01)</td>
<td>0.76</td>
<td>30.5 (1.5)</td>
<td>39.9 (2.0)</td>
<td>27.8 (1.3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C: Pooled Data Validation</td>
<td>Clinical</td>
<td>0.63 (0.01)</td>
<td>0.67</td>
<td>36.1 (1.1)</td>
<td>53.3 (1.5)</td>
<td>31.9 (1.0)</td>
<td>Figure 4.2e</td>
<td></td>
</tr>
<tr>
<td>C: Pooled Data Validation</td>
<td>Radiomic</td>
<td>0.71 (0.01)</td>
<td>0.78</td>
<td>32.4 (1.1)</td>
<td>38.7 (1.7)</td>
<td>30.9 (1.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C: Pooled Data Validation</td>
<td>All</td>
<td>0.71 (0.01)</td>
<td>0.78</td>
<td>31.3 (1.2)</td>
<td>39.7 (1.8)</td>
<td>29.3 (1.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D: Feature Harmonization</td>
<td>Clinical</td>
<td>0.62 (0.01)</td>
<td>-</td>
<td>37.2 (1.5)</td>
<td>50.3 (1.9)</td>
<td>34.0 (1.3)</td>
<td>Figure C.2a</td>
<td></td>
</tr>
<tr>
<td>D: Feature Harmonization</td>
<td>Radiomic</td>
<td>0.69 (0.01)</td>
<td>-</td>
<td>33.6 (1.4)</td>
<td>38.3 (2.0)</td>
<td>32.5 (1.3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D: Feature Harmonization</td>
<td>All</td>
<td>0.70 (0.01)</td>
<td>-</td>
<td>32.7 (1.4)</td>
<td>39.1 (2.0)</td>
<td>31.2 (1.3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D: Feature Harmonization</td>
<td>Clinical</td>
<td>0.59 (0.02)</td>
<td>-</td>
<td>33.8 (1.5)</td>
<td>61.3 (2.4)</td>
<td>27.1 (1.3)</td>
<td>Figure C.2f</td>
<td></td>
</tr>
<tr>
<td>D: Feature Harmonization</td>
<td>Radiomic</td>
<td>0.73 (0.01)</td>
<td>-</td>
<td>29.2 (1.5)</td>
<td>43.8 (2.5)</td>
<td>25.6 (1.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D: Feature Harmonization</td>
<td>All</td>
<td>0.72 (0.01)</td>
<td>-</td>
<td>30.6 (1.5)</td>
<td>41.0 (2.5)</td>
<td>28.0 (1.3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E: Feature Harmonization &amp; Consolidation</td>
<td>ComBat Radiomic</td>
<td>0.63</td>
<td>-</td>
<td>37.6</td>
<td>50.0</td>
<td>35.1</td>
<td>Figure C.2c</td>
<td></td>
</tr>
<tr>
<td>E: Feature Harmonization &amp; Consolidation</td>
<td>Radiomic</td>
<td>0.61</td>
<td>-</td>
<td>43.9</td>
<td>44.4</td>
<td>43.8</td>
<td>Figure C.2d</td>
<td></td>
</tr>
<tr>
<td>E: Feature Harmonization &amp; Consolidation</td>
<td>n = 1 ComBat Radiomic</td>
<td>0.66</td>
<td>-</td>
<td>42.7</td>
<td>35.0</td>
<td>44.3</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>E: Feature Harmonization &amp; Consolidation</td>
<td>n = 2 ComBat Radiomic</td>
<td>0.63</td>
<td>-</td>
<td>48.6</td>
<td>18.5</td>
<td>58.3</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>E: Feature Harmonization &amp; Consolidation</td>
<td>n = 3 ComBat Radiomic</td>
<td>0.61</td>
<td>-</td>
<td>48.7</td>
<td>30.0</td>
<td>52.6</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>E: Feature Harmonization &amp; Consolidation</td>
<td>n = 4 ComBat Radiomic</td>
<td>0.59</td>
<td>-</td>
<td>36.6</td>
<td>63.0</td>
<td>29.2</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>E: Feature Harmonization &amp; Consolidation</td>
<td>n = 5 ComBat Radiomic</td>
<td>0.70</td>
<td>-</td>
<td>33.3</td>
<td>40.0</td>
<td>32.0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>E: Feature Harmonization &amp; Consolidation</td>
<td>n = 6 ComBat Radiomic</td>
<td>0.65</td>
<td>-</td>
<td>30.1</td>
<td>63.0</td>
<td>20.8</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>E: Feature Harmonization &amp; Consolidation</td>
<td>n = 7 ComBat Radiomic</td>
<td>0.67</td>
<td>-</td>
<td>45.3</td>
<td>40.0</td>
<td>46.4</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>E: Feature Harmonization &amp; Consolidation</td>
<td>n = 8 ComBat Radiomic</td>
<td>0.70</td>
<td>-</td>
<td>26.0</td>
<td>33.3</td>
<td>24.0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>E: Feature Harmonization &amp; Consolidation</td>
<td>n = 9 ComBat Radiomic</td>
<td>0.68</td>
<td>-</td>
<td>45.3</td>
<td>35.0</td>
<td>47.4</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>E: Feature Harmonization &amp; Consolidation</td>
<td>n = 10 ComBat Radiomic</td>
<td>0.65</td>
<td>-</td>
<td>36.6</td>
<td>37.0</td>
<td>36.5</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>E: Feature Harmonization &amp; Consolidation</td>
<td>n = 11 ComBat Radiomic</td>
<td>0.46</td>
<td>-</td>
<td>66.7</td>
<td>35.0</td>
<td>73.2</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>E: Feature Harmonization &amp; Consolidation</td>
<td>n = 12 ComBat Radiomic</td>
<td>0.61</td>
<td>-</td>
<td>46.3</td>
<td>37.0</td>
<td>49.0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>E: Feature Harmonization &amp; Consolidation</td>
<td>n = 13 ComBat Radiomic</td>
<td>0.60</td>
<td>-</td>
<td>35.0</td>
<td>50.0</td>
<td>32.0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>E: Feature Harmonization &amp; Consolidation</td>
<td>n = 14 ComBat Radiomic</td>
<td>0.62</td>
<td>-</td>
<td>37.4</td>
<td>44.4</td>
<td>35.4</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.2: Error metrics values across the validation experiments performed. Notes: The rows of the table are divided into five sections (A–E), corresponding to the three types of validation performed, along with two sections regarding the feature harmonization and consolidation results. For sections A–C, row highlighting is based on the available features to allow for easier comparisons between experiments using clinical and/or radiomic features. The row highlighting in sections D and E is based on whether Centre A or B’s dataset was used for testing. For section E, the “Features” column value refers to using the top 1 to n consolidated radiomic features with ComBat harmonization also applied. Rows marked with “□” represent model external validation experiments for comparison with Centre B as the test dataset using only radiomic features with no or different feature harmonization and consolidation techniques applied. Rows marked with “■” are represent a similar comparison, but with Centre A as the test dataset. The “ROC Figure” column provides a means to cross-reference error metric values with the ROC curves from which they were derived (see supplementary figure C.2). Where bootstrapped experiments were performed, 95% CI values are supplied in parentheses, along with AUC\textsubscript{0.632+} values.

4.3.4 Feature importance analysis

Methodology external validation radiomic feature importance analysis found commonalities and differences between Centre A and B. Features with an importance score of 0.75 or greater were deemed “highly important”, with 14 such features revealed for Centre A and five for Centre B, as shown in figure 4.3’s left and right-most tables (features above dashed line). At both centres, primarily higher-order texture features were important, and in particular, GLCM features (33.3% of Centre A’s important features; 40.0% of Centre B’s). First-order statistical features were in the minority at both centres, and only Centre A used a highly important shape and size-based feature. Only two features, GLSZM zone entropy and GLCM inverse difference normalized, were highly important at both centres. Importance scores for all radiomic features are provided in the supplementary data (see section 4.6).

4.3.5 Feature harmonization and consolidation

ComBat feature harmonization was found to have a minimal impact on model performance. ComBat harmonization on the radiomic features caused model external validation AUC
<table>
<thead>
<tr>
<th>Rank</th>
<th>Score</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>GLRLM - Gray-Level Non-Uniformity Normalized</td>
</tr>
<tr>
<td>2</td>
<td>0.94</td>
<td>NGTDM - Contrast</td>
</tr>
<tr>
<td>3</td>
<td>0.93</td>
<td>GLSZM - Zone Entropy</td>
</tr>
<tr>
<td>4</td>
<td>0.92</td>
<td>GLCM - Inverse Difference Normalized</td>
</tr>
<tr>
<td>5</td>
<td>0.91</td>
<td>GLCM - Inverse Variance</td>
</tr>
<tr>
<td>6</td>
<td>0.88</td>
<td>GLSZM - Zone Percentage</td>
</tr>
<tr>
<td>7</td>
<td>0.85</td>
<td>GLDM - Dependence Entropy</td>
</tr>
<tr>
<td>8</td>
<td>0.83</td>
<td>Shape &amp; Size - Surface Volume Ratio</td>
</tr>
<tr>
<td>9</td>
<td>0.82</td>
<td>GLRLM - Gray-Level Non-Uniformity Normalized</td>
</tr>
<tr>
<td>10</td>
<td>0.81</td>
<td>GLCM - Correlation</td>
</tr>
<tr>
<td>11</td>
<td>0.81</td>
<td>First-Order - Kurtosis</td>
</tr>
<tr>
<td>12</td>
<td>0.82</td>
<td>First-Order - 10th Percentile</td>
</tr>
<tr>
<td>13</td>
<td>0.79</td>
<td>GLCM - Informational Measure of Correlation 2</td>
</tr>
<tr>
<td>14</td>
<td>0.75</td>
<td>GLCM - Cluster Shade</td>
</tr>
<tr>
<td>15</td>
<td>0.72</td>
<td>NGTDM - Busyness</td>
</tr>
<tr>
<td>16</td>
<td>0.72</td>
<td>GLSZM - Large Area Low Gray-Level Emphasis</td>
</tr>
<tr>
<td>17</td>
<td>0.70</td>
<td>Shape &amp; Size - Flatness</td>
</tr>
<tr>
<td>18</td>
<td>0.68</td>
<td>GLCM - Informational Measure of Correlation 1</td>
</tr>
<tr>
<td>19</td>
<td>0.66</td>
<td>GLCM - Contrast</td>
</tr>
<tr>
<td>20</td>
<td>0.64</td>
<td>GLSZM - Size Zone Non-Uniformity</td>
</tr>
<tr>
<td>21</td>
<td>0.64</td>
<td>GLSZM - Gray-Level Non-Uniformity Normalized</td>
</tr>
<tr>
<td>22</td>
<td>0.63</td>
<td>NGTDM - Strength</td>
</tr>
<tr>
<td>23</td>
<td>0.63</td>
<td>GLSZM - Size Zone Non-Uniformity Normalized</td>
</tr>
<tr>
<td>24</td>
<td>0.62</td>
<td>GLDM - Dependence Non-Uniformity Normalized</td>
</tr>
<tr>
<td>25</td>
<td>0.61</td>
<td>GLRLM - Run Entropy</td>
</tr>
<tr>
<td>26</td>
<td>0.61</td>
<td>First-Order - 90th Percentile</td>
</tr>
<tr>
<td>27</td>
<td>0.61</td>
<td>GLDM - Small Dependence Low Gray-Level Emphasis</td>
</tr>
<tr>
<td>28</td>
<td>0.61</td>
<td>GLCM - Cluster Prominence</td>
</tr>
</tbody>
</table>

**Centre B Features**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Score</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>First-Order - Range</td>
</tr>
<tr>
<td>2</td>
<td>0.92</td>
<td>GLSZM - Zone Entropy</td>
</tr>
<tr>
<td>3</td>
<td>0.88</td>
<td>GLCM - Cluster Prominence</td>
</tr>
<tr>
<td>4</td>
<td>0.76</td>
<td>GLCM - Inverse Difference Normalized</td>
</tr>
<tr>
<td>5</td>
<td>0.75</td>
<td>NGTDM - Strength</td>
</tr>
<tr>
<td>6</td>
<td>0.70</td>
<td>Shape &amp; Size - Surface Volume Ratio</td>
</tr>
<tr>
<td>7</td>
<td>0.69</td>
<td>GLCM - Correlation</td>
</tr>
<tr>
<td>8</td>
<td>0.68</td>
<td>GLSZM - Small Area High Gray-Level Emphasis</td>
</tr>
<tr>
<td>9</td>
<td>0.63</td>
<td>Shape &amp; Size - Sphericity</td>
</tr>
<tr>
<td>10</td>
<td>0.63</td>
<td>GLSZM - Size Zone Non-Uniformity</td>
</tr>
<tr>
<td>11</td>
<td>0.62</td>
<td>GLSZM - Large Area High Gray-Level Emphasis</td>
</tr>
<tr>
<td>12</td>
<td>0.61</td>
<td>GLSZM - Gray-Level Non-Uniformity Normalized</td>
</tr>
<tr>
<td>13</td>
<td>0.60</td>
<td>GLRLM - Run Entropy</td>
</tr>
<tr>
<td>14</td>
<td>0.59</td>
<td>GLDM - Dependence Non-Uniformity Normalized</td>
</tr>
</tbody>
</table>

**Figure 4.3**: Per-centre radiomic feature importance analysis and associated feature consolidation. The left and right-most tables show the most important radiomic features for Centres A and B, respectively, as computed using the methodology external validation results. The central table shows the ranking of the top 10 consolidated features from Centres A and B, with a feature’s ranking determined by taking the minimum of the importance score it received at each centre. Only the features down to the least important feature required to create the top 10 consolidated features are shown, with an exhaustive feature list provided in the supplementary data (see section 4.6). The dashed lines represent the 0.75 cut-off for “highly important” features.

Values to remain unchanged or rise slightly by 0.02 (see indicated comparison between values in table 4.2, sections A, D). The chosen ROC operating points chosen were more appropriate, however (table 4.2, section D). Feature harmonization was pursued only with
radiomic features, as the univariate predictive value of clinical features differed greatly between the two centres (see table 4.1), and clinical features added no benefit in Centre B’s methodology external validation.

A consolidated set of the top 10 mutually important features was then established, as shown in the middle table of figure 4.3. These consolidated features consisted almost entirely of higher-order texture features, with one shape and size feature included. GLCM and GLSZM features made up the majority of the set, including four of the top five features. To construct the feature set, features with importance scores down to near 0.60 were required, as shown through the linkages between tables in figure 4.3.

When using the top \( n = 4 \) of the consolidated features, the highest average AUC across testing on Centre A or B was achieved (figure 4.4a; further error metrics in table 4.2, section E). The AUC values from each centre were within 0.03, while the associated MCR, FPR, and FNR were quite different. Figure 4.4b demonstrates this is due to ROC operating point
selection, with the ROC curves demonstrating multiple points of equivalent FPR and FNR across centres.

### 4.4 Discussion

Our study demonstrates the generalizability across two independent centres of post-SRS progression predictive models based on clinical and MRI radiomic features, as well as the methodology to produce these models. Methodology external validation showed maximal accuracy when models were retrained per-centre using a locked methodology developed externally at another centre. The achieved \( \text{AUC}_{0.632+} = 0.80 \) is comparable to previous results, with most studies reporting AUCs in the range of 0.73–0.87 [12, 14–18]. Model external validation with feature harmonization and consolidation showed inferior accuracy with a maximum \( \text{AUC} = 0.70 \) achieved. Pooled data validation displayed a promising \( \text{AUC}_{0.632+} = 0.78 \), with balanced per-centre performance.

#### 4.4.1 Model external validation

The poor model external validation results before feature harmonization and consolidation indicate that critical differences exist between the two centres’ study samples. The clinical features predicting SRS response varied significantly between centres, with only GTV volume being a common predictor, as has been reported in other studies [35, 36]. All clinical features except sex were also significantly different between the two study samples, demonstrating that the samples are not comparable clinically. This is likely due to differences in year of treatment, clinical practice, and population demographics. These differences could also be caused by small sample sizes, as some less common primary cancer sites present at Centre A were not present at Centre B, whereas larger sample sizes could be more balanced. Further validation of clinical feature differences is required, especially considering the 5–17 year difference in SRS treatment dates between centres A and B. Other studies in
the field have all been conducted using modern study samples that are therefore likely to be more similar to Centre’s B dataset [12–16, 18–24]. Performing model external validation with Centre B’s dataset and an external centre dataset more representative of modern practice than Centre A’s, could potentially show enhanced accuracy when models are transferred between centres.

4.4.2 Methodology external validation

The ability of Centre A’s methodology to create an accurate model for Centre B is an important step towards clinical translation, showing that a centre could train a predictive model using a now validated methodology. Given that clinical features across centres vary significantly and offered no benefit to Centre B’s results, the use of only radiomic features is recommended for clinical implementation. This also eliminates the need for resource-intensive collection of clinical data. Furthermore, it is highly encouraging that methodology external validation was successful despite large inter-centre differences.

4.4.3 Pooled data validation

Pooled data model overall AUC values were always directly halfway between the AUCs for the models training for each centre. The pooled model also provided AUCs within 0.04 when comparing accuracy between testing samples from Centre A and Centre B. This shows that multi-centre models generalize across the centres on which they were trained, but a third centre’s data is required to investigate the generalizability of these models.

4.4.4 Feature importance analysis

Feature importance analysis revealed some features are predictive of SRS outcomes and resistant to inter-centre variability. The number of highly important features also differed between the two centres (14 for Centre A; five for Centre B), which is likely due to Centre
A’s more diverse set of MR scanners compared to Centre B.

Limited overlap in important radiomic features exists between our study’s centres and previous studies. Centre A had a total of four highly important features overlap with other studies (kurtosis with two studies [15, 16], and GLCM cluster shade and GLCM information measure correlation 2 both with one study [12]). Centre B had two highly important features, range and NGTDM strength, which overlap with another study [16]. Across previous SRS studies, there are only examples of studies having a single important radiomic feature in common with another study [12–16, 18]. MR scanner, utilized MR sequences and ROIs, clinical feature diversity, image pre-processing, and ML methodology variability likely all contribute to this finding, as our study controlled some of these variables and found greater radiomic feature overlap between centres.

### 4.4.5 Feature harmonization and consolidation

ComBat harmonization did not improve model performance, and so it is not a viable solution for correcting inter-centre variability, with other studies also demonstrating mixed effects [37–39]. Before harmonization, nearly 80% of radiomic features were significantly different between centres, and so ComBat clearly brought these features into closer alignment, as demonstrated through more appropriate ROC operating point selection. ComBat’s utility despite performing MRI pre-processing shows that MRI differences were not fully corrected for by pre-processing. Our previous sensitivity study demonstrated that MR scanner variability within a single centre negatively impacts model accuracy despite MRI pre-processing including Z-score intensity normalization [17]. Most previous studies investigating predicting post-SRS/SRT progression using radiomics have a single MR scanner represented in their datasets [12–14, 18, 23, 24], while the remaining studies have two scanners represented, with the effect of scanner variability not examined [15, 16, 19, 22]. MRI pre-processing and intensity normalization techniques vary widely across these previous studies as well, motivating the robust comparison of currently employed MRI nor-
malization techniques and development of enhanced techniques to possibly improve model external validation accuracy.

The per-centre feature importance analysis shows that ComBat failed because despite features being better aligned, the important radiomic features at each centre were not necessarily important at both centres. Therefore, it was hypothesized that using a consolidated set of mutually important features was required, which did offer enhanced model external validation performance. Optimal performance occurring at \( n = 4 \) consolidated features likely indicates that using too few features does not provide enough degrees of freedom for accurate prediction, while too many features enables overfitting to a given centre. This demonstrates a possible method for creating more generalizable models without retraining per centre, but external validation is needed, as the consolidated feature set was constructed using information from both centres.

4.4.6 Limitations and future work

It is important to interpret the results of this study in the context of its limitations. First, both centres’ datasets were retrospective, leading to implicit inclusion and exclusion criteria being applied based on a centre’s local population demographics and SRS treatment selection process. Since one follow-up MRI was required for study inclusion, patients who passed away shortly after SRS and those that declined or were not well enough for MRI were also systematically removed. Second, both centres’ datasets only included patients treated with linac-based SRS. No immediate conclusions can then be drawn about whether the methodologies, models, or results of this study would transfer to SRS-specific modalities (e.g. Gamma Knife), and so future inclusion of datasets with patients treated with SRS-specific modalities is required. Third, a limited set of clinical features and MRI sequences was included to allow for both centres’ datasets to contain equivalent variables. Other clinical and MRI features have had predictive value in other studies [14–16, 18], motivating further external validation with datasets of greater feature diversity. Fourth, interpretability of
radiomic systems remains a barrier to clinical adoption. While our finding of consistently predictive features across two centres builds confidence in radiomic methods, discovering relationships between features and underlying biological mechanisms must be investigated.

The definition of our BM progression post-SRS endpoint is another limitation, as our method was non-standard. This was required to include Centre A’s dataset and also is a more robust volumetric method compared to the standard RANO-BM approach [26], but it does limit comparison to other studies. Our successful external validation reinforces our method’s validity, though future experiments should report both measures. The differentiation between true and pseudo-progression is also notoriously difficult [40]. Our positive results suggest reasonable control of pseudo-progression, but prospective trials would allow for more reliable pseudo-progression determination, which may enhance model performance due to more accurate training datasets.

4.5 Conclusions

We have demonstrated dual-centre generalizability of models and modeling methodologies for predicting SRS outcomes using clinical and MRI radiomic features. We revealed that without radiomic feature harmonization and consolidation, a model trained at one centre performs poorly at another. Next, we showed our model training methodology is generalizable by producing accurate models across both centres. We also showed that a pooled data model offers strong performance overall and balanced performance across the centres on which it was trained. The results of this external validation study provide crucial motivation for continued research to ultimately improve clinical outcomes in the treatment of BMs.
4.6 Data and code availability

The ground truth label and predicted progression probability data from each reported experiment needed to replicate this study’s analysis are available at the following URL: https://github.com/baines-imaging-research-laboratory/radiomics-for-srs-external-validation-data-share

The computer code used to perform the reported ML experiments and subsequent analysis are available at the following URL: https://github.com/baines-imaging-research-laboratory/radiomics-for-srs-external-validation-code

Supplementary data files are available at the following URL: https://github.com/baines-imaging-research-laboratory/radiomics-for-srs-external-validation-supplementary

4.7 References


Chapter 5

Conclusions and future work

The research presented in this thesis aimed to address the gaps in knowledge and unmet needs in the field of predicting BM response to SRS using MRI radiomics-based ML models identified in section 1.4.6. Before the presented work, the sensitivity of radiomics-based ML models to clinical factors such as primary cancer site, BM volume, and MR scanner were unknown, as well as if the results from radiomics-based ML models would extend to linac-based SRS. Using radiomics-based ML to predictive SRS outcomes had also not been compared to previous predictive models using qualitative BM appearance, which is critical to motivate the use of ML approaches. Robust interpretation of radiomics-based models had also not been performed, which when successful, can aid in motivating the clinical translation of ML approaches. Lastly, external validation of ML models across multiple centres had not been performed until first reported March 2023, and this first external validation was limited in terms of the external dataset’s size and MR scanner variability, as well as the external validation techniques applied [1]. To address these gaps in knowledge and unmet needs, six research objectives were proposed in section 1.5 and three research projects were performed to meet these objectives (chapters 2, 3, and 4), producing the following advances in knowledge and technology.
5.1 Advances in knowledge and technology

5.1.1 Characterization of machine learning model sensitivity to clinical factors

The analysis in chapter 2 revealed that radiomics-based ML model accuracy is dependent on multiple factors that are clinically relevant including primary cancer site and BM volume. Primary cancer site and BM volume have been important factors to consider in other BM prognostic and predictive models [2, 3], and this indicates the possibility of having to create models specific to these factors. This may be especially true as treatment guidelines become specific to primary cancer type with respect to immunotherapy usage, and the choice of using SRS, SRT, or surgical resection is guided by BM size [4–6]. The difference in BM volumes being treated with SRS across the two centres in chapter 4 also revealed that smaller BMs are being treated with SRS in recent years, likely due to enhanced primary cancer screening and MRI surveillance for BMs. SRS is typically highly effective against very small BMs (< 2 cm in diameter), with failure rates as low as 5% [3, 7, 8]. Therefore specific models for different BM sizes, or possibly only for larger BMs, may be necessary.

Chapter 2 also indicated a small, but statistically significant, advantage to using clinical and radiomic features to make predictions. In contrast, chapter 3 indicated that using only SRS prescription and radiomic features was optimal for risk stratification, while chapter 4 showed that clinical features offered no benefit during external validation at a second centre. This observed mixed benefit of clinical features, combined with similar mixed observations across previous studies, demonstrates a large source of variability between centres. The causal connection between the prescribed SRS dose and fractionation and TCP has been well-described and modeled [3, 9, 10] and therefore should remain in future models, as was shown in chapter 3’s optimal risk stratification model. The inclusion of other clinical features may have limited or no benefit, while also increasing the complexity
of data collection and comparing study results, and therefore these clinical features should be included cautiously moving forward.

5.1.2 Method to remove predictive model brain metastasis volume bias developed for and applied to stereotactic radiosurgery

After characterizing the effect of BM volume on ML model accuracy in chapter 2, a method to evaluate the effect of gradually “blinding” a radiomics-based ML model to BM volume was developed and applied to SRS. This method reduced the ML model’s bias to misclassify large BMs, while overall performance was minimally affected. While this method was beneficial in chapter 2, the second, more modern dataset used for external validation in chapter 4 would derive no such benefit given that all of the dataset’s BMs fell under the 7.5 cc threshold used to define large BMs in chapter 2. Modern guidelines suggest using SRS up to BMs of approximately 2 cm in diameter (approximately 4.2 cc for an assumed spherical BM), above which SRT is recommended [6]. Therefore the BM volume bias removal technique demonstrated in chapter 2 may have limited SRS applications, but may be beneficial if models are eventually developed for pooled SRS and SRT data that would then include a greater range of BM volumes.

5.1.3 Comparison of clinician-based qualitative analysis versus radiomics-based quantitative analysis for outcome prediction

In chapter 3, the first comparison was presented between SRS outcome predictive models that used qualitative or quantitative approaches to T1w-CE MRI analysis. While previous work had compared radiomics-based ML model to a single qualitative appearance score [11, 12], chapter 3’s comparison was against a predictive model developed using multiple predictive features and a more comprehensive qualitative appearance scoring method. The results demonstrated that using qualitative appearance to predict SRS outcome is not only
less capable of stratifying BMs at risk for post-SRS progression compared to ML models, but high interobserver variability in appearance scoring makes the technique infeasible for clinical use. This advance in knowledge firmly demonstrates that radiomics-based ML models are currently the most capable and feasible models developed for predicting BM response to SRS, and therefore should be strongly considered for clinical translation.

5.1.4 Method to interpret radiomics-based machine learning models using brain metastasis qualitative appearance developed for and applied to stereotactic radiosurgery

The interpretation of radiomics-based ML model operation in chapter 3 indicated that radiomic models predicting post-SRS progression independently identified BMs of heterogeneous or necrotic appearance as being at the highest risk for progression. Furthermore, the change in the radiomic models’ predicted probability of progression with respect to different radiomic feature values was highly correlated with necrosis as well. The development of these methods to use the five-way BMs qualitative appearance scoring provided novel interpretation of radiomic ML models that also agrees with current radiobiological understanding. This advancement in model interpretability provides a critical insight into radiomic ML models for SRS outcome predictions that is complimentary to the strong predictive accuracy of ML models reported in this work and the broader literature. It is intended that these model interpretability results will motivate further research in radiomics-based outcome prediction and eventual clinical translation.
5.1.5 External validation of extending radiomics-based machine learning outcome prediction to linear accelerator stereotactic radio-surgery

In chapter 2, radiomics-based ML techniques were demonstrated to extend to linac-based SRS. Given that linac-based SRS offers smaller centres without specialized equipment the ability to run SRS treatment programs and provides all centres the ability to concurrently treat multiple BMs, this advance in knowledge greatly widens the number of centres and patients for which ML predictive models have been specifically investigated. The results in chapter 2 were also externally validated in chapter 4 for a second centre also using linac-based SRS. By showing repeatable model performance across two centres with highly diverse patient cohorts, this research lays a foundation for further centres offering linac-based SRS to develop, evaluate, and deploy radiomics-based ML for treatment decision-making.

5.1.6 External validation of methodology to produce per-centre machine learning models

A methodology for training radiomics-based ML models to predict SRS outcomes was first developed in chapter 2. While the results of chapter 2’s analysis showed a capable predictive model, it was unknown if similar results were possible if the same model training methodology was used to produce a model for an external centre. This is especially important to consider as chapter 4 revealed that a model trained at one centre does not achieve high accuracy when tested at another centre, suggesting that training models per centre may be necessary. External validation of the model training methodology in chapter 4 demonstrated that using the locked methodology from chapter 2 allowed for training a predictive model at an external centre with similar error metrics performance. If the transfer of
models between centres without retraining remains unrealized, the model training methodology presented in this work will be critical for the widespread adoption and evaluation of radiomics-based ML models across multiple centres.

5.1.7 Characterization of the effect of magnetic resonance scanner model variability on radiomics-based machine learning model performance

The two datasets used in this work’s research each contained a greater variety of MR scanner models compared to previous studies. This unique aspect of the datasets allowed for examination of the effect MR scanner model variability has on radiomic-based ML models. Chapter 2 found that intra-centre scanner model variability had a significant impact on model performance, despite the MRI normalization pre-processing performed. Chapter 4 found inter-centre scanner model variability may negatively impact the performance of models transferred between centres without retraining. This may explain the difference in the poor external validation results reported in chapter 4 when models were not retrained per centre and the stronger external validation results reported in Du et al.’s external validation [1], as Du et al.’s external validation featured the same MR scanner model at both centres examined.

5.2 Suggestions for future work

The advances in knowledge and technology achieved in this thesis lay a foundation for future work in the field of radiomics-based prediction of SRS outcomes. Given the generally consistent ML model performance reported across multiple studies, future work is suggested to be firmly directed towards clinical applications and translation.
5.2.1 Further external validation studies and prospective data collection

The dual-centre validation conducted in chapter 4 provided only an initial step in external validation of radiomics-based ML models predicting SRS outcomes. First, external validation of models and model training methodologies should be conducted across datasets with greater similarity in treatment era, as this may improve external validation results due to closer alignment of treatment techniques, imaging technologies, typical length of patient follow-up, and the size of BMs treated. Second, external validation across datasets treated with different modalities (Gamma Knife, CyberKnife, or linac) would ensure that ML models and methodologies developed with one modality are transferable within the field of SRS/SRT outcome prediction more generally. Third, a wide variety of data from MRI sequences, SRS dose maps, and other clinical features have been explored in single centres with often positive results (see table 1.6), and so external validation of these more diverse feature sets is recommended. Fourth, performing external validation with more than two centres would allow for a model to be trained with pooled data from multiple centres and still be tested with an external dataset. Performing this experiment would allow for validation of the generalizability of models created using pooled data that is likely more representative of the general SRS/SRT patient population.

The suggested future external validation studies, as well as all other research in the field, would benefit from prospective data collection. All current studies, and the work presented in this thesis, have featured retrospective datasets. While the rapid data collection afforded by retrospective analysis has enabled the recent growth of research in the area of ML SRS outcome prediction, it comes at the cost of data quality and patient selection bias. Prospective data collection would first allow for more comprehensive and complete collection of clinical data points. This may be especially valuable as new systemic therapies are used to specifically treat BMs, possibly confounding the measurement of SRS effects [4, 5].
New data from histopathology and genetic sequencing may also become more widely available, and should be considered for predictive use [13–15]. Most critically, prospective data collection would allow for far greater consistency in endpoint data collection for both post-SRS progression and toxicity. The RANO-BM protocol provides a measurement method for standardized post-SRS progression data collection in retrospective datasets [16], but prospective data collection would allow for greater control of the frequency and protocol of radiology follow-ups. The confounding effects of pseudo-progression, ARE, and RN (see figure 1.7) are a common weakness for all studies examining BM outcomes, and so prospective data collection would possibly allow for enhanced screening for these outcomes using advanced imaging or biopsy [17, 18].

5.2.2 Development and evaluation of enhanced magnetic resonance imaging normalization techniques

MR scanner model variability and the inability of Z-score MRI intensity normalization pre-processing to completely account for this variability were identified in chapters 2 and 4 as a cause for poor intra and inter-centre model performance. Alternative MRI intensity normalization techniques do exist in the literature [19–21] and should be systematically evaluated to optimize model accuracy and generalizability. Furthermore, the application of MRI analysis for SRS outcome prediction could be used as motivation for the development of enhanced intensity normalization techniques, if required. NNs could also be applied for MRI normalization. CNNs could be used to avoid radiomic feature extraction and their sensitivity to MRI variability altogether by directly predicting SRS outcomes by taking raw MRI data as input, as has been first demonstrated by Jalalifar et al. [22]. NNs could also be trained to generate normalized MRI data for use in traditional radiomic feature extraction [23]. To maximize the likelihood of success of these efforts, the assembly of as diverse of a dataset as possible will be necessary, likely by pooling data across centres with various MR scanner models. Open-source MRI datasets of BM or other brain tumours
without associated treatment outcome data could also be leveraged for this task [24, 25].

5.2.3 Interpretation of radiomic signatures through advanced imaging and pathology

The model interpretation performed in chapter 3 provided novel insights into the operation of radiomics ML models using T1wCE MRI, but as identified in section 3.4.4, qualitative appearance labels do not offer a conclusive representation of the microbiological environment within a BM. Therefore, the field would benefit from future interpretation analysis based on more reliable biological data. Advanced imaging modalities such as PET or magnetic resonance spectroscopy (MRS) would offer more precise identification and quantification of relevant biological data such as hypoxia or cancer metabolism [18, 26]. PET and MRS are not typically performed before SRS/SRT, and so prospective data collection would be required. In the shorter term, similar qualitative appearance analysis techniques as used in this work may be beneficial for extending model interpretation beyond T1wCE MRI, such as to the appearance of edema in T2w imaging [27].

Pathology tissue samples would also provide more precise biological data from BMs, though tissue samples are also not commonly acquired before SRS/SRT. BMs treated with surgical resection may provide a solution, as MRI is consistently acquired before resection. The resected BM’s tissue would be available for pathology analysis (either through prospective data collection or possibly through tissue banks), allowing for correlations between the MRI and pathology data to be discovered, as has been explored in other fields [28, 29].
Figure 5.1: Schematic of using pre-treatment patient and BM data with radiomics-based ML models to choose an optimal SRS/SRT prescription. Tumour control and toxicity ML models are both trained, and then patient and BM specific features are provided to each, along with multiple options for the SRS/SRT prescription. For each SRS/SRT prescription option provided to the ML models, a probability of tumour control and a probability of treatment toxicity are predicted. The predicted probabilities from each prescription option can then be assembled to produce estimates of patient/BM-specific TCP and NTCP functions. An optimal SRS/SRT prescription can then be chosen (shown graphically as the highlighted region), and potential tradeoffs between tumour control and toxicity examined (e.g. in the example shown, if the next less aggressive SRS/SRT prescription was chosen, the TCP is predicted to drop slightly, while the risk of toxicity decrease substantially).

5.2.4 Combination and application of predictive models for stereotactic radiosurgery tumour control and normal tissue toxicity

To ultimately realize the clinical benefits of using ML predictive models for SRS/SRT patients, the BM response models examined in this work will have to be combined with ML models predicting SRS/SRT toxicity. As discussed previously, current decisions around the use and prescription of SRS/SRT are guided by the HyTEC guidelines’ underlying generalized TCP and NTCP models used for all patients and BMs [3, 30] (see section 1.4.2). ML models predicting post-SRS progression for BMs are effectively producing TCP values as well, but the TCP values are individualized to the patient and BM. As shown in figure 5.1, a TCP function individualized to a specific patient and BM can be produced by providing pre-treatment radiomic and clinical features to an ML model predicting BM response and then consecutively changing the SRS prescription input to the model over a range of values. The same process could the be repeated for a ML model predicting
SRS/SRT toxicities, such as RN, resulting in a NTCP curve that is also specific to the patient and BM (figure 5.1). Equipped with these individualized TCP and NTCP curves, along with a OS prognosis from a model such as the DS-GPA, clinicians and patients could make more informed decisions around trade-offs between risk of BM progression post-SRS and toxicity.

To realize this vision of a complete predictive modeling system for clinical use, a unique dataset would be required. First, this dataset would require dual-endpoint collection of BM post-SRS progression and toxicity. The same prospective techniques described previously in section 5.2.1 to improve post-SRS progression scoring accuracy would inherently provide this toxicity scoring. Second, this dataset would require a cohort of patients treated with a wide range of SRS prescriptions. Without a large prescription range, the ML models would be forced to extrapolate when producing predictions of BM response and toxicity for prescriptions at very low or high doses. To acquire such a dataset, pooling of data across centres would be largely beneficial, as some centres are generally more conservative with SRS/SRT prescriptions to avoid toxicity, while other centres are willing to prescribe higher doses to maximize tumour control. Data from clinical trials examining dose escalation or deescalation could also aid in extending the range of represented prescriptions [31, 32].

A ML study using BM response and toxicity models to estimate individualized TCP and NTCP estimates would allow for comparison with existing TCP and NTCP models, and if successful, an eventual clinical trial. The individualized TCP and NTCP models produced per BM could be compared to the generalized TCP and NTCP currently available to see which approach better predicts the known ground truth outcomes of BM response and toxicity. This comparison would allow for the first true goalposts for ML model accuracy to be set, as it could be evaluated how much accuracy from the BM response and toxicity ML models is required to outperform the generalized TCP and NTCP models. If enhanced accuracy was observed from ML models, a clinical trial could be performed with one patient arm prescribed SRS/SRT using the ML-generated TCP and NTCP values, and
the other control arm being prescribed using the current generalized models. BM control rates, toxicity rates, and patient satisfaction in the treatment decision-making process and treatment outcome could then be compared to assess if the ML-based approach offers true clinical benefit. If successful, this proposed trial would complete the clinical translation of using MRI radiomics and ML to predict BM response to SRS, marking a new paradigm of using data-driven approaches to provide individualized care to patients that maximizes their QOL and agency in healthcare decisions.

5.3 References


Appendix A

Supplementary material related to
Chapter 2: “Performance sensitivity analysis of brain metastasis stereotactic radiosurgery outcome prediction using magnetic resonance imaging radiomics”
<table>
<thead>
<tr>
<th>#</th>
<th>Type</th>
<th>Feature Name</th>
<th>#</th>
<th>Type</th>
<th>Feature Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>First Order</td>
<td>10th Percentile</td>
<td>57</td>
<td>GLRLM</td>
<td>GL Non-Uniformity</td>
</tr>
<tr>
<td>2</td>
<td>First Order</td>
<td>90th Percentile</td>
<td>58</td>
<td>GLRLM</td>
<td>GL Non-Uniformity Normalized</td>
</tr>
<tr>
<td>3</td>
<td>First Order</td>
<td>Energy</td>
<td>59</td>
<td>GLRLM</td>
<td>GL Variance</td>
</tr>
<tr>
<td>4</td>
<td>First Order</td>
<td>Entropy</td>
<td>60</td>
<td>GLRLM</td>
<td>High GL Run Emphasis</td>
</tr>
<tr>
<td>5</td>
<td>First Order</td>
<td>Interquartile Range</td>
<td>61</td>
<td>GLRLM</td>
<td>Long Run Emphasis</td>
</tr>
<tr>
<td>6</td>
<td>First Order</td>
<td>Kurtosis</td>
<td>62</td>
<td>GLRLM</td>
<td>Long Run High GL Emphasis</td>
</tr>
<tr>
<td>7</td>
<td>First Order</td>
<td>Maximum</td>
<td>63</td>
<td>GLRLM</td>
<td>Long Run Low GL Emphasis</td>
</tr>
<tr>
<td>8</td>
<td>First Order</td>
<td>Mean Absolute Deviation</td>
<td>64</td>
<td>GLRLM</td>
<td>Low GL Run Emphasis</td>
</tr>
<tr>
<td>9</td>
<td>First Order</td>
<td>Mean</td>
<td>65</td>
<td>GLRLM</td>
<td>Run Entropy</td>
</tr>
<tr>
<td>10</td>
<td>First Order</td>
<td>Median</td>
<td>66</td>
<td>GLRLM</td>
<td>Run Length Non-Uniformity</td>
</tr>
<tr>
<td>11</td>
<td>First Order</td>
<td>Minimum</td>
<td>67</td>
<td>GLRLM</td>
<td>Run Length Non-Uniformity Normalized</td>
</tr>
<tr>
<td>12</td>
<td>First Order</td>
<td>Range</td>
<td>68</td>
<td>GLRLM</td>
<td>Run Percentage</td>
</tr>
<tr>
<td>13</td>
<td>First Order</td>
<td>Robust Mean Absolute Deviation</td>
<td>69</td>
<td>GLRLM</td>
<td>Run Variance</td>
</tr>
<tr>
<td>14</td>
<td>First Order</td>
<td>Root Mean Squared</td>
<td>70</td>
<td>GLRLM</td>
<td>Short Run Emphasis</td>
</tr>
<tr>
<td>15</td>
<td>First Order</td>
<td>Skewness</td>
<td>71</td>
<td>GLRLM</td>
<td>Short Run High GL Emphasis</td>
</tr>
<tr>
<td>16</td>
<td>First Order</td>
<td>Total Energy</td>
<td>72</td>
<td>GLRLM</td>
<td>Short Run Low GL Emphasis</td>
</tr>
<tr>
<td>17</td>
<td>First Order</td>
<td>Uniformity</td>
<td>73</td>
<td>GLDM</td>
<td>Dependence Entropy</td>
</tr>
<tr>
<td>18</td>
<td>First Order</td>
<td>Variance</td>
<td>74</td>
<td>GLDM</td>
<td>Dependence Non-Uniformity</td>
</tr>
<tr>
<td>19</td>
<td>Shape &amp; Size</td>
<td>Elongation</td>
<td>75</td>
<td>GLDM</td>
<td>Dependence Non-Uniformity Normalized</td>
</tr>
<tr>
<td>20</td>
<td>Shape &amp; Size</td>
<td>Flatness</td>
<td>76</td>
<td>GLDM</td>
<td>Dependence Variance</td>
</tr>
<tr>
<td>21</td>
<td>Shape &amp; Size</td>
<td>Least Axis Length</td>
<td>77</td>
<td>GLDM</td>
<td>GL Non-Uniformity</td>
</tr>
<tr>
<td>22</td>
<td>Shape &amp; Size</td>
<td>Major Axis Length</td>
<td>78</td>
<td>GLDM</td>
<td>GL Variance</td>
</tr>
<tr>
<td>23</td>
<td>Shape &amp; Size</td>
<td>Maximum 2D Diameter Column</td>
<td>79</td>
<td>GLDM</td>
<td>High GL Emphasis</td>
</tr>
<tr>
<td>24</td>
<td>Shape &amp; Size</td>
<td>Maximum 2D Diameter Row</td>
<td>80</td>
<td>GLDM</td>
<td>Large Dependence Emphasis</td>
</tr>
<tr>
<td>25</td>
<td>Shape &amp; Size</td>
<td>Maximum 2D Diameter Slice</td>
<td>81</td>
<td>GLDM</td>
<td>Large Dependence High GL Emphasis</td>
</tr>
<tr>
<td>26</td>
<td>Shape &amp; Size</td>
<td>Maximum 3D Diameter</td>
<td>82</td>
<td>GLDM</td>
<td>Large Dependence Low GL Emphasis</td>
</tr>
<tr>
<td>27</td>
<td>Shape &amp; Size</td>
<td>Mesh Volume</td>
<td>83</td>
<td>GLDM</td>
<td>Low GL Emphasis</td>
</tr>
<tr>
<td>28</td>
<td>Shape &amp; Size</td>
<td>Minor Axis Length</td>
<td>84</td>
<td>GLDM</td>
<td>Small Dependence Emphasis</td>
</tr>
<tr>
<td>29</td>
<td>Shape &amp; Size</td>
<td>Sphericity</td>
<td>85</td>
<td>GLDM</td>
<td>Small Dependence High GL Emphasis</td>
</tr>
<tr>
<td>30</td>
<td>Shape &amp; Size</td>
<td>Surface Area</td>
<td>86</td>
<td>GLDM</td>
<td>Small Dependence Low GL Emphasis</td>
</tr>
<tr>
<td>31</td>
<td>Shape &amp; Size</td>
<td>Surface Volume Ratio</td>
<td>87</td>
<td>GLSZM</td>
<td>GL Non-Uniformity</td>
</tr>
<tr>
<td>32</td>
<td>Shape &amp; Size</td>
<td>Voxel Volume</td>
<td>88</td>
<td>GLSZM</td>
<td>GL Non-Uniformity Normalized</td>
</tr>
<tr>
<td>33</td>
<td>GLCM</td>
<td>Autocorrelation</td>
<td>89</td>
<td>GLSZM</td>
<td>GL Variance</td>
</tr>
<tr>
<td>34</td>
<td>GLCM</td>
<td>Cluster Prominence</td>
<td>90</td>
<td>GLSZM</td>
<td>High GL Zone Emphasis</td>
</tr>
<tr>
<td>35</td>
<td>GLCM</td>
<td>Cluster Shade</td>
<td>91</td>
<td>GLSZM</td>
<td>Large Area Emphasis</td>
</tr>
<tr>
<td>36</td>
<td>GLCM</td>
<td>Cluster Tendency</td>
<td>92</td>
<td>GLSZM</td>
<td>Large Area High GL Emphasis</td>
</tr>
<tr>
<td>37</td>
<td>GLCM</td>
<td>Contrast</td>
<td>93</td>
<td>GLSZM</td>
<td>Large Area Low GL Emphasis</td>
</tr>
<tr>
<td>38</td>
<td>GLCM</td>
<td>Correlation</td>
<td>94</td>
<td>GLSZM</td>
<td>Low GL Zone Emphasis</td>
</tr>
<tr>
<td>39</td>
<td>GLCM</td>
<td>Difference Average</td>
<td>95</td>
<td>GLSZM</td>
<td>Size Zone Non-Uniformity</td>
</tr>
<tr>
<td>40</td>
<td>GLCM</td>
<td>Difference Entropy</td>
<td>96</td>
<td>GLSZM</td>
<td>Size Zone Non-Uniformity Normalized</td>
</tr>
<tr>
<td>41</td>
<td>GLCM</td>
<td>Difference Variance</td>
<td>97</td>
<td>GLSZM</td>
<td>Small Area Emphasis</td>
</tr>
<tr>
<td>42</td>
<td>GLCM</td>
<td>Inverse Difference</td>
<td>98</td>
<td>GLSZM</td>
<td>Small Area High GL Emphasis</td>
</tr>
<tr>
<td>43</td>
<td>GLCM</td>
<td>Inverse Difference Moment</td>
<td>99</td>
<td>GLSZM</td>
<td>Small Area Low GL Emphasis</td>
</tr>
<tr>
<td>44</td>
<td>GLCM</td>
<td>Inverse Difference Moment Normalized</td>
<td>100</td>
<td>GLSZM</td>
<td>Zone Entropy</td>
</tr>
<tr>
<td>45</td>
<td>GLCM</td>
<td>Inverse Difference Normalized</td>
<td>101</td>
<td>GLSZM</td>
<td>Zone Percentage</td>
</tr>
<tr>
<td>46</td>
<td>GLCM</td>
<td>Informational Measure of Correlation</td>
<td>102</td>
<td>GLSZM</td>
<td>Zone Variance</td>
</tr>
<tr>
<td>47</td>
<td>GLCM</td>
<td>Informational Measure of Correlation 2</td>
<td>103</td>
<td>NGTDM</td>
<td>Busyness</td>
</tr>
<tr>
<td>48</td>
<td>GLCM</td>
<td>Inverse Variance</td>
<td>104</td>
<td>NGTDM</td>
<td>Coarseness</td>
</tr>
<tr>
<td>49</td>
<td>GLCM</td>
<td>Joint Average</td>
<td>105</td>
<td>NGTDM</td>
<td>Complexity</td>
</tr>
<tr>
<td>50</td>
<td>GLCM</td>
<td>Joint Energy</td>
<td>106</td>
<td>NGTDM</td>
<td>Contrast</td>
</tr>
<tr>
<td>51</td>
<td>GLCM</td>
<td>Joint Entropy</td>
<td>107</td>
<td>NGTDM</td>
<td>Strength</td>
</tr>
<tr>
<td>52</td>
<td>GLCM</td>
<td>Maximal Correlation Coefficient</td>
<td>108</td>
<td>Clinical</td>
<td>Gender</td>
</tr>
<tr>
<td>53</td>
<td>GLCM</td>
<td>Maximum Probability</td>
<td>109</td>
<td>Clinical</td>
<td>Age</td>
</tr>
<tr>
<td>54</td>
<td>GLCM</td>
<td>Sum Average</td>
<td>110</td>
<td>Clinical</td>
<td>Primary Cancer Active</td>
</tr>
<tr>
<td>55</td>
<td>GLCM</td>
<td>Sum Entropy</td>
<td>111</td>
<td>Clinical</td>
<td>Primary Cancer Site</td>
</tr>
<tr>
<td>56</td>
<td>GLCM</td>
<td>Sum Squares</td>
<td>112</td>
<td>Clinical</td>
<td>Primary Cancer Histology</td>
</tr>
<tr>
<td>57</td>
<td>GLCM</td>
<td>Non-Uniformity</td>
<td>113</td>
<td>Clinical</td>
<td>Systemic Metastases Status</td>
</tr>
<tr>
<td>58</td>
<td>GLCM</td>
<td>Non-Uniformity Normalized</td>
<td>114</td>
<td>Clinical</td>
<td>Systemic Therapy Status</td>
</tr>
<tr>
<td>59</td>
<td>GLCM</td>
<td>Variance</td>
<td>115</td>
<td>Clinical</td>
<td>Steroid Status</td>
</tr>
<tr>
<td>60</td>
<td>GLCM</td>
<td>Run Entropy</td>
<td>116</td>
<td>Clinical</td>
<td>WHO Score</td>
</tr>
<tr>
<td>61</td>
<td>GLCM</td>
<td>Run Length Non-Uniformity</td>
<td>117</td>
<td>Clinical</td>
<td>GTV Volume</td>
</tr>
<tr>
<td>62</td>
<td>GLCM</td>
<td>Run Length Non-Uniformity Normalized</td>
<td>118</td>
<td>Clinical</td>
<td>BM Location</td>
</tr>
<tr>
<td>63</td>
<td>GLCM</td>
<td>Run Percentage</td>
<td>119</td>
<td>Clinical</td>
<td>Dose and Fractionation</td>
</tr>
</tbody>
</table>
Table A.1: Complete catalogue of the 119 features included within the study. The number for each feature is the “Feature Look-up Number” used elsewhere such as figure A.2 and table A.3. All features not with feature type “Clinical” are radiomic features computed from the pre-treatment T1w-CE MRI, with complete documentation of the features provided by the PyRadiomics project (https://pyradiomics.readthedocs.io/en/latest/features.html). Underlined numbers are only to aid the reader in looking-up feature numbers.
Figure A.1: Schematic diagram of the ML experimental method technique used in the study. (a) shows the overall experimental method, while (b) provides enhanced detail of the model training to show how the out-of-bag samples from each tree in the RDF are used to produce the aggregated training dataset prediction probabilities. The colouring of the objects in the diagram is used to illustrate the separation of the entire dataset into training and testing datasets, the isolation of objects derived from each of the datasets during feature filtering and hyperparameter optimization (to prevent overfitting), and the recombination of objects from the datasets only during model testing and error metrics calculation. The inter-feature correlation filter used hierarchical clustering on the training dataset alone to determine groups of correlated features with a correlation coefficient $> 0.8$. For each group of correlated features, the feature most strongly correlated to BM progression was retained. This filtering of features was then transferred to the testing dataset, ensuring the testing dataset did not inform the selected features. For the experiments investigating primary cancer site and BM volume effects, the methodology shown would remain identical, except that after the “Repetitions Aggregation” processes shown in (a), the training and testing dataset prediction probabilities would be grouped by the associated BMs’ primary cancers and volumes. This would lead to an optimal ROC operating point, and average ROC, AUC, MCR, FNR, and FPR per primary cancer or volume group.
<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Optimization Domain</th>
<th>Optimization Domain Search Transform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of trees</td>
<td>[10, 1000]</td>
<td>logarithmic</td>
</tr>
<tr>
<td>Number of features to sample</td>
<td>[1, number of features]</td>
<td>linear</td>
</tr>
<tr>
<td>Minimum leaf size</td>
<td>[1, number of features / 2]</td>
<td>logarithmic</td>
</tr>
<tr>
<td>Maximum number of decision splits</td>
<td>[1, number of samples – 1]</td>
<td>logarithmic</td>
</tr>
<tr>
<td>Feature selection</td>
<td>curvature, interaction curvature</td>
<td>categorical</td>
</tr>
<tr>
<td>Decision split criterion</td>
<td>Gini’s diversity index, deviance</td>
<td>categorical</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-bag fraction</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>produces in-bag dataset that is the same size as the training dataset</td>
</tr>
<tr>
<td>Sample with replacement</td>
<td>on</td>
</tr>
<tr>
<td></td>
<td>allows for out-of-bag samples to be reserved for evaluating trained model using the training dataset</td>
</tr>
<tr>
<td>Cost per SRS response</td>
<td>equal for each response</td>
</tr>
<tr>
<td></td>
<td>false negatives and positives given equal cost</td>
</tr>
<tr>
<td>Prior</td>
<td>empirical</td>
</tr>
<tr>
<td></td>
<td>allows priors to be optimized for the study population</td>
</tr>
<tr>
<td>Algorithm for categorical features</td>
<td>exact</td>
</tr>
<tr>
<td></td>
<td>all combinations of categories for categorical features investigated at decision splits</td>
</tr>
<tr>
<td>Merge leaves</td>
<td>off</td>
</tr>
<tr>
<td></td>
<td>leaf merging not needed as trees are not pruned</td>
</tr>
<tr>
<td>Prune</td>
<td>off</td>
</tr>
<tr>
<td></td>
<td>tree pruning not needed as the maximum number of decision splits hyperparameter was optimized</td>
</tr>
<tr>
<td>Surrogate decision splits</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>not all surrogate splits investigated to decrease model training time</td>
</tr>
<tr>
<td>Weights</td>
<td>equal for each sample</td>
</tr>
<tr>
<td></td>
<td>all samples given equal importance during training</td>
</tr>
</tbody>
</table>

Table A.2: Hyperparameters for the RDF model used. For hyperparameters that underwent optimization, the optimization domain and search transform are provided. Numerical domains are indicated with minimum and maximum values in square brackets. Hyperparameter optimization was performed using 50 iterations of Bayesian optimization using the expected-improvement-plus acquisition function. The AUC on the out-of-bag samples was used as the optimization objective function. For hyperparameters that were not optimized, their value and justification are provided. For further descriptions of the hyperparameters, see the documentation for the *TreeBagger* function provided in Matlab 2019b.
Figure A.2: Comparison of normalized feature importance scores for each feature (y-axis) and for each feature volume-correlation coefficient threshold value (lower x-axis). Each colour-coded box represents the score for the given feature (row) and correlation threshold (column), with blue (1) representing a highly important feature and red (0) representing a highly unimportant feature. A gray box represents a feature that was removed due to it being dependent on or closely correlated to BM volume or diameter, and so was not available at a given threshold. The upper x-axis shows the number of features available for each threshold after the removal of volume-correlated features was performed. The feature number provided on the left y-axis can be cross-referenced with table A.1 to retrieve the given feature’s name.
<table>
<thead>
<tr>
<th>Feature Look-up Number</th>
<th>Feature Type</th>
<th>Feature Name</th>
<th>Correlation Threshold = 1 (No Feature Removal)</th>
<th>Correlation Threshold = 0.25</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Importance Score</td>
<td>Importance Rank</td>
</tr>
<tr>
<td>6</td>
<td>Radiomic - First Order</td>
<td>Kurtosis</td>
<td>0.878</td>
<td>10</td>
</tr>
<tr>
<td>48</td>
<td>Radiomic - GLCM</td>
<td>Inverse Variance</td>
<td>0.908</td>
<td>6</td>
</tr>
<tr>
<td>111</td>
<td>Clinical</td>
<td>Primary Cancer Site</td>
<td>0.990</td>
<td>2</td>
</tr>
<tr>
<td>112</td>
<td>Clinical</td>
<td>Primary Cancer Histology</td>
<td>0.926</td>
<td>4</td>
</tr>
<tr>
<td>106</td>
<td>Radiomic - NGTDM</td>
<td>Contrast</td>
<td>0.929</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>Radiomic - First Order</td>
<td>10th Percentile</td>
<td>0.811</td>
<td>15</td>
</tr>
<tr>
<td>35</td>
<td>Radiomic - GLCM</td>
<td>Cluster Shade</td>
<td>0.736</td>
<td>19</td>
</tr>
<tr>
<td>34</td>
<td>Radiomic - GLCM</td>
<td>Cluster Prominence</td>
<td>0.654</td>
<td>28</td>
</tr>
<tr>
<td>37</td>
<td>Radiomic - GLCM</td>
<td>Contrast</td>
<td>0.720</td>
<td>21</td>
</tr>
<tr>
<td>5</td>
<td>Radiomic - First Order</td>
<td>Interquartile Range</td>
<td>0.122</td>
<td>90</td>
</tr>
<tr>
<td>57</td>
<td>Radiomic - GLRLM</td>
<td>GL Non-Uniformity</td>
<td>1.000</td>
<td>1</td>
</tr>
<tr>
<td>45</td>
<td>Radiomic - GLCM</td>
<td>Inverse Difference Normalized</td>
<td>0.915</td>
<td>5</td>
</tr>
<tr>
<td>101</td>
<td>Radiomic - GLSZM</td>
<td>Zone Percentage</td>
<td>0.902</td>
<td>7</td>
</tr>
<tr>
<td>73</td>
<td>Radiomic - GLDM</td>
<td>Dependence Entropy</td>
<td>0.884</td>
<td>8</td>
</tr>
<tr>
<td>100</td>
<td>Radiomic - GLSZM</td>
<td>Zone Entropy</td>
<td>0.882</td>
<td>9</td>
</tr>
<tr>
<td>117</td>
<td>Clinical</td>
<td>GTV Volume</td>
<td>0.847</td>
<td>11</td>
</tr>
<tr>
<td>31</td>
<td>Radiomic - Shape/Size</td>
<td>Surface Volume Ratio</td>
<td>0.832</td>
<td>12</td>
</tr>
<tr>
<td>93</td>
<td>Radiomic - GLSZM</td>
<td>Large Area Low GL Emphasis</td>
<td>0.824</td>
<td>13</td>
</tr>
<tr>
<td>38</td>
<td>Radiomic - GLCM</td>
<td>Correlation</td>
<td>0.818</td>
<td>14</td>
</tr>
<tr>
<td>103</td>
<td>Radiomic - NGTDM</td>
<td>Busyness</td>
<td>0.803</td>
<td>16</td>
</tr>
<tr>
<td>58</td>
<td>Radiomic - GLRLM</td>
<td>GL Non-Uniformity Normalized</td>
<td>0.800</td>
<td>17</td>
</tr>
<tr>
<td>47</td>
<td>Radiomic - GLCM</td>
<td>Informational Measure of Correlation 2</td>
<td>0.795</td>
<td>18</td>
</tr>
</tbody>
</table>

Table A.3: Comparison of the normalized feature importance score and rank for all features with an importance score ≥ 0.75 for either correlation threshold = 1 (no features removed) or threshold = 0.25 (volume-correlated features removed). Shaded rows indicate features that were not highly important at threshold = 1, but became highly important at threshold = 0.25. Clinical features are underlined only to differentiate them more clearly against the radiomic features. The feature look-up number is provided to cross-reference this table with figure A.2, in which these look-up numbers correspond to the indices along the y-axis. Lastly, it should be noted that rows in this table are sorted according to the feature importance score/rank for correlation threshold = 0.25, and since the relative importance of the features changes at each threshold, the features are not in sorted order with respect to the threshold = 1 importance scores.
Appendix B

Supplementary material related to Chapter 3: “Assessment of brain metastasis qualitative appearance interobserver variability and comparison to magnetic resonance imaging radiomics for stereotactic radiosurgery outcome prediction”
Table B.1: Clinical feature distributions for number of BMs, BMs progressing post-SRS, and patients (where applicable) for the study sample. The “Neurological Symptoms Corticosteroid Response” feature qualitatively scores the improvement of neurological symptoms after the administration of corticosteroids.

<table>
<thead>
<tr>
<th>Clinical Features</th>
<th># Patients</th>
<th># BMs (% Progression)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>55</td>
<td>66 (21.2%)</td>
</tr>
<tr>
<td>Male</td>
<td>44</td>
<td>57 (22.8%)</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (Range)</td>
<td>58.0 (38.4-86.0) years</td>
<td></td>
</tr>
<tr>
<td><strong>Primary Cancer Active</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>44</td>
<td>55 (9.1%)</td>
</tr>
<tr>
<td>No</td>
<td>55</td>
<td>68 (32.4%)</td>
</tr>
<tr>
<td><strong>Primary Cancer Site</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung</td>
<td>59</td>
<td>70 (12.9%)</td>
</tr>
<tr>
<td>Breast</td>
<td>10</td>
<td>14 (35.7%)</td>
</tr>
<tr>
<td>Renal</td>
<td>10</td>
<td>15 (13.3%)</td>
</tr>
<tr>
<td>Colorectal</td>
<td>8</td>
<td>10 (40.0%)</td>
</tr>
<tr>
<td>Skin</td>
<td>8</td>
<td>9 (66.7%)</td>
</tr>
<tr>
<td>Other</td>
<td>4</td>
<td>5 (20.0%)</td>
</tr>
<tr>
<td><strong>Primary Cancer Histology</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>49</td>
<td>65 (20.0%)</td>
</tr>
<tr>
<td>NSCLC</td>
<td>31</td>
<td>36 (11.1%)</td>
</tr>
<tr>
<td>Melanoma</td>
<td>8</td>
<td>9 (66.7%)</td>
</tr>
<tr>
<td>Squamous Carcinoma</td>
<td>7</td>
<td>8 (50.0%)</td>
</tr>
<tr>
<td>Other</td>
<td>4</td>
<td>5 (0.0%)</td>
</tr>
<tr>
<td><strong>Extracranial Systemic Metastases</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>39</td>
<td>50 (20.0%)</td>
</tr>
<tr>
<td>No</td>
<td>60</td>
<td>73 (23.3%)</td>
</tr>
<tr>
<td><strong>Systemic Therapy Status</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Radical</td>
<td>51</td>
<td>10 (20.0%)</td>
</tr>
<tr>
<td>Palliative</td>
<td>41</td>
<td>60 (31.7%)</td>
</tr>
<tr>
<td>None</td>
<td>7</td>
<td>53 (11.3%)</td>
</tr>
<tr>
<td><strong>Neurological Symptoms Corticosteroid Response</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fully resolved</td>
<td>48</td>
<td>31 (16.1%)</td>
</tr>
<tr>
<td>Improvement</td>
<td>7</td>
<td>4 (50.0%)</td>
</tr>
<tr>
<td>Limited improvement</td>
<td>4</td>
<td>56 (25.0%)</td>
</tr>
<tr>
<td>No improvement</td>
<td>26</td>
<td>11 (9.1%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>14</td>
<td>21 (23.8%)</td>
</tr>
<tr>
<td><strong>ECOG Performance Score</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>31</td>
<td>39 (17.9%)</td>
</tr>
<tr>
<td>1</td>
<td>60</td>
<td>73 (21.9%)</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>9 (33.3%)</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2 (50.0%)</td>
</tr>
<tr>
<td><strong>GTV Volume</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (Range)</td>
<td>3.07 (0.02-30.23) cc</td>
<td></td>
</tr>
<tr>
<td>&lt; 7.5 cc</td>
<td>-</td>
<td>94 (17.0%)</td>
</tr>
<tr>
<td>&gt; 7.5 cc</td>
<td>-</td>
<td>29 (37.9%)</td>
</tr>
<tr>
<td><strong>BM Location</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supratentorial</td>
<td>-</td>
<td>96 (22.9%)</td>
</tr>
<tr>
<td>Infratentorial</td>
<td>-</td>
<td>27 (18.5%)</td>
</tr>
<tr>
<td><strong>SRS Prescription</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 Gy in 1 fraction</td>
<td>-</td>
<td>5 (0.0%)</td>
</tr>
<tr>
<td>18 Gy in 1 fraction</td>
<td>-</td>
<td>36 (30.6%)</td>
</tr>
<tr>
<td>21 Gy in 1 fraction</td>
<td>-</td>
<td>72 (13.9%)</td>
</tr>
<tr>
<td>24 Gy in 3 fractions</td>
<td>-</td>
<td>10 (60.0%)</td>
</tr>
<tr>
<td>Scanner Model and Field Strength</td>
<td>Acquisition Orientation</td>
<td>Voxel Size (mm³)</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>-------------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Siemens Magnetom Vision (1.5 T)</td>
<td>Sagittal</td>
<td>1×1×1.5</td>
</tr>
<tr>
<td>Siemens Avanto (1.5 T)</td>
<td>Sagittal</td>
<td>0.5×0.5×1</td>
</tr>
<tr>
<td></td>
<td>Axial</td>
<td>0.5×0.5×2</td>
</tr>
<tr>
<td>Siemens Magnetom Expert (1.0 T)</td>
<td>Sagittal</td>
<td>1×1×1.5</td>
</tr>
<tr>
<td>Siemens Sonata (1.5 T)</td>
<td>Axial</td>
<td>1×1×1.5</td>
</tr>
<tr>
<td></td>
<td>Axial</td>
<td>1×1×2</td>
</tr>
<tr>
<td>General Electric Signa HDxt (1.5 T)</td>
<td>Sagittal</td>
<td>1×1×1.5</td>
</tr>
</tbody>
</table>

Table B.2: Number of patients and BMs scanned by each of the MR scanner models and acquisition parameter configurations. Siemens (Erlangen, Germany); General Electric (Chicago, USA)
Table B.3: Complete catalogue of the 107 radiomic features included within the study. All features were computed on the pre-treatment T1w-CE MRI, with complete documentation of the features provided by the PyRadiomics project (https://pyradiomics.readthedocs.io/en/latest/features.html).
Figure B.1: Schematic diagram of the ML experimental method technique used in the study. (a) shows the overall experimental method, while (b) provides enhanced detail of the model training to show how the out-of-bag samples from each tree in the RDF are used to produce the aggregated training dataset prediction probabilities. The colouring of the objects in the diagram is used to illustrate the separation of the entire dataset into training and testing datasets, the isolation of objects derived from each of the datasets during feature filtering and hyperparameter optimization (to prevent overfitting), and the recombination of objects from the datasets only during model testing and error metrics calculation. The inter-feature correlation filter used hierarchical clustering on the training dataset alone to determine groups of correlated features with a correlation coefficient > 0.8. For each group of correlated features, the feature most strongly correlated to the model output was retained. This filtering of features was then transferred to the testing dataset, ensuring the testing dataset did not inform the selected features. Depending on whether the intent of the ML experiment was to predict the probability of progression or to replicate an observer’s qualitative appearance labelling, one of the six different model outputs would be provided, as shown. All experiments used the set of 107 radiomic features (table B.3), but only the radiomic and clinical progression experiment (figure 3.1d) included the set of 12 clinical features (table B.1). Table B.4 provides further detail on the hyperparameter optimization, while figure B.2 shows how this ML experiment template was applied to produce the reported results and analysis.
<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Optimization Domain</th>
<th>Optimization Domain Search Transform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of trees</td>
<td>[10, 1000]</td>
<td>logarithmic</td>
</tr>
<tr>
<td>Number of features to sample</td>
<td>[1, number of features]</td>
<td>linear</td>
</tr>
<tr>
<td>Minimum leaf size</td>
<td>[1, number of features / 2]</td>
<td>logarithmic</td>
</tr>
<tr>
<td>Maximum number of decision splits</td>
<td>[1, number of samples – 1]</td>
<td>logarithmic</td>
</tr>
<tr>
<td>Feature selection</td>
<td>curvature, interaction curvature</td>
<td>categorical</td>
</tr>
<tr>
<td>Decision split criterion</td>
<td>Gini’s diversity index, deviance</td>
<td>categorical</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-bag fraction</td>
<td>1</td>
</tr>
<tr>
<td>Sample with replacement</td>
<td>on</td>
</tr>
<tr>
<td>Cost per SRS response</td>
<td>equal for each response</td>
</tr>
<tr>
<td>Prior</td>
<td>empirical</td>
</tr>
<tr>
<td>Algorithm for categorical features</td>
<td>exact</td>
</tr>
<tr>
<td>Merge leaves</td>
<td>off</td>
</tr>
<tr>
<td>Prune</td>
<td>off</td>
</tr>
<tr>
<td>Surrogate decision splits</td>
<td>10</td>
</tr>
<tr>
<td>Weights</td>
<td>equal for each sample</td>
</tr>
</tbody>
</table>

Table B.4: Hyperparameters for the RDF model used. For hyperparameters that underwent optimization, the optimization domain and search transform are provided. Numerical domains are indicated with minimum and maximum values in square brackets. Hyperparameter optimization was performed using 50 iterations of Bayesian optimization using the expected-improvement-plus acquisition function. The AUC on the out-of-bag samples was used as the optimization objective function. For hyperparameters that were not optimized, their value and justification are provided. For further descriptions of the hyperparameters, see the documentation for the `TreeBagger` function provided in Matlab 2019b.
Feature Importance & ALE Analysis (Table 3.3, Fig. 3.5)

Model Input: Radiomic Features

Entire Dataset

Model Output: Homogeneous* "or other qualitative appearance label:
- Heterogeneous
- Cystic (Simple)
- Cystic (Complex)
- Necrotic
(from Expert 1 or expert consensus)

Feature Importance Scores

Optimal ROC Operating Point

Training Dataset Prediction Probabilities (Per Iteration)

Average MCR, FNR & FPR

Optimal ROC Operating Point (Per BM)

Distance Of Average Probability From Optimal ROC Operating Point (Per BM)

Distances from Other Qualitative Appearance Label ML Model Experiments* Qualitative Appearance Label (Per BM)

RPA Model with Qualitative Appearance ML Models (Fig. 3.1b)

Legend

Input Object → Process → Output Object

Comparison of Predicted Probability of Progression Across Qualitative Appearance Labels (Fig. 3.4)

RPA Model with Progression ML Model (Fig. 3.1c)

Choose BMs Treated with 21 Gy in 1 fx

Training Dataset Prediction Probabilities (21 Gy in 1 fx)

Average ROC & AUC

Optimal ROC Operating Point (21 Gy in 1 fx)

Choose BMs ("Low" BMs)

Training Dataset Prediction Probabilities ("Low" BMs)

Optimal ROC Operating Point ("Low" BMs)

Choose BMs ("High" BMs)

Training Dataset Prediction Probabilities ("High" BMs)

Optimal ROC Operating Point ("High" BMs)

Direct Progression ML Model (Fig. 3.1d)

Comparison of Predicted Probability of Progression Across Qualitative Appearance Labels (Fig. 3.4)

RPA Model with Progression ML Model (Fig. 3.1c)

Choose BMs Treated with 21 Gy in 1 fx

Training Dataset Prediction Probabilities (21 Gy in 1 fx)

Average ROC & AUC

Optimal ROC Operating Point (21 Gy in 1 fx)

Choose BMs ("Low" BMs)

Training Dataset Prediction Probabilities ("Low" BMs)

Optimal ROC Operating Point ("Low" BMs)

Choose BMs ("High" BMs)

Training Dataset Prediction Probabilities ("High" BMs)

Optimal ROC Operating Point ("High" BMs)

Direct Progression ML Model (Fig. 3.1d)

Comparison of Predicted Probability of Progression Across Qualitative Appearance Labels (Fig. 3.4)

RPA Model with Progression ML Model (Fig. 3.1c)

Choose BMs Treated with 21 Gy in 1 fx

Training Dataset Prediction Probabilities (21 Gy in 1 fx)

Average ROC & AUC

Optimal ROC Operating Point (21 Gy in 1 fx)

Choose BMs ("Low" BMs)

Training Dataset Prediction Probabilities ("Low" BMs)

Optimal ROC Operating Point ("Low" BMs)

Choose BMs ("High" BMs)

Training Dataset Prediction Probabilities ("High" BMs)

Optimal ROC Operating Point ("High" BMs)

Direct Progression ML Model (Fig. 3.1d)
Figure B.2: Schematic of the ML experiments and analysis associated with (a) the radiomic appearance experiments (as needed for figure 3.1b), (c) the radiomic progression experiment (as needed for figure 3.1c), (d) the radiomic and clinical progression experiment (as needed for figure 3.1d), and (b) the analysis performed for ML model interpretation using the results from (a) and (c). The “Bootstrapped Resampling Machine Learning Experiment” process blocks in (a), (b), (d) each represent an instance of the ML experiment common template described in the article text under heading “Machine learning experimental design” (section 3.2.4) and shown schematically above in figure B.1 of this supplementary document. The specific model inputs/outputs and experimental results associated with each experiment instance are shown in (a), (b), and (d), on which further analysis is performed. From each ML experiment instance, there are training and testing dataset prediction probabilities per bootstrapped resampling iteration. To get the “Average Prediction Probability (Per BM)”, the testing dataset probabilities are iterated through to find all instances in which a given BM was randomly chosen to be in the testing dataset for that bootstrapped iteration. The prediction probabilities for the BM from all these instances are then aggregated and their average taken. This process is then repeated for each BM in the entire dataset. A similar process is performed on the “Training Dataset Prediction Probabilities” when the “Choose...” process blocks are used, except in these cases the training dataset probabilities are just aggregated according to the “Choose...” process rule, with no average taken, allowing for the creation of an average ROC from the chosen training dataset probabilities instead. The specific details of the model interpretation analysis processes outlined in (b) are provided in the article text under the heading “Post-SRS progression machine learning model interpretation” (section 3.2.8).
Table B.5: Confusion matrices across appearance labels (A–E), for each pairwise comparison between observers. The highlighted cells indicate instances of agreement between observers, allowing for the calculation of the agreement rate when summed and divided by the total number of BMs ($n = 123$). The diagonal of the larger “observer-level” matrix only contains these highlighted cells as each observer is in perfect agreement with themselves, and so these values provide the number of each appearance label an observer called.

<table>
<thead>
<tr>
<th>Expert 1</th>
<th>Expert 2</th>
<th>Expert 3</th>
<th>Trainee 1</th>
<th>Trainee 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
</tr>
<tr>
<td>40</td>
<td>19</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>23</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>7</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
</tr>
<tr>
<td>21</td>
<td>14</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>59</td>
<td>2</td>
<td>2</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>11</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>35</td>
<td>4</td>
<td>23</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>7</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>39</td>
<td>1</td>
<td>9</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
</tr>
<tr>
<td>17</td>
<td>7</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>45</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>33</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
</tr>
<tr>
<td>11</td>
<td>14</td>
<td>50</td>
<td>6</td>
<td>26</td>
</tr>
</tbody>
</table>

Appearance Legend:
A: Homogeneous
B: Heterogeneous
C: Cystic (Simple)
D: Cystic (Complex)
E: Necrotic
Figure B.3: KM analysis plots for progressive disease for each individual observer for comparison against Expert 1 (see figure 3.2a). The risk group number for each risk curve is labelled on the right y-axis, and the number of BMs at risk per 3-month follow-up interval is given below each x-axis. The stated $p$-values are from the log-rank test performed over all risk groups.
Figure B.4: Error metrics from the radiomic Expert 1 appearance experiments using the observer labels from the original study. As each appearance label (e.g. “homogeneous”) had specific models trained to make a binary labelling decision (e.g. “homogeneous” or “not homogeneous”), error metrics for each appearance label are presented. The error bars for the non-$\text{AUC}_{0.632^+}$ error metrics represent the 95% CI of each value determined from the 250 bootstrapped resampling iterations.
Figure B.5: ALE plots for the highly importance features from table 3.3 that were not included in figure 3.5, as these features were not also highly important for any of the radiomic appearance label experiments. The Pearson correlation coefficient values for each plot’s pair of ALE curve are given in table 3.3.
Appendix C

Supplementary material related to Chapter 4: “Dual-centre validation of using magnetic resonance imaging radiomics to predict stereotactic radiosurgery outcomes”
<table>
<thead>
<tr>
<th>Scanner Model and Field Strength</th>
<th>Acquisition Orientation</th>
<th>Voxel Size (mm³)</th>
<th># BMs (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centre A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Siemens Magnetom Vision (1.5 T)</td>
<td>Sagittal</td>
<td>1×1×1.5</td>
<td>39 (28.2%)</td>
</tr>
<tr>
<td>Siemens Avanto (1.5 T)</td>
<td>Sagittal</td>
<td>0.5×0.5×1</td>
<td>37 (13.5%)</td>
</tr>
<tr>
<td></td>
<td>Axial</td>
<td>0.5×0.5×2</td>
<td>8 (0.0%)</td>
</tr>
<tr>
<td>Siemens Magnetom Expert (1.0 T)</td>
<td>Sagittal</td>
<td>1×1×1.5</td>
<td>29 (31.0%)</td>
</tr>
<tr>
<td></td>
<td>Sagittal</td>
<td>1×1×1.5</td>
<td>5 (40.0%)</td>
</tr>
<tr>
<td>Siemens Sonata (1.5 T)</td>
<td>Axial</td>
<td>1×1×1.5</td>
<td>1 (0.0%)</td>
</tr>
<tr>
<td></td>
<td>Axial</td>
<td>1×1×2</td>
<td>3 (0.0%)</td>
</tr>
<tr>
<td>General Electric Signa HDxt (1.5 T)</td>
<td>Sagittal</td>
<td>1×1×1.5</td>
<td>1 (0.0%)</td>
</tr>
<tr>
<td>Centre B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General Electric Optima MR450w (1.5 T)</td>
<td>Axial</td>
<td>0.5×0.5×2</td>
<td>58 (12.1%)</td>
</tr>
<tr>
<td></td>
<td>Axial</td>
<td>0.4×0.4×2</td>
<td>6 (16.7%)</td>
</tr>
<tr>
<td>General Electric Signa HDxt (1.5 T)</td>
<td>Axial</td>
<td>0.5×0.5×2</td>
<td>17 (23.5%)</td>
</tr>
<tr>
<td></td>
<td>Axial</td>
<td>0.4×0.4×2</td>
<td>13 (23.1%)</td>
</tr>
<tr>
<td></td>
<td>Axial</td>
<td>0.4×0.4×1</td>
<td>5 (0.0%)</td>
</tr>
<tr>
<td>Siemens Magnetom Aera (1.5 T)</td>
<td>Sagittal</td>
<td>0.7×0.7×5</td>
<td>2 (0.0%)</td>
</tr>
<tr>
<td></td>
<td>Axial</td>
<td>0.4×0.4×0.9</td>
<td>3 (66.7%)</td>
</tr>
<tr>
<td></td>
<td>Axial</td>
<td>1×1×1</td>
<td>3 (66.7%)</td>
</tr>
<tr>
<td>Siemens Avanto (1.5 T)</td>
<td>Axial</td>
<td>0.4×0.4×4</td>
<td>1 (0.0%)</td>
</tr>
<tr>
<td>Phillips Achieva (1.5 T)</td>
<td>Axial</td>
<td>0.4×0.4×3</td>
<td>1 (0.0%)</td>
</tr>
</tbody>
</table>

Table C.1: Number of BMs imaged by each imaging configuration dependent on scanner model, acquisition orientation, and voxel size. The first two voxel size dimensions for each imaging configuration are the in-plane voxel dimensions, while the last voxel size dimension represents the slice thickness. Scanner manufacturers: Siemens (Erlangen, Germany), General Electric (Chicago, USA), Phillips (Amsterdam, The Netherlands)
### Table C.2: Post-SRS progression, no progression and pseudo-progression calls for Centre B’s dataset, broken down according to patterns of post-SRS BM size over time.

For BMs with multi-step size over time patterns, sub-groups of BMs are described where clinical judgement was used in some cases to make a progression vs. pseudo-progression call. Progression scores marked with a "*" indicate BMs with complex clinical factors that make definitive progression calls difficult. These five BMs were possible confounders in model training, and so were removed for a specific test to gauge their effect. No effect was observed, and so the BMs remained in the dataset for the entire study.
<table>
<thead>
<tr>
<th>#</th>
<th>Type</th>
<th>Feature Name</th>
<th>#</th>
<th>Type</th>
<th>Feature Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>First Order</td>
<td>10th Percentile</td>
<td>57</td>
<td>GLRLM</td>
<td>GL Non-Uniformity</td>
</tr>
<tr>
<td>2</td>
<td>First Order</td>
<td>90th Percentile</td>
<td>58</td>
<td>GLRLM</td>
<td>GL Non-Uniformity Normalized</td>
</tr>
<tr>
<td>3</td>
<td>First Order</td>
<td>Energy</td>
<td>59</td>
<td>GLRLM</td>
<td>GL Variance</td>
</tr>
<tr>
<td>4</td>
<td>First Order</td>
<td>Entropy</td>
<td>60</td>
<td>GLRLM</td>
<td>High GL Run Emphasis</td>
</tr>
<tr>
<td>5</td>
<td>First Order</td>
<td>Interquartile Range</td>
<td>61</td>
<td>GLRLM</td>
<td>Long Run Emphasis</td>
</tr>
<tr>
<td>6</td>
<td>First Order</td>
<td>Kurtosis</td>
<td>62</td>
<td>GLRLM</td>
<td>Long Run High GL Emphasis</td>
</tr>
<tr>
<td>7</td>
<td>First Order</td>
<td>Maximum</td>
<td>63</td>
<td>GLRLM</td>
<td>Long Run Low GL Emphasis</td>
</tr>
<tr>
<td>8</td>
<td>First Order</td>
<td>Mean Absolute Deviation</td>
<td>64</td>
<td>GLRLM</td>
<td>Low GL Run Emphasis</td>
</tr>
<tr>
<td>9</td>
<td>First Order</td>
<td>Mean</td>
<td>65</td>
<td>GLRLM</td>
<td>Run Entropy</td>
</tr>
<tr>
<td>10</td>
<td>First Order</td>
<td>Median</td>
<td>66</td>
<td>GLRLM</td>
<td>Run Length Non-Uniformity</td>
</tr>
<tr>
<td>11</td>
<td>First Order</td>
<td>Minimum</td>
<td>67</td>
<td>GLRLM</td>
<td>Run Length Non-Uniformity Normalized</td>
</tr>
<tr>
<td>12</td>
<td>First Order</td>
<td>Range</td>
<td>68</td>
<td>GLRLM</td>
<td>Run Percentage</td>
</tr>
<tr>
<td>13</td>
<td>First Order</td>
<td>Robust Mean Absolute Deviation</td>
<td>69</td>
<td>GLRLM</td>
<td>Run Variance</td>
</tr>
<tr>
<td>14</td>
<td>First Order</td>
<td>Root Mean Squared</td>
<td>70</td>
<td>GLRLM</td>
<td>Short Run Emphasis</td>
</tr>
<tr>
<td>15</td>
<td>First Order</td>
<td>Skewness</td>
<td>71</td>
<td>GLRLM</td>
<td>Short Run High GL Emphasis</td>
</tr>
<tr>
<td>16</td>
<td>First Order</td>
<td>Total Energy</td>
<td>72</td>
<td>GLRLM</td>
<td>Short Run Low GL Emphasis</td>
</tr>
<tr>
<td>17</td>
<td>First Order</td>
<td>Uniformity</td>
<td>73</td>
<td>GLDM</td>
<td>Dependence Entropy</td>
</tr>
<tr>
<td>18</td>
<td>First Order</td>
<td>Variance</td>
<td>74</td>
<td>GLDM</td>
<td>Dependence Non-Uniformity</td>
</tr>
<tr>
<td>19</td>
<td>Shape &amp; Size</td>
<td>Elongation</td>
<td>75</td>
<td>GLDM</td>
<td>Dependence Non-Uniformity Normalized</td>
</tr>
<tr>
<td>20</td>
<td>Shape &amp; Size</td>
<td>Flatness</td>
<td>76</td>
<td>GLDM</td>
<td>Dependence Variance</td>
</tr>
<tr>
<td>21</td>
<td>Shape &amp; Size</td>
<td>Least Axis Length</td>
<td>77</td>
<td>GLDM</td>
<td>GL Non-Uniformity</td>
</tr>
<tr>
<td>22</td>
<td>Shape &amp; Size</td>
<td>Major Axis Length</td>
<td>78</td>
<td>GLDM</td>
<td>GL Variance</td>
</tr>
<tr>
<td>23</td>
<td>Shape &amp; Size</td>
<td>Maximum 2D Diameter Column</td>
<td>79</td>
<td>GLDM</td>
<td>High GL Emphasis</td>
</tr>
<tr>
<td>24</td>
<td>Shape &amp; Size</td>
<td>Maximum 2D Diameter Row</td>
<td>80</td>
<td>GLDM</td>
<td>Large Dependence Emphasis</td>
</tr>
<tr>
<td>25</td>
<td>Shape &amp; Size</td>
<td>Maximum 2D Diameter Slice</td>
<td>81</td>
<td>GLDM</td>
<td>Large Dependence High GL Emphasis</td>
</tr>
<tr>
<td>26</td>
<td>Shape &amp; Size</td>
<td>Maximum 3D Diameter</td>
<td>82</td>
<td>GLDM</td>
<td>Large Dependence Low GL Emphasis</td>
</tr>
<tr>
<td>27</td>
<td>Shape &amp; Size</td>
<td>Mesh Volume</td>
<td>83</td>
<td>GLDM</td>
<td>Low GL Emphasis</td>
</tr>
<tr>
<td>28</td>
<td>Shape &amp; Size</td>
<td>Minor Axis Length</td>
<td>84</td>
<td>GLDM</td>
<td>Small Dependence Emphasis</td>
</tr>
<tr>
<td>29</td>
<td>Shape &amp; Size</td>
<td>Sphericity</td>
<td>85</td>
<td>GLDM</td>
<td>Small Dependence High GL Emphasis</td>
</tr>
<tr>
<td>30</td>
<td>Shape &amp; Size</td>
<td>Surface Area</td>
<td>86</td>
<td>GLDM</td>
<td>Small Dependence Low GL Emphasis</td>
</tr>
<tr>
<td>31</td>
<td>Shape &amp; Size</td>
<td>Surface Volume Ratio</td>
<td>87</td>
<td>GLSZM</td>
<td>GL Non-Uniformity</td>
</tr>
<tr>
<td>32</td>
<td>Shape &amp; Size</td>
<td>Voxel Volume</td>
<td>88</td>
<td>GLSZM</td>
<td>GL Non-Uniformity Normalized</td>
</tr>
<tr>
<td>33</td>
<td>GLCM</td>
<td>Autocorrelation</td>
<td>89</td>
<td>GLSZM</td>
<td>GL Variance</td>
</tr>
<tr>
<td>34</td>
<td>GLCM</td>
<td>Cluster Prominence</td>
<td>90</td>
<td>GLSZM</td>
<td>High GL Zone Emphasis</td>
</tr>
<tr>
<td>35</td>
<td>GLCM</td>
<td>Cluster Shade</td>
<td>91</td>
<td>GLSZM</td>
<td>Large Area Emphasis</td>
</tr>
<tr>
<td>36</td>
<td>GLCM</td>
<td>Cluster Tendency</td>
<td>92</td>
<td>GLSZM</td>
<td>Large Area High GL Emphasis</td>
</tr>
<tr>
<td>37</td>
<td>GLCM</td>
<td>Contrast</td>
<td>93</td>
<td>GLSZM</td>
<td>Large Area Low GL Emphasis</td>
</tr>
<tr>
<td>38</td>
<td>GLCM</td>
<td>Correlation</td>
<td>94</td>
<td>GLSZM</td>
<td>Low GL Zone Emphasis</td>
</tr>
<tr>
<td>39</td>
<td>GLCM</td>
<td>Difference Average</td>
<td>95</td>
<td>GLSZM</td>
<td>Size Zone Non-Uniformity</td>
</tr>
<tr>
<td>40</td>
<td>GLCM</td>
<td>Difference Entropy</td>
<td>96</td>
<td>GLSZM</td>
<td>Size Zone Non-Uniformity Normalized</td>
</tr>
<tr>
<td>41</td>
<td>GLCM</td>
<td>Difference Variance</td>
<td>97</td>
<td>GLSZM</td>
<td>Small Area Emphasis</td>
</tr>
<tr>
<td>42</td>
<td>GLCM</td>
<td>Inverse Difference</td>
<td>98</td>
<td>GLSZM</td>
<td>Small Area High GL Emphasis</td>
</tr>
<tr>
<td>43</td>
<td>GLCM</td>
<td>Inverse Difference Moment</td>
<td>99</td>
<td>GLSZM</td>
<td>Small Area Low GL Emphasis</td>
</tr>
<tr>
<td>44</td>
<td>GLCM</td>
<td>Inverse Difference Moment Normalized</td>
<td>100</td>
<td>GLSZM</td>
<td>Zone Entropy</td>
</tr>
<tr>
<td>45</td>
<td>GLCM</td>
<td>Inverse Difference Moment Normalized</td>
<td>101</td>
<td>GLSZM</td>
<td>Zone Percentage</td>
</tr>
<tr>
<td>46</td>
<td>GLCM</td>
<td>Informational Measure of Correlation1</td>
<td>102</td>
<td>GLSZM</td>
<td>Zone Variance</td>
</tr>
<tr>
<td>47</td>
<td>GLCM</td>
<td>Informational Measure of Correlation2</td>
<td>103</td>
<td>NGTDM</td>
<td>Busyness</td>
</tr>
<tr>
<td>48</td>
<td>GLCM</td>
<td>Inverse Entropy</td>
<td>104</td>
<td>NGTDM</td>
<td>Coarseness</td>
</tr>
<tr>
<td>49</td>
<td>GLCM</td>
<td>Joint Average</td>
<td>105</td>
<td>NGTDM</td>
<td>Complexity</td>
</tr>
<tr>
<td>50</td>
<td>GLCM</td>
<td>Joint Energy</td>
<td>106</td>
<td>NGTDM</td>
<td>Contrast</td>
</tr>
<tr>
<td>51</td>
<td>GLCM</td>
<td>Joint Entropy</td>
<td>107</td>
<td>NGTDM</td>
<td>Strength</td>
</tr>
<tr>
<td>52</td>
<td>GLCM</td>
<td>Maximal Correlation Coefficient</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>GLCM</td>
<td>Maximum Probability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>GLCM</td>
<td>Sum Average</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>GLCM</td>
<td>Sum Entropy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>GLCM</td>
<td>Sum Squares</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table C.3: Complete catalogue of the 107 radiomic features included within the study. All radiomic features were computed from the pre-treatment T1w-CE MRI, with complete documentation of the features provided by the PyRadiomics project (https://pyradiomics.readthedocs.io/en/latest/features.html).
(a) Model External Validation

Switch

Centre A Dataset → Testing Dataset → Transfer Correlation Filter → Filtered Testing Dataset

Centre B Dataset → Training Dataset → Inter-Feature Correlation Filter → Filtered Training Dataset

Optimal ROC Operating Point → Training Dataset Prediction Probabilities → Model Training → Optimal Hyperparameters

FNR, FPR & MCR → Average ROC & AUC → Testing Dataset Prediction Probabilities → Model Testing → Filtered Testing Dataset

(b) Methodology External Validation or Pooled Data Validation

Centre A Dataset

Centre B Dataset

Pooled Dataset

Switch

Centre A Dataset

Centre B Dataset

Bootstrapped Repetition Split

Filtering Correlation Filter

Optimal ROC Operating Point

Model Training

Optimal Hyperparameters

FNR, FPR & MCR

Average ROC & AUC

250x Repetitions

Repetitions Aggregation

Model Testing

Filtered Testing Dataset

(c) Detailed Schematic of “Model Training”

Filtered Training Dataset

Hyperparameters

Trained Model

Training Dataset Prediction Probabilities

Sample Selection Per Tree

Out-Of-Bag Samples

In-Bag Samples

Tree Training

Trained Tree

Aggregated Trees

Aggregation Across Trees

Out-Of-Bag Prediction Probabilities

Tree Testing

Legend

For n trees in forest

Input Object

Process

Output Object

201
Figure C.1: Schematic of the experimental design for (a) model external validation and (b) methodology external validation or pooled data validation. (a) can be performed with either Centre A or B’s dataset as the training or testing dataset. (b) can be performed with Centre A’s dataset, Centre B’s dataset, or the pooled dataset. (c) shows a detailed view of the “Model Training” process present in (a) and (b). The schematics of (a) and (b) both demonstrate how training and testing datasets are kept separate until model testing to prevent data leakage.
<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Optimization Domain</th>
<th>Optimization Domain Search Transform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of trees</td>
<td>[10, 1000]</td>
<td>logarithmic</td>
</tr>
<tr>
<td>Number of features to sample</td>
<td>[1, number of features]</td>
<td>linear</td>
</tr>
<tr>
<td>Minimum leaf size</td>
<td>[1, number of features / 2]</td>
<td>logarithmic</td>
</tr>
<tr>
<td>Maximum number of decision splits</td>
<td>[1, number of samples – 1]</td>
<td>logarithmic</td>
</tr>
<tr>
<td>Feature selection</td>
<td>curvature, interaction curvature</td>
<td>categorical</td>
</tr>
<tr>
<td>Decision split criterion</td>
<td>Gini’s diversity index, deviance</td>
<td>categorical</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-bag fraction</td>
<td>1 produces in-bag dataset that is the same size as the training dataset</td>
</tr>
<tr>
<td>Sample with replacement</td>
<td>on allows for out-of-bag samples to be reserved for evaluating trained model using the training dataset</td>
</tr>
<tr>
<td>Cost per SRS response</td>
<td>equal for each response allows priors to be optimized for the study population</td>
</tr>
<tr>
<td>Prior</td>
<td>empirical</td>
</tr>
<tr>
<td>Algorithm for categorical features</td>
<td>exact all combinations of categories for categorical features investigated at decision splits</td>
</tr>
<tr>
<td>Merge leaves</td>
<td>off leaf merging not needed as trees are not pruned</td>
</tr>
<tr>
<td>Prune</td>
<td>off tree pruning not needed as the maximum number of decision splits hyperparameter was optimized</td>
</tr>
<tr>
<td>Surrogate decision splits</td>
<td>10 not all surrogate splits investigated to decrease model training time</td>
</tr>
<tr>
<td>Weights</td>
<td>equal for each sample all samples given equal importance during training</td>
</tr>
</tbody>
</table>

Table C.4: Hyperparameters for the RDF model used. For hyperparameters that underwent optimization, the optimization domain and search transform are provided. Numerical domains are indicated with minimum and maximum values in square brackets. Hyperparameter optimization was performed using 50 iterations of Bayesian optimization using the expected-improvement-plus acquisition function. The AUC on the out-of-bag samples was used as the optimization objective function. For hyperparameters that were not optimized, their value and justification are provided. For further descriptions of the hyperparameters, see the documentation for the `TreeBagger` function provided in Matlab 2019b.
Pooled Data Validation
Per Centre Sub-Analysis

Feature Harmonization
Test Dataset: B

Feature Harmonization
Test Dataset: A

n = 3 Consolidated Features

n = 5 Consolidated Features
Figure C.2: ROC curves and operating points for additional validation experiments. (a) and (b) show similar per centre sub-analysis as shown in figure 4.2f, but for using only clinical features and using all features, respectively. (c) and (d) show the results of performing only ComBat harmonization for model external validation experiments using only radiomic features. (e) and (f) are identical to figure 4.4b, but for the $n = 3$ and $n = 5$ cases of performing model external validation when feature harmonization and consolidation is applied.
Appendix D

Machine learning software framework

In order to perform the experiments within this thesis, a ML software framework of over 170,000 lines of code and documentation was constructed in Matlab 2019b through the efforts of Ryan Alfano, Salma Dammak, Carol Johnson, Aaron Ward, and myself. This framework was designed to be highly flexible and modular, allowing for customized ML experiments to be performed with minimal additional programming required. Object-oriented programming (OOP) was therefore used to encapsulate data within classes and define functional interfaces for using each class. OOP class abstraction and inheritance allowed for the minimization of repeated code and for new features to be added to the framework quickly and efficiently. The framework was also thoroughly tested using over 3,000 per class unit tests and framework-wide “end-to-end” tests. Unified Modelling Language (UML) class diagrams are provided in this appendix to show the structure and capabilities of the ML software framework. Given the size of the framework, a simplified overview is provided in figure D.2, while the remaining figures offer more detailed diagrams. In the more detailed diagrams, some auxiliary classes shown may have their classes properties and methods hidden for clarity, with another figure providing full details. Only public methods are shown for each class, as protected and private methods are invisible to the user and only used internally by each class.
<table>
<thead>
<tr>
<th>Symbol Legend</th>
<th>Property Name Legend</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td>bVariableName</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>dVariableName</td>
</tr>
<tr>
<td><strong>E</strong></td>
<td>iVariableName</td>
</tr>
<tr>
<td></td>
<td>chVariableName</td>
</tr>
<tr>
<td></td>
<td>sVariableName</td>
</tr>
<tr>
<td></td>
<td>dtVariableName</td>
</tr>
<tr>
<td></td>
<td>eVariableName</td>
</tr>
<tr>
<td></td>
<td>oVariableName</td>
</tr>
<tr>
<td></td>
<td>v.VariableName</td>
</tr>
<tr>
<td></td>
<td>mN.VariableName</td>
</tr>
<tr>
<td></td>
<td>cN.VariableName</td>
</tr>
<tr>
<td></td>
<td>cNx.VariableName</td>
</tr>
</tbody>
</table>

- **A** Abstract class
- **C** Concrete class
- **E** Enumeration
- Diamond: Private class property
- Diamond with pentagon: Protected class property
- Circle: Public class method
- Arrows: Inheritance (B inherits from A)
- Diamond with arrow: Aggregation (A contains objects of class B, which may exist without A)†
- Diamond with pentagon arrow: Composition (A contains objects of class B, which cease exist without A)†
- Arrows with bar: Association (A and B are associated)

†Numbers or "*" on relationships indicate "1-to-n" or "1-to-many" relationships, with "*" indicating "many"

![Figure D.1: Symbol and property name legends for UML class diagrams.](image-url)
Figure D.2: UML class diagram providing an overview of the ML software framework. Class properties and methods omitted for clarity, but are provided in subsequent figures. Legend provided in figure D.1.
Geometrical Imaging Object Class Structure

- PostTransformImageVolumeGeometry
- EqualityBound
- chObjMatFileVarName
- chImageDataMatFileVarName
- bDefaultRenderShowEdges
- chObjMatFileVarName
- dOrthogonalPrecisionBound
- dAcquisitionSliceThickness_mm
- dPrecisionBound

- GetOriginalFilePath()
- Load()
- ImageVolume()
- View()
- LoadVolumeData()
- UnloadVolumeData()
- IsVolumeDataLoaded()
- Save()
- SetMatFilePath()
- SaveToNIfTI()
- MustHaveRegionsOfInterest()

- GetImageData()
- GetCroppedImageData()
- GetImageDataMinimumValue()
- GetImageDataMaximumValue()
- GetMatFilePath()
- ForceApplyAllTransforms()
- InterpolateOntoTargetGeometry()
- ReassignFirstVoxel()
- ReassignFirstVoxelToAlignWithRASCoordinateSystem()
- NormalizeIntensityLinearly()
- TransformImageDataWithCustomFunction()
- CastImageDataToType()
- SetRegionsOfInterest()
- RemoveRegionsOfInterest()
- GetNumberOfRegionsOfInterest()
- GetMinimallyBoundedImageDataAndMaskByRegionOfInterestNumber()

- disp()
- oRegionsOfInterest
- dImageDataMaximumValue
- dImageDataMinimumValue
- chMatFilePath
- m3xImageData

- View()
- LoadVolumeData()
- UnloadVolumeData()
- IsVolumeDataLoaded()
- Save()
- SetMatFilePath()
- SaveToNIfTI()
- MustHaveRegionsOfInterest()
- MustBeValidRegionOfInterestNumbers()

- GetImageDataMaximumWindow()
- GetDefaultImageDisplayBounds()
- GetMatFilePath()
- ForceApplyAllTransforms()
- RemoveAllTransforms()
- InterpolateToIsotropicVoxelResolution()
- ReassignFirstVoxelToAlignWithRASCoordinateSystem()
- MatchHistogramToReferenceImageVolume()
- NormalizeIntensityLinearly()
- NormalizeIntensityWithZScoreTransform()
- CastImageDataToType()
- Crop()
- GetRegionsOfInterest()
- RemoveRegionsOfInterest()
- dImageDataMaximumValue
- chMatFilePath
- chImageDataMatFileVarName

Figure D.3: UML class diagram of the “GeometricalImagingObject” portion of the ML software framework. The “GeometricalImagingObject” class allows for integration of imaging data (images and ROIs) into the framework. Legend provided in figure D.1.
Figure D.4: UML class diagram of the “ImageVolume” portion of the ML software framework. The “ImageVolume” class allows for the loading, viewing, and manipulation of images. Legend provided in figure D.1.
Figure D.5: UML class diagram of the “RegionsOfInterest” portion of the ML software framework. The “RegionsOfInterest” class allows for the loading, viewing, and manipulation of ROIs. Legend provided in figure D.1.
Figure D.6: UML class diagram of the “ImagingObjectTransform” portion of the ML software framework. The “ImagingObjectTransform” class allows for the application of transforms to imaging data. Legend provided in figure D.1.
Figure D.7: UML class diagram of the “ImageVolumeViewer” portion of the ML software framework. The ML framework included a basic 3D image and ROIs viewer for validation of data and transforms. Legend provided in figure D.1.
Figure D.8: UML class diagram of the “FeatureExtractionImageVolumeHandler” portion of the ML software framework. The “FeatureExtractionImageVolumeHandler” class allows for the selection, labelling, and viewing of images and ROIs in preparation for radiomic feature extraction. Legend provided in figure D.1.
Figure D.9: UML class diagram of the “Feature” portion of the ML software framework. The “Feature” class allows for the programming of radiomic features and then their extraction from images. Only three concrete classes for each feature type (e.g. “FirstOrderFeature” or “GLCMFeature”) are shown for clarity. Legend provided in figure D.1.
Figure D.10: UML class diagram of the “FeatureExtractionRecord” portion of the ML software framework. The “FeatureExtractionRecord” class allows for storing how features (radiomic or otherwise) were gathered. These records persist throughout the framework, allowing for a final testing dataset predicted classification to be traced back to a specific image and ROI or other data source. Legend provided in figure D.1.
Figure D.11: UML class diagram of the “FeatureValues” portion of the ML software framework. The “FeatureValues” class allows for the storage of feature values for a given set of samples and features. The “LabelledFeatureValues” class adds the ability to store ground truth labels alongside the feature data for each sample. The “ClassificationGuessResult” class adds the ability to store the predicted probability produced by a trained classifier during testing, which can then be used for the calculation of error metrics. “FeatureValues” classes present to the user as a matrix of data (samples along rows, features along columns) that can be manipulated as any other matrix in Matlab is, except with additional validation being performed. Legend provided in figure D.1.
Figure D.12: UML class diagram of the “Classifier” portion of the ML software framework. The “Classifier” class allows for the optimization, training, and testing of ML classifier models. Only three concrete classes for “MATLABClassifier” and “PRToolsClassifier” are shown for clarity. Legend provided in figure D.1.
Figure D.13: UML class diagram of the “FeatureSelector” portion of the ML software framework. The “FeatureSelector” class allows for the selection of a subset of features using a variety of optimization methods and objective functions. Legend provided in figure D.1.
Figure D.14: UML class diagram of the “HyperParameterOptimizer” portion of the ML software framework. The “HyperParameterOptimizer” class allows for optimization of a classifier’s hyperparameters using a variety of optimization methods and objective functions. Legend provided in figure D.1.
Figure D.15: UML class diagram of the “ErrorMetric” and “ObjectiveFunction” portions of the ML software framework. The “ObjectiveFunction” class allows for defining customized objective functions for feature selection and hyperparameter optimization. The “ErrorMetric” and “ErrorMetricCalculator” classes allows for calculating a variety of error metrics for objective functions and classifier testing. Legend provided in figure D.1.
Figure D.16: Class methods and properties of the “Experiment” class in the ML software framework. The UML class diagram of the “Experiment” portion of the ML software framework is provided in figure D.17. Legend provided in figure D.1.
Figure D.17: UML class diagram of the “Experiment” portion of the ML software framework. The “Experiment” class enables complete reproduction of a ML experiment, including automated experiment journaling, storage of experimental results, and creating a complete copy of the code used to run the experiment. Random number generation is made reproducible independent of whether the experiment is repeated on a local single-core, local multi-core, or computing cluster environment. An integrated computing resource manager for distributing workloads between multiple users of a Matlab distributed parallel computing pool is also included. The class methods and properties of the “Experiment” class are provided in figure D.16 for clarity. Legend provided in figure D.1.
Appendix E

Clinical trial data collection and management overview

To collect and process the LHSC SRS dataset used in chapter 4, a clinical trial data collection and management system was needed. As shown in figure E.1 below, various types of patient data was collected from the RT record and verify system (RVS), multiple RT treatment planning systems (TPSs), the LHSC picture archiving and communication system (PACS), and the LHSC electronic health record (EHR) system. All imaging data was stored in the Digital Imaging and Communications in Medicine (DICOM) format, and so could be transferred between PACSs using DICOM transfer protocols over the LHSC network to the clinical research MIM PACS and DICOM workstation. Non-imaging data was collected manually and stored either in the REDCap clinical trial platform or, for patient overview data (used for study population identification) and SRS treatment data, stored directly in a laboratory MySQL relational database. The data collection forms used in the REDCap platform are provided in appendix F. The imaging data was anonymized in the MIM DICOM workstation and any required corrections to DICOM files were applied.

In order to process the data, it had to be stored locally within the imaging laboratory. To store the imaging data, a virtual PACS server was created using the open-source Orthanc
PACS project. The anonymized DICOM data could then be directly transferred to the Orthanc PACS using DICOM protocols over the LHSC network. The non-imaging data in the REDCap platform was accessed directly using the REDCap application programming interface (API) over the LHSC network. Using the REDCap API, only anonymized data could be selected to be transferred. The raw data from REDCap platform was then processed and inserted into the MySQL relational database, according to the MySQL database schema provided in appendix G. The Orthanc PACS and MySQL database were linked by having DICOM object identifiers stored in the MySQL database, which could then be used to automatically retrieve the appropriate DICOM files from the Orthanc PACS.

To then produce a dataset that could be used for ML experiments, a software framework was built in Matlab to interface with the clinical trial database. As shown in appendix H, this framework was built using OOP such that the data from the Orthanc PACS and MySQL database could be stored and accessed. By using the software framework, the MySQL database could be queried, a population of patients identified, and then their clinical data points and treatment outcomes could be automatically gathered into a dataset. The T1w-CE MRI and ROIs DICOM files could also be automatically pulled for each patient from the Orthanc PACS, at which point radiomic features could be extracted and added to the dataset. The ML experiments outlined in chapter 4 could then be performed.
Figure E.1: Schematic representation of the data sources and data management systems used to collect and process the LHSC SRS dataset. The “Clinical Research Systems” represent research systems within LHSC, and therefore the data is not immediately available for processing and is not anonymized. The “Laboratory Research Systems” are research systems within the imaging laboratory, and therefore the data can be processed directly from these systems. Only anonymized data is stored on the “Laboratory Research Systems”, as the DICOM data is anonymized within the MIM DICOM workstation and the REDCap platform exported only anonymized data.
Appendix F

Clinical trial data collection forms

The following figures show the data collection forms created within the REDCap platform that were used to collect the LHSC SRS dataset in chapter 4. The figures show all the possible data fields that could be collected per form, but REDCap branching logic (not shown) was used to selectively hide data fields if they were not relevant (e.g. “Estrogen Receptor Status” in the “Primary Cancer Histology” form would only be shown if “Breast Cancer” was the selected “Histopathology Type”). Forms collecting per BM data are truncated to show fields up to “BM 3”, but the system was capable of collecting data for up to 10 BMs per patient. Only the fields for however many BMs a patient had would be displayed using REDCap’s branching logic.
### Patient Information

<table>
<thead>
<tr>
<th>Study ID</th>
<th>____________________________</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Male&lt;br&gt;Female</td>
</tr>
<tr>
<td>Date of Birth</td>
<td>____________________________</td>
</tr>
<tr>
<td>Patient Deceased</td>
<td>Yes&lt;br&gt;No&lt;br&gt;Likely</td>
</tr>
<tr>
<td>Date of Death</td>
<td>____________________________</td>
</tr>
<tr>
<td>Last Interaction Date</td>
<td>____________________________</td>
</tr>
</tbody>
</table>

Figure F.1: “Patient Information” REDCap data collection form.

<table>
<thead>
<tr>
<th>Study ID</th>
<th>____________________________</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis Date</td>
<td>____________________________</td>
</tr>
<tr>
<td>Diagnosis ICD-10-CM Code</td>
<td>____________________________</td>
</tr>
<tr>
<td>Diagnosis ICD-10-CM Description</td>
<td>____________________________</td>
</tr>
</tbody>
</table>

Figure F.2: “Aria Disease Diagnosis” REDCap data collection form. A patient could have multiple copies of this form completed depending on their diagnosis.
| Study ID |

| Site | ○ Lung  
○ Breast  
○ Renal  
○ Skin  
○ Colorectal  
○ Oesophageal  
○ Thyroid  
○ Oral  
○ GI  
○ Liver  
○ Pancreas  
○ Gynecological  
○ Ovarian  
○ Prostate  
○ Testicular  
○ Urinary  
○ Head & Neck  
○ Other  
○ Unknown |

| Histopathology Date |

| ○ Primary Cancer  
○ Extracranial Metastasis |

| Histopathology Source  
○ Poorly Differentiated  
○ Moderately Differentiated  
○ Highly Differentiated  
○ Unknown |

| Histopathology Type  
○ Carcinoma (Adeno)  
○ Carcinoma (Squamous)  
○ Carcinoma (Basal)  
○ Carcinoma (Papillary)  
○ Carcinoma (Urothelial)  
○ Carcinoma (Renal)  
○ Carcinoma (Mammary)  
○ Melanoma  
○ Sarcoma  
○ Neuroendocrine  
○ Germ Cell  
○ Unknown |

| Histopathology Type  
○ SCLC (Pure)  
○ SCLC (Combined)  
○ NSCLC (Adeno)  
○ NSCLC (Squamous)  
○ NSCLC (Large Cell)  
○ NSCLC (Other)  
○ Unknown |
<table>
<thead>
<tr>
<th>Status</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD-L1 Status</td>
<td>Strongly Positive, Weakly Positive, Negative, Unknown</td>
</tr>
<tr>
<td>ALK Status</td>
<td>Positive, Negative, Unknown</td>
</tr>
<tr>
<td>EGFR Status</td>
<td>Positive, Negative, Unknown</td>
</tr>
<tr>
<td>ROS1 Status</td>
<td>Positive, Negative, Unknown</td>
</tr>
<tr>
<td>BRAF Status</td>
<td>Positive, Negative, Unknown</td>
</tr>
<tr>
<td>KRAS Status</td>
<td>Positive, Negative, Unknown</td>
</tr>
<tr>
<td>Estrogen Receptor Status</td>
<td>Positive, Negative, Unknown</td>
</tr>
<tr>
<td>Progesterone Receptor Status</td>
<td>Positive, Negative, Unknown</td>
</tr>
<tr>
<td>Her2/neu Receptor Status</td>
<td>Positive, Negative, Unknown</td>
</tr>
</tbody>
</table>

Figure F.3: “Primary Cancer Histopathology” REDCap data collection form (two pages). A patient could have multiple copies of this form completed if they had multiple pathology samples taken.
### Brain Metastasis Histopathology

<table>
<thead>
<tr>
<th>Study ID</th>
<th>______________________</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site</td>
<td>Lung</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Histopathology Date</th>
<th>______________________</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Sampled Brain Metastasis Number</th>
<th>______________________</th>
</tr>
</thead>
</table>

| Malignancy Present | Yes | No | |
|--------------------|-----|----| |

| Necrosis Present | Yes | No | |
|------------------|-----|----| |

<table>
<thead>
<tr>
<th>Histopathology Differentiation</th>
<th>Poorly Differentiated</th>
<th>Moderately Differentiated</th>
<th>Highly Differentiated</th>
<th>Unknown</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Histopathology Type</th>
<th>Carcinoma (Adeno)</th>
<th>Carcinoma (Squamous)</th>
<th>Carcinoma (Basal)</th>
<th>Carcinoma (Papillary)</th>
<th>Carcinoma (Urothelial)</th>
<th>Carcinoma (Renal)</th>
<th>Carcinoma (Mammary)</th>
<th>Melanoma</th>
<th>Sarcoma</th>
<th>Neuroendocrine</th>
<th>Unknown</th>
</tr>
</thead>
</table>

232
<table>
<thead>
<tr>
<th>Histopathology Type</th>
<th>SCLC (Pure)</th>
<th>SCLC (Combined)</th>
<th>NSCLC (Adeno)</th>
<th>NSCLC (Squamous)</th>
<th>NSCLC (Large Cell)</th>
<th>NSCLC (Other)</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD-L1 Status</td>
<td>Strongly Positive</td>
<td>Weakly Positive</td>
<td>Negative</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALK Status</td>
<td>Positive</td>
<td>Negative</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EGFR Status</td>
<td>Positive</td>
<td>Negative</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROS1 Status</td>
<td>Positive</td>
<td>Negative</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRAF Status</td>
<td>Positive</td>
<td>Negative</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRAS Status</td>
<td>Positive</td>
<td>Negative</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estrogen Receptor Status</td>
<td>Positive</td>
<td>Negative</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Progesterone Receptor Status</td>
<td>Positive</td>
<td>Negative</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Her2/neu Receptor Status</td>
<td>Positive</td>
<td>Negative</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure F.4: “Brain Metastasis Histopathology” REDCap data collection form (two pages). A patient could have multiple copies of this form completed if they had multiple pathology samples taken.
### Systemic Therapy

<table>
<thead>
<tr>
<th>Study ID</th>
<th>____________________________</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Systemic Therapy Type</th>
<th>Chemotherapy</th>
<th>Targeted Therapy</th>
<th>Hormone Therapy</th>
<th>Immunotherapy</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Systemic Therapy Used For Radiosensitizing</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Start Date Known</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Start Date</th>
<th>____________________________</th>
</tr>
</thead>
<tbody>
<tr>
<td>Therapy Agent Known</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Therapy Agent</th>
<th>____________________________</th>
</tr>
</thead>
</table>

Figure F.5: “Systemic Therapy” REDCap data collection form. A patient could have multiple copies of this form completed if they had multiple systemic therapies.
<table>
<thead>
<tr>
<th>Study ID</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Treatment Type</th>
<th>Surgery</th>
<th>WBRT</th>
<th>SRS</th>
<th>SRT</th>
<th>Other</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Treatment Type</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>New BMs Targeted</td>
<td>Yes</td>
</tr>
<tr>
<td>BM 1 Targeted</td>
<td>Yes</td>
</tr>
<tr>
<td>BM 2 Targeted</td>
<td>Yes</td>
</tr>
<tr>
<td>BM 3 Targeted</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Figure F.6: “Salvage Therapy” REDCap data collection form. A patient could have multiple copies of this form completed if they had multiple salvage treatments.
<table>
<thead>
<tr>
<th>Study ID</th>
<th>______________________</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>______________________</td>
</tr>
</tbody>
</table>

**Treatment Type**

- Dexamethasone
- Hyperbaric Oxygen
- Resection
- Other

**New BMs Targeted**

- Yes
- No

**BM 1 Targeted**

- Yes
- No

**BM 2 Targeted**

- Yes
- No

**BM 3 Targeted**

- Yes
- No

**BM 4 Targeted**

- Yes
- No

**BM 5 Targeted**

- Yes
- No

**BM 6 Targeted**

- Yes
- No

**BM 7 Targeted**

- Yes
- No

**BM 8 Targeted**

- Yes
- No

**BM 9 Targeted**

- Yes
- No

**BM 10 Targeted**

- Yes
- No

Figure F.7: “Radionecrosis Treatment” REDCap data collection form. A patient could have multiple copies of this form completed if they had multiple RN treatments.
### Brain Radiation Course

<table>
<thead>
<tr>
<th>Study ID</th>
<th>____________________________</th>
</tr>
</thead>
</table>
| Course Date | ____________________________  
(Only Y-M collected, D defaulted to 1) |
| Intent |  
- Palliative  
- Curative  
- Curative with Chemotherapy  
- Radical  
- Unknown |
| Number of BMs Treated | ____________________________ |
| Number of Beam Sets in Treatment | ____________________________ |

#### Beam Set 1
- Prescribed Dose [Gy] | ____________________________
- Calculated Dose [Gy] | ____________________________
- Number of Fractions Prescribed | ____________________________
- BMs Targeted | ____________________________

#### Beam Set 2
- Prescribed Dose [Gy] | ____________________________
- Calculated Dose [Gy] | ____________________________
- Number of Fractions Prescribed | ____________________________
- BMs Targeted | ____________________________
**Beam Set 3**

<table>
<thead>
<tr>
<th>Prescribed Dose [Gy]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculated Dose [Gy]</td>
</tr>
<tr>
<td>Number of Fractions Prescribed</td>
</tr>
<tr>
<td>BMs Targeted</td>
</tr>
</tbody>
</table>

**Beam Set 4**

<table>
<thead>
<tr>
<th>Prescribed Dose [Gy]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculated Dose [Gy]</td>
</tr>
<tr>
<td>Number of Fractions Prescribed</td>
</tr>
<tr>
<td>BMs Targeted</td>
</tr>
</tbody>
</table>

**Beam Set 5**

<table>
<thead>
<tr>
<th>Prescribed Dose [Gy]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculated Dose [Gy]</td>
</tr>
<tr>
<td>Number of Fractions Prescribed</td>
</tr>
<tr>
<td>BMs Targeted</td>
</tr>
</tbody>
</table>

Figure F.8: “Brain Radiation Course” REDCap data collection form (two pages). A patient could have multiple copies of this form completed if they had multiple BM RT treatments.
<table>
<thead>
<tr>
<th>Study ID</th>
<th>__________________________</th>
</tr>
</thead>
<tbody>
<tr>
<td>Course Date</td>
<td>__________________________</td>
</tr>
<tr>
<td>(Only Y-M collected, D defaulted to 1)</td>
<td></td>
</tr>
</tbody>
</table>
| Intent | Palliative  
Curative  
Curative with Chemotherapy  
Radical  
Unknown |
| Number of Beam Sets in Treatment | __________________________ |

### Beam Set 1

| Site | LUNL  
LUNR  
LUNB  
CHEL  
CHER  
CHEB  
MEDI |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculated Dose [Gy]</td>
<td>__________________________</td>
</tr>
<tr>
<td>Number of Fractions Prescribed</td>
<td>__________________________</td>
</tr>
</tbody>
</table>

### Beam Set 2

| Site | LUNL  
LUNR  
LUNB  
CHEL  
CHER  
CHEB  
MEDI |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculated Dose [Gy]</td>
<td>__________________________</td>
</tr>
<tr>
<td>Number of Fractions Prescribed</td>
<td>__________________________</td>
</tr>
</tbody>
</table>
### Beam Set 3

<table>
<thead>
<tr>
<th>Site</th>
<th>LUNL</th>
<th>LUNR</th>
<th>LUNB</th>
<th>CHEL</th>
<th>CHER</th>
<th>CHEB</th>
<th>MEDI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Calculated Dose [Gy]

__________________________________

Number of Fractions Prescribed

__________________________________

### Beam Set 4

<table>
<thead>
<tr>
<th>Site</th>
<th>LUNL</th>
<th>LUNR</th>
<th>LUNB</th>
<th>CHEL</th>
<th>CHER</th>
<th>CHEB</th>
<th>MEDI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Calculated Dose [Gy]

__________________________________

Number of Fractions Prescribed

__________________________________

### Beam Set 5

<table>
<thead>
<tr>
<th>Site</th>
<th>LUNL</th>
<th>LUNR</th>
<th>LUNB</th>
<th>CHEL</th>
<th>CHER</th>
<th>CHEB</th>
<th>MEDI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Calculated Dose [Gy]

__________________________________

Number of Fractions Prescribed

__________________________________

Figure F.9: “Lung Radiation Course” REDCap data collection form (two pages). A patient could have multiple copies of this form completed if they had multiple lung RT treatments.
<table>
<thead>
<tr>
<th>Study ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scan Date</td>
</tr>
</tbody>
</table>

### BM 1

| **RANO-BM Measurement [mm]**
using axial plane | (longest diameter in axial slice passing through only the BM (e.g. not through healthy tissue, cysts or surgical cavities) | enter 0.1mm if BM is visible, but too small to be measurable) |
|------------------|-------------------------------------------------|-------------------------------------------------------------|
| **Anterior-Posterior Diameter [mm]**
using axial plane | (enter 0.1mm if BM is visible, but too small to be measurable) |
| **Mediolateral Diameter [mm]**
using coronal plane | (enter 0.1mm if BM is visible, but too small to be measurable) |
| **Craniocaudal Diameter [mm]**
using sagittal plane | (enter 0.1mm if BM is visible, but too small to be measurable) |
| **BM Is Parenchymal** | ○ Yes  
○ No |
| **Surgical Cavity Present** | ○ Yes  
○ No |
| **Edema Present** | ○ Yes  
○ No |
| **Mass Effect Present (Moderate or High)** | ○ Yes  
○ No |
| **BM Appearance** | ○ Homogeneous  
○ Heterogeneous  
○ Rim-Enhancing |
| **Rim-Enhancement Type** | ○ Single Cystic  
○ Multi-Cystic  
○ Necrotic Core |
### BM 2

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RANO-BM Measurement [mm]</td>
<td>(longest diameter in axial slice passing through only the BM (e.g. not through healthy tissue, cysts or surgical cavities)</td>
</tr>
<tr>
<td>Anterior-Posterior Diameter [mm]</td>
<td>(enter 0.1mm if BM is visible, but too small to be measurable)</td>
</tr>
<tr>
<td>Mediolateral Diameter [mm]</td>
<td>(enter 0.1mm if BM is visible, but too small to be measurable)</td>
</tr>
<tr>
<td>Craniocaudal Diameter [mm]</td>
<td>(enter 0.1mm if BM is visible, but too small to be measurable)</td>
</tr>
</tbody>
</table>

- BM Is Parenchymal
  - Yes
  - No

- Surgical Cavity Present
  - Yes
  - No

- Edema Present
  - Yes
  - No

- Mass Effect Present (Moderate or High)
  - Yes
  - No

- BM Appearance
  - Homogeneous
  - Heterogeneous
  - Rim-Enhancing

- Rim-Enhancement Type
  - Single Cystic
  - Multi-Cystic
  - Necrotic Core

### BM 3

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RANO-BM Measurement [mm]</td>
<td>(longest diameter in axial slice passing through only the BM (e.g. not through healthy tissue, cysts or surgical cavities)</td>
</tr>
<tr>
<td>Anterior-Posterior Diameter [mm]</td>
<td>(enter 0.1mm if BM is visible, but too small to be measurable)</td>
</tr>
<tr>
<td>Measured Parameter</td>
<td>Measurement Method</td>
</tr>
<tr>
<td>---------------------------------------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>Mediolateral Diameter [mm]</td>
<td>using coronal plane</td>
</tr>
<tr>
<td>Craniocaudal Diameter [mm]</td>
<td>using sagittal plane</td>
</tr>
<tr>
<td>BM Is Parenchymal</td>
<td>Yes</td>
</tr>
<tr>
<td>Surgical Cavity Present</td>
<td>Yes</td>
</tr>
<tr>
<td>Edema Present</td>
<td>Yes</td>
</tr>
<tr>
<td>Mass Effect Present (Moderate or High)</td>
<td>Yes</td>
</tr>
<tr>
<td>BM Appearance</td>
<td>Homogeneous</td>
</tr>
<tr>
<td></td>
<td>Rim-Enhancing</td>
</tr>
<tr>
<td>Rim-Enhancement Type</td>
<td>Single Cystic</td>
</tr>
<tr>
<td></td>
<td>Necrotic Core</td>
</tr>
</tbody>
</table>

Figure F.10: “Brain Radiology Pre-Radiation” REDCap data collection form (three pages).
<table>
<thead>
<tr>
<th>Study ID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Scan Date</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

| Number of New BMs | ○ Uncountable  
|                   | ○ Countable  
|                   | (Leptomeningeal disease should be included here. Bone metastases should NOT be recorded here, but should be included in the “Notes” section below.) |

<table>
<thead>
<tr>
<th>Number of New BMs</th>
<th></th>
</tr>
</thead>
</table>

| Number of Suspected New BMs | ○ Uncountable  
|                             | ○ Countable  |

<table>
<thead>
<tr>
<th>Number of Suspected New BMs</th>
<th></th>
</tr>
</thead>
</table>

| BMs Present That Were Missed In Previous Follow-up | ○ Yes  
|                                                 | ○ No  |

<table>
<thead>
<tr>
<th>BM 1</th>
</tr>
</thead>
</table>

| RANO-BM Measurement [mm]  
| using axial plane | (longest diameter in axial slice passing through only the BM (e.g. not through healthy tissue, cysts or surgical cavities) | enter 0.1mm if BM is visible, but too small to be measurable) |
|-----------------|----------|

| Anterior-Posterior Diameter [mm]  
<table>
<thead>
<tr>
<th>using axial plane</th>
<th>(enter 0.1mm if BM is visible, but too small to be measurable)</th>
</tr>
</thead>
</table>

| Mediolateral Diameter [mm]  
<table>
<thead>
<tr>
<th>using coronal plane</th>
<th>(enter 0.1mm if BM is visible, but too small to be measurable)</th>
</tr>
</thead>
</table>

| Craniocaudal Diameter [mm]  
<table>
<thead>
<tr>
<th>using sagittal plane</th>
<th>(enter 0.1mm if BM is visible, but too small to be measurable)</th>
</tr>
</thead>
</table>

| Pseudo-Progression | ○ None  
|                   | ○ Suspected  
|                   | ○ Likely  |
Figure F.11: “Brain Radiology Follow-up” REDCap data collection form (two pages). A patient would have a copy of this form completed for each follow-up MRI after their SRS treatment.
<table>
<thead>
<tr>
<th>Study ID</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BM 1</td>
<td></td>
</tr>
<tr>
<td>Pseudo-Progression Confirmed</td>
<td>No&lt;br&gt;Yes&lt;br&gt;Unknown&lt;br&gt;No Progression/Pseudo-Progression Was Observed</td>
</tr>
<tr>
<td>Pseudo-Progression was Radiation Necrosis</td>
<td>No&lt;br&gt;Suspected&lt;br&gt;Yes&lt;br&gt;Unknown</td>
</tr>
<tr>
<td>Pseudo-Progression was Adverse Radiation Effect (ARE)</td>
<td>No&lt;br&gt;Suspected&lt;br&gt;Yes&lt;br&gt;Unknown</td>
</tr>
<tr>
<td>How was pseudo-progression confirmed?</td>
<td></td>
</tr>
<tr>
<td>BM 2</td>
<td></td>
</tr>
<tr>
<td>Pseudo-Progression Confirmed</td>
<td>No&lt;br&gt;Yes&lt;br&gt;Unknown&lt;br&gt;No Progression/Pseudo-Progression Was Observed</td>
</tr>
<tr>
<td>Pseudo-Progression was Radiation Necrosis</td>
<td>No&lt;br&gt;Suspected&lt;br&gt;Yes&lt;br&gt;Unknown</td>
</tr>
<tr>
<td>Pseudo-Progression was Adverse Radiation Effect (ARE)</td>
<td>No&lt;br&gt;Suspected&lt;br&gt;Yes&lt;br&gt;Unknown</td>
</tr>
<tr>
<td>How was pseudo-progression confirmed?</td>
<td></td>
</tr>
</tbody>
</table>
### BM 3

| Pseudo-Progression Confirmed | ☐ No  
|                             | ☐ Yes  
|                             | ☐ Unknown  
|                             | ☐ No Progression/Pseudo-Progression Was Observed  

| Pseudo-Progression was Radiation Necrosis | ☐ No  
|                                         | ☐ Suspected  
|                                         | ☐ Yes  
|                                         | ☐ Unknown  

| Pseudo-Progression was Adverse Radiation Effect (ARE) | ☐ No  
|                                                      | ☐ Suspected  
|                                                      | ☐ Yes  
|                                                      | ☐ Unknown  

How was pseudo-progression confirmed?

__________________________

Figure F.12: “Brain Radiology Pseudo-Progression Conclusion” REDCap data collection form (two pages).
Appendix G

Clinical trial database schema

The data collected via the REDCap platform (appendix F) to form the LHSC SRS dataset in chapter 4 was stored in a MySQL relational database. The database schema is presented in this appendix via multiple schematics to show portions of the database.
Figure G.1: Legend for database schema figures.

Figure G.2: Database schema representing the storage of data collected from the Aria RVS. Legend provided in figure G.1.
Figure G.3: Database schema representing the storage of BM SRS treatment data. Legend provided in figure G.1.
Figure G.4: Database schema representing the storage of imaging metadata. The raw imaging data was stored separately in research Orthanc PACS (figure E.1), which the metadata stored in the MySQL database referenced via the “orthanc_pacs_uuid” values. Legend provided in figure G.1.
Figure G.5: Database schema representing the storage of systemic, RN, and SRS salvage treatment data. Legend provided in figure G.1.
Figure G.6: Database schema representing the storage of histopathology data. Legend provided in figure G.1.
Figure G.7: Database schema representing the storage of BM radiologic assessment data. Legend provided in figure G.1.
Appendix H

Clinical trial database software framework

The data stored in the MySQL relational database (appendix G) was accessed to perform the experiments in chapter 4 using an OOP software framework built in Matlab 2019b. The UML class diagrams for this framework are presented in this appendix. The UML class diagram legend used in appendix D is used for this appendix as well (figure D.1). As with appendix D, only public methods are shown for each class.
Figure H.1: UML class diagram for the objects accessing patient and BM data. The properties and methods for these classes are only shown in this figure to simplify the further figures. Legend provided in figure D.1.
Figure H.2: UML class diagram for the objects accessing data collected from the Aria RVS. Legend provided in figure D.1.
Figure H.3: UML class diagram for the objects accessing BM SRS treatment data. Legend provided in figure D.1.
Figure H.4: UML class diagram for the objects accessing imaging data. Legend provided in figure D.1.
Figure H.5: UML class diagram for the objects accessing systemic, RN, and SRS salvage treatment data. Legend provided in figure D.1.
Figure H.6: UML class diagram for the objects accessing histopathology data. Legend provided in figure D.1.
Figure H.7: UML class diagram for the objects accessing radiologic assessment data. Legend provided in figure D.1.
Appendix I

Clinical trial research ethics board approval

The collection of the LHSC SRS dataset used in chapter 4 was approved by the Western University Health Sciences Research Ethics Board. The ethics approval is documented in figure I.1, with the ethics application documented in figures I.2 and I.3.
Dear Dr. Aaron Ward

The Western University Health Science Research Ethics Board (HSREB) has reviewed and approved the above mentioned study as described in the WREM application form, as of the HSREB Initial Approval Date noted above. This research study is to be conducted by the investigator noted above. All other required institutional approvals and mandated training must also be obtained prior to the conduct of the study.

Documents Approved:

<table>
<thead>
<tr>
<th>Document Name</th>
<th>Document Type</th>
<th>Document Date</th>
<th>Document Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>REB Research Plan</td>
<td>Protocol</td>
<td>09/Feb/2021</td>
<td>v4.0</td>
</tr>
<tr>
<td>REB REDCap Instruments</td>
<td>Other Data Collection Instruments</td>
<td>09/Feb/2021</td>
<td>v2.0</td>
</tr>
</tbody>
</table>

Documents Acknowledged:

<table>
<thead>
<tr>
<th>Document Name</th>
<th>Document Type</th>
<th>Document Date</th>
<th>Document Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>REB Study Budget</td>
<td>Study budget</td>
<td>08/Jan/2021</td>
<td>v1</td>
</tr>
<tr>
<td>REDCap Approval Documentation</td>
<td>Technology Review document</td>
<td>12/Jan/2021</td>
<td>v1</td>
</tr>
<tr>
<td>Orthanc Approval Documentation</td>
<td>Technology Review document</td>
<td>26/Jan/2021</td>
<td>v1</td>
</tr>
</tbody>
</table>

No deviations from, or changes to, the protocol or WREM application should be initiated without prior written approval of an appropriate amendment from Western HSREB, except when necessary to eliminate immediate hazard(s) to study participants or when the change(s) involves only administrative or logistical aspects of the trial.

REB members involved in the research project do not participate in the review, discussion or decision.

The Western University HSREB operates in compliance with, and is constituted in accordance with, the requirements of the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2); the International Conference on Harmonisation Good Clinical Practice Consolidated Guideline (ICH GCP); Part C, Division 5 of the Food and Drug Regulations; Part 4 of the Natural Health Products Regulations; Part 3 of the Medical Devices Regulations and the provisions of the Ontario Personal Health Information Protection Act (PHIPA 2004) and its applicable regulations. The HSREB is registered with the U.S. Department of Health & Human Services under the IRB registration number.

Please do not hesitate to contact us if you have any questions.

Sincerely,

Ms. Jhananiee Subendran, Ethics Coordinator on behalf of Dr. Joseph Gilbert, HSREB Vice-Chair

Note: This correspondence includes an electronic signature (validation and approval via an online system that is compliant with all regulations).

Figure I.1: Study approval letter from the Western University Health Sciences Research Ethics Board. Sensitive information redacted.
1.1 *If this is the first time you are submitting this particular application to the REB, select “Initial Submission”. If this application form has already been reviewed by the REB and they issued recommendations, select “Response to REB recommendations”:

- Initial Submission
- Response to REB recommendations

1.2 *Does this study involve the London hospitals (see HELP text if you are unsure):

- No this study does not involve the London hospitals
- Yes this study involves the London hospitals and this form has been exported from ReDA.
- This study involves the London Hospitals but a ReDA application has not been completed. NOTE: You cannot submit this application until the ReDA application has FIRST been completed and you exported from ReDA to WREM.

*What is the Lawson ReDA number associated with this study?


*As this study IS taking place in the hospital, copy and paste the following in the below email text box:

Email
Once the PI is added to this form you MUST also add them into the ROLES tile (See ROLES tile in the actions items on the left hand side of your screen).

1.3 Use the Search field to enter the Principal Investigator (PI) details from the WREM user directory:

*Prefix       *First Name       *Last Name
Dr.           Aaron           Ward

*Telephone

*Email

*Indicate the PI's Western Academic Faculty/Department:

Schulich-Medical Biophysics

Indicate the PI's Hospital Department/Division:

Oncology - London Regional Cancer Program

1.4 *Are there any additional study team members (incl. students, postdocs, coordinators, managers, etc.) from Western and/or its affiliate institutions working on this study?

☐ Yes there are additional study team members
☐ No other study team members involved
1.4 *Complete the following information for additional study team members (from Western and or its affiliate institutions) who are working on this study:

<table>
<thead>
<tr>
<th>Prefix</th>
<th>*First Name</th>
<th>*Last Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mr</td>
<td>David</td>
<td>DeVries</td>
</tr>
</tbody>
</table>

Telephone

*Email

1.4 Specify ROLE, DUTIES, and DEPARTMENT/FACULTY. (E.g. John Doe - Research Assistant - responsible for recruitment, interviews and analysis of data; Psychology/Social Sciences.):

Role: Research Coordinator
Duties: Responsible for managing practical aspects of the study, maintaining REDCap database, and analysis of data
Department/Faculty: Department of Medical Biophysics, Schulich School of Medicine & Dentistry, Western University

1.4a *Are there any additional study team members (incl. students, postdocs, coordinators, managers, etc.) from Western and/or its affiliate institutions working on this study?

☐ Yes there are other study team members
☐ No other study team members

1.4a *Use the Search field to enter the following information for additional study team members (from Western and or its affiliate institutions) who are working on this study:

<table>
<thead>
<tr>
<th>Prefix</th>
<th>*First Name</th>
<th>*Last Name</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Joanna</td>
<td>Laba</td>
</tr>
</tbody>
</table>

Telephone

*Email

1.4a *Specify ROLE, DUTIES, and DEPARTMENT/FACULTY. (E.g. John Doe - Research Assistant - responsible for recruitment, interviews and analysis of data; Psychology/Social Sciences.):

Role: Co-investigator from radiation oncology
Department/Faculty: Department of Oncology, Schulich School of Medicine & Dentistry, Western University; Department of Radiation Oncology, London Regional Cancer Program, London Health Sciences Centre
### 1.4b Are there any additional study team members (incl. students, postdocs, coordinators, managers, etc.) from Western and/or its affiliate institutions working on this study?

- [ ] Yes there are other study team members
- [x] No other study team members

**Use the Search field to enter the following information for additional study team members (from Western and or its affiliate institutions) who are working on this study:**

<table>
<thead>
<tr>
<th>Prefix</th>
<th>*First Name</th>
<th>*Last Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr</td>
<td>Andrew</td>
<td>Leung</td>
</tr>
</tbody>
</table>

**Telephone**

**Email**

### 1.4b Specify ROLE, DUTIES, and DEPARTMENT/FACULTY. (E.g. John Doe - Research Assistant - responsible for recruitment, interviews and analysis of data; Psychology/Social Sciences.):

- Role: Co-investigator from medical imaging
- Department/Faculty: Department of Medical Imaging, Schulich School of Medicine & Dentistry, Western University; Department of Medical Imaging, London Health Sciences Centre

### 1.4c Are there any additional study team members (incl. students, postdocs, coordinators, managers, etc.) from Western and/or its affiliate institutions working on this study?

- [ ] Yes there are other study team members
- [x] No other study team members

**Use the Search field to enter the following information for additional study team members (from Western and or its affiliate institutions) who are working on this study:**

<table>
<thead>
<tr>
<th>Prefix</th>
<th>*First Name</th>
<th>*Last Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr</td>
<td>George</td>
<td>Hajdok</td>
</tr>
</tbody>
</table>

**Telephone**

**Email**
1.4c *Specify ROLE, DUTIES, and DEPARTMENT/FACULTY. (E.g. John Doe - Research Assistant - responsible for recruitment, interviews and analysis of data; Psychology/Social Sciences.):

Role: Co-investigator from medical physics
Department/Faculty: Department of Medical Biophysics, Schulich School of Medicine & Dentistry, Western University; Department of Radiation Oncology, London Regional Cancer Program, London Health Sciences Centre

1.4d *Are there any additional study team members (incl. students, postdocs, coordinators, managers, etc.) from Western and/or its affiliate institutions working on this study?

☑ Yes there are other study team members
☐ No other study team members

1.4d *Use the Search field to enter the following information for additional study team members (from Western and or its affiliate institutions) who are working on this study:

Prefix  *First Name  *Last Name
Dr  Ghada  Alqaidy

Telephone

*Email

1.4d *Specify ROLE, DUTIES, and DEPARTMENT/FACULTY. (E.g. John Doe - Research Assistant - responsible for recruitment, interviews and analysis of data; Psychology/Social Sciences.):

Role: Research Assistant
Duties: Collection of retrospective patient medical imaging data
Department/Faculty: Department of Medical Imaging, London Health Sciences Centre

1.4e *Are there any additional study team members (incl. students, postdocs, coordinators, managers, etc.) from Western and/or its affiliate institutions working on this study?

☑ Yes there are other study team members
☐ No other study team members
1.4e *Use the Search field to enter the following information for additional study team members (from Western and or its affiliate institutions) who are working on this study:

Prefix *First Name *Last Name
Dr Terence Tang

Telephone

*Email

1.4e *Specify ROLE, DUTIES, and DEPARTMENT/FACULTY. (E.g. John Doe - Research Assistant - responsible for recruitment, interviews and analysis of data; Psychology/Social Sciences.):

Role: Research Assistant
Duties: Collection of retrospective patient radiation oncology data
Department/Faculty: Department of Radiation Oncology, London Regional Cancer Program, London Health Sciences Centre

1.4f *Are there any additional study team members (incl. students, postdocs, coordinators, managers, etc.) from Western and/or its affiliate institutions working on this study?

☐ Yes there are other study team members
☐ No other study team members

1.5 *Enter the Complete Study Title:

Retrospective Outcome Analysis of Brain Metastasis Patients Treated with Stereotactic Radiosurgery/Radiotherapy (SRS/SRT)

1.6 *What is the acronym or nickname/short title for the study? (NOTE: The acronym or nickname/short title will be used to identify the study and will be included in all notifications and REB applications associated with this project.):

Retrospective Outcome Analysis of Brain Metastasis Patients Treated with SRS/SRT
1.7

1.7 *What type of REB submission is this?

- Full Board
- Delegated Level 2 - Prospective data collection
- Delegated Level 1 - Retrospective study data and/or biological sample collection

1.8

1.8 *Are any of the investigator(s) based at any of the sites below or will the study utilize any patient data/biological specimens, staff resources or facilities within any of these sites? (Please indicate all applicable sites):

- No

LHSC Sites

- Adult Eating Disorder Service (Riverview)
- Byron Family Medical Centre
- Children’s Hospital
- Fowler Kennedy Sports Medicine
- Kidney Care Centre (Westmount)
- London Regional Cancer Program (LRCP)
- Southwestern Ontario Regional Base Hospital Program
- Stroke Prevention & Atherosclerosis Research Centre
- University Hospital (UH)
- Victoria Family Medical Centre
- Victoria Hospital (VH)

St Joseph’s Sites

- Mount Hope Centre for Long Term Care
- Parkwood Institute – Main Building
- Parkwood Institute Mental Health Care
- Southwest Centre for Forensic Mental Health Care
- St. Joseph’s Family Medical and Dental Centre
- St. Joseph’s Hospital

1.9

1.9 *Is this study directly related to a study at this institution (e.g., is this study a sub-study, extension, rollover, subsequent to a pilot study)?

Please note that not only does this provide context for the reviewers at large but also impacts Ethics Officer assignment for continuity with a particular program of research.

- Yes – This study relates to a previously approved study at this institution
- Yes – This study relates to a study currently under Western’s REB review, but has not yet been approved
- No - This study does not relate to a previous study at this institution
1.10 *Upload the protocol/research plan for this study. NOTE: ALL HSREB submissions require a protocol/research plan:

<table>
<thead>
<tr>
<th>Type</th>
<th>Document Name</th>
<th>File Name</th>
<th>Version Date</th>
<th>Version</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protocol</td>
<td>REB Research Plan</td>
<td>REB Research Plan.pdf</td>
<td>09/Feb/2021</td>
<td>v4.0</td>
<td>187.3 KB</td>
</tr>
</tbody>
</table>

Note that your document name will appear on the approval notices. Ensure you name your document something that reflects what the document is (e.g., debriefing script, date). Avoid using slang, student names, etc. Upload only the clean version here (i.e., not the tracked copy). Do not include “clean” in the document name.

1.11 *Is this an Investigator-initiated study?*

- ☐ Yes
- ☑ No

1.12 *Who is the Study Sponsor?*

- ☐ Industry Sponsored
- ☐ External Non-Profit
- ☐ External PI (outside of Western)
- ☐ Local PI (Western-affiliated team member other than PI on this REB application)
- ☑ Self (PI on this REB application)

1.13 *Is this primarily a student project?*

- ☐ No
- ☐ Yes - Resident/Fellow
- ☐ Yes - MD
- ☐ Yes - Post-doctoral Fellow
- ☐ Yes - PhD
- ☐ Yes - Masters
- ☐ Yes - Undergraduate
- ☐ Yes - Other
1.14

*Has the study undergone a formal scientific or peer review (i.e., internal peer review or external review (e.g., CIHR, NSERC, NIH, etc.))?

- Yes
- No

1.15

*Has the study been reviewed and approved by another REB in Canada?

- Yes
- No
- pending

1.16

*Has the study been rejected by any other REB?

- Yes
- No

1.17

*Is this research study supported/funded by the United States federal government or regulated by the FDA (Food and Drug Administration)?

- Yes
- No

1.18

*Is this a multi-centre study?

- Yes
- No
1.21

1.21 *Is there an external third party (Coordinating or Contract Research Organization) overseeing the study?

- Yes
- No

1.22

1.22 *Are there any associated sub-studies or companion studies?

- Yes
- No

1.23

1.23 *Indicate how the results will be communicated to participants and other stakeholders (e.g.; advocacy groups, scientific community).

*To Participants:

- Debriefing Script
- Group debriefing
- End of study letter
- Publication(s)
- Other
- No Plan

*To Other Stakeholders:

- Presentation(s)
- Publication
- Other
- No plan

1.24
1.24 *Provide a brief lay/non-scientific summary of the study (max 250 words)

This study aims to investigate the prediction of treatment outcomes of cancer patients that have developed brain metastases. The treatment type being investigated is the treatment of brain metastases with stereotactic radiosurgery (SRS) or stereotactic radiotherapy (SRT). This study will begin by gathering data retrospectively from cancer patients treated with SRS/SRT for brain metastases. Importantly, this data will include imaging and clinical data available before treatment, as well as treatment outcome metrics gathered after treatment. With this data set, machine learning techniques will be used to investigate if the data available before treatment is predictive of what the treatment outcomes will be. If the machine learning techniques provide strong predictions of treatment outcomes, this study will be a step towards providing clinicians with improved decision support tools when deciding to use SRS/SRT treatments for brain metastases.

2.1

2.1 *Does this retrospective study include the collection of (select all that apply):

- Chart/Record collection
- Biological Specimens
- Registry data
- Existing research dataset
- Bioarchaeological Human Remains

*How many participants charts/records will be accessed? 350

*Describe Other:

Medical imaging and radiotherapy treatment planning electronic files (DICOM format)

2.2

2.2 *Provide the Retrospective start date for study data and/or biological sample collection. NOTE: To qualify as retrospective, study data and/or biological sample(s) must have been collected prior to your initial submission.

January 1, 2014

2.3
2.3 *Provide the Retrospective end date for study data and/or biological sample collection. NOTE: To qualify as retrospective, study data and/or biological sample(s) must have been collected prior to your initial submission.

January 12, 2021

2.5

2.5 *What are the study hypotheses or research question(s) or purpose of this study?

This study aims to answer the research question of if the clinical and imaging data available before the treatment of brain metastasis with SRS/SRT is able to accurately predict:
1. Overall survival of the patient
2. Intracranial local recurrence
3. Intracranial distant progression
4. Time to salvage treatment
5. Necrosis occurrence

2.6

2.6 *What is the rationale for this study (why is it being done)? In your response ensure to include relevant background information from previous studies that have been done. Cite references where appropriate and add LIST as a separate attachment (do not include within your response).

To provide beneficial personalized treatment for brain metastasis patients treated with stereotactic radiosurgery (SRS) or stereotactic radiotherapy (SRT), it is critical for treatment decisions to be made with as much relevant data as is available. Along with readily available clinical data points, such as primary cancer location and histopathology, brain metastasis patients often have extensive medical imaging performed, including x-ray computed tomography (CT) and magnetic resonance imaging (MRI). Currently in the clinical setting, medical imaging reliably provides brain metastasis location and size, but is not used as a predictor of treatment outcomes. Some studies in the literature have shown that having clinicians qualitatively categorize brain metastases using MRI data leads to statistically significant differences in SRS treatment response between the categories [1-3]. While these studies demonstrate a promising link between metastasis appearance in medical imaging and SRS treatment outcomes, the technique involved is limited by inter-observer variability and from the imaging data being reduced to a single categorical interpretation.

An alternative technique is to use machine learning algorithms that take clinical and imaging data in as input and then produce a predicted treatment outcome. These predictions are based upon training the algorithm with a database of patients already treated and observed post-treatment. This computerized technique eliminates inter-observer variability and allows imaging data to be quantitatively analyzed. Some preliminary results from using this machine learning approach have been published in the literature [4-6]. While these studies show promising results, they have small study populations (<100 patients), are not externally validated using data from another clinical centre, use only one or two imaging modalities, and do not examine differences between SRS and SRT or results for patients with a single primary cancer type.

The rationale beyond this study is to assemble a database to provide the most robust machine learning study to date on brain metastases treated with SRS or SRT. This study’s database will produce the largest population of patients treated with SRS and the largest population treated with SRT. Furthermore, the analysis performed on the SRS population will be able to be externally validated by data available through a collaboration with the Amsterdam University Medical Center in The Netherlands. The scope of the imaging data being collected for this study will also be the most extensive to date, incorporating x-ray CT data and data from four MRI modalities. The size and scope of the study database will also allow for direct comparison between patients treated with SRS and SRT at the same centre with identical clinical and imaging data collected. Lastly, it is anticipated the study database will be large enough that a cohort of brain metastasis patients with only primary lung cancer will be produced to study independently. This will represent the first machine learning analysis of brain metastasis patients treated with SRS/SRT of a single primary cancer type. It is hypothesized this focused analysis will allow for more accurate treatment outcome prediction for patients with primary lung cancer, which represent the majority of brain metastasis SRS/SRT patients.

Upload a list of references used in your rationale above (if applicable):
2.7

*Provide a brief summary of the study design type and methodology being employed in this study.

DO NOT include Information about objectives, inclusion/exclusion criteria, study procedures, sample size calculations and data analysis here.

This study is a machine learning study, and so the study design consists of first constructing a database of patient data and then using this database to train and test machine learning models. The database consists of two parts: the inputs and outputs for the machine learning models. The inputs consist of the data available pre-treatment for each patient, including clinical and imaging data points. For each patient, a set of outputs are also collected, consisting of the treatment outcomes outlined in Sections 2.24 and 2.25. With the inputs and outputs collected for all patients, a portion of the patients in the database can be selected to train a machine learning model. This training is completed by providing the model both the inputs and outputs for the selected patients, with the model being trained to find the mathematical relationship between the inputs and outputs. The accuracy of the model can then be tested using the patients in the database that were not selected for training. The model is tested by providing only the inputs from the patients selected for testing and recording the predicted outputs from the model. These predicted outputs are then compared to the known outputs from the database for each patient to assess the accuracy of the model. The more specific machine learning techniques that will be applied in this study are provided in Section 2.21.

Upload a flow diagram (if applicable):

2.15

*Does this study include a non-patient group (e.g., caregiver, student, employee, etc.)-SEE HELP TEXT?

☐ Yes
☐ No

2.21

*Is this a collaborative community-based project?

☐ Yes
☐ No

2.22
2.22 *Indicate your data collection tools/forms by selecting the relevant option(s) below:

- Paper Survey(s)/Questionnaire(s)
- Online Survey(s)/Questionnaire(s)
- Interview Guide(s)
- Focus Group Guide(s)
- Non-Participant Observation Guide(s)
- Participant Observation Guide(s)
- Case Report Form(s)
- Other (e.g., visual stimuli, participant diary, data collection forms, etc.)

*Upload "Other" instrument(s) that will be used during this study:

Note that your document name will appear on the approval notices. Ensure you name your document something that reflects what the document is (e.g., debriefing script, date). Avoid using slang, student names, etc. Upload only the clean version here (i.e., not the tracked copy). Do not include “clean” in the document name. Avoid using slang, student names, etc.

*Upload "Other" instrument(s) that will be used during this study:

<table>
<thead>
<tr>
<th>Documents</th>
<th>Instrument Name</th>
<th>Version Date</th>
<th>Version</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>REB REDCap Instruments</td>
<td>REB REDCap Instruments.pdf</td>
<td>09/Feb/2021</td>
<td>v2.0</td>
<td>218.4 KB</td>
</tr>
</tbody>
</table>

*Describe "Other" instrument(s) and how they will be used in this study:

Lawson REDCap database instruments
Exporting of medical imaging and radiotherapy treatment planning electronic files (DICOM files) to a research PACS (Picture archiving and communication system) server

2.23

2.23 *Will any technological tool(s)/platform(s)/software/device(s) be used (beyond an institutional network or hard drive) throughout the project (e.g., data collection, analysis, transfer, storage, etc.)?

- Yes
- No

*Specify the tool(s)/platform(s)/software(s)/device(s):

Lawson REDCap Database
Orthanc PACS Server

*Has the tool(s)/platform(s)/software(s)/device(s) received any of the following:

- Technology Risk Assessment by Western’s Technology Risk Assessment Committee (TRAC)?
- Authorized Technology Review at LHSC or SJHC?
- Review by a local institutional privacy office (e.g., Western’s privacy office, hospital privacy, or other collaborating institution privacy office)?

- Yes
- No
- unsure
Upload any relevant reports and/or other review/approval documentation (For Lawson-hosted REDCap and WebEx or Western-hosted’s Qualtrics, Zoom, Office 365 and OWL, no documentation is required):

<table>
<thead>
<tr>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type</strong></td>
</tr>
<tr>
<td>Technology Review document</td>
</tr>
<tr>
<td>Technology Review document</td>
</tr>
</tbody>
</table>

*Specify what information will be collected through and/or entered into the tool(s)/platform(s)/software(s)/device(s) and for what purpose:

*If personal identifiers will be shared using the technology, please list them and ensure consistency with Q13.11

All patient data except for medical imaging data will be collected and stored within the Lawson REDCap database system for the purposes of this study.

All patient imaging data will be collected and stored within the Orthanc PACS server system on a research server within the LHSC computer network for the purposes of this study.

*Specify who will have access to this information and for what purpose (incl. third party vendors and any future use, if applicable):

Only the PI and study team will have access to the information on these systems.

The Lawson REDCap system administrators will have access to the information stored with REDCap database as required for technical support and maintenance.

*Specify how long will the information be accessible in the tool(s)/platform(s)/software/device(s):

The data will be retained for 15 years at which point it will deleted from the Lawson REDCap and Orthanc PACS systems.

*Specify how the information will be removed from the tool(s)/platform(s)/software/device(s):

Data will be deleted from the Lawson REDCap system via the REDCap administrative staff.

Data will be deleted from the Orthanc PACS system by deleting the associated medical image database created by the Orthanc PACS system.

2.26

2.26 *Is the sample size justified in the study protocol/research plan or sponsor protocol?

☐ Yes

☐ No
2.27

2.27 Describe the method(s) for data analysis.

For each of the endpoints/objectives of the study, a variety of machine learning experiments will be performed, including the following techniques:
• Radiomics feature extraction from imaging data
• Feature selection algorithms
• Conventional machine learning models such as support vector machines and random decision forests
• Deep learning machine learning models such as conventional neural networks and convolutional neural networks

2.28

2.28 *Provide the inclusion criteria:

• Patients with a primary cancer that is not located within the brain or central nervous system
• Patients with one or more brain metastases (pathologic confirmation not required, but probable primary cancer source is required)
• Patient brain metastases treated with either SRS or SRT

2.29

2.29 *Provide the exclusion criteria.

• Patient SRS or SRT treatment occurred before January 1, 2014
• Patient did not receive MRI and CT brain imaging before their SRS/SRT treatment
• Patient date of death is unknown

2.30

2.30 *What is/are the primary objective(s) of the study and briefly describe how it/they will be measured. NOTE: For qualitative research studies-if this is not applicable indicate "NA"

The primary objective of the study is to use clinical and imaging data available before the treatment of brain metastasis patients with SRS/SRT to accurately predict overall survival of the patient.
2.31 What is/are the secondary objective(s) (if applicable) of the study and briefly describe how it/they will be measured.

The secondary objectives of the study are to use clinical and imaging data available before the treatment of brain metastasis patients with SRS/SRT to accurately predict the following endpoints:
1. Intracranial local recurrence
2. Intracranial distant progression
3. Time to salvage treatment
4. Radio-necrosis occurrence

11.1 Describe any direct benefits to the study participants. If there are no direct benefits to the participants themselves, please state as such:

As this study is a retrospective study including the overall survival of brain metastasis patients, the vast majority of included patients will be deceased, and so will not directly benefit.

11.2 What is the overall anticipated public and scientific benefits of the study?

This study will provide a scientific benefit by providing insight between the connection of clinical and imaging data known before the treatment of brain metastasis patients with SRS/SRT and the outcomes for the patient. If accurate predictions are possible, this study will explain which data points are most relevant to the predictions, and so important relationships between biological processes and treatment outcomes may be discovered. This study will also provide a foundation and procedure for future studies, encouraging similar studies to be performed that could be validated against this study's dataset. The techniques developed and validated in this study could be beneficial for similar research in other clinical scenarios beyond the area of brain metastasis patients treated with SRS/SRT. Lastly, some of the data collected for this study will be used with a similar retrospective data set available from another institution to perform external validation. Performing external validation is critical in machine learning studies to truly evaluate the effectiveness of a proposed technique, and so this study will provide a rare opportunity to perform such a validation.

The public benefits of this study will occur when the results of this study allow for the creation of a system that can provide clinicians with enhanced decision support during treatment planning. If implemented, the clinician would be provided with information that would aid them in deciding to use SRS or SRT to treat brain metastases and if so, how to best plan and deliver the treatment. This would allow for treatments to be optimally customized for patients, providing treatments that are the most likely to ensure adequate treatment of the brain metastases, while minimizing damage to healthy areas of the brain.
12.1 *Will Personal Information (PI) and/or Personal Health Information (PHI) be used to identify potential participants (pre-screening)?

- [ ] Yes
- [X] No

*Who is accessing the PI and/or PHI for pre-screening and under whose authorization?

David DeVries under the authorization of Dr. George Hajdok, Dr. Joanna Laba and Dr. Andrew Leung

*Describe what PI and/or PHI will be used or accessed to identify potential participants?

- Hospital patient identification number (LRCP and LHSC PINs)
- Sex
- Date of birth
- Date of death
- Radiation treatment dates and parameters
- Primary cancer diagnosis
- Brain metastasis imaging types and date
- Pathology reports
- Systemic therapy reports

12.2 *Is a waiver of the requirement to obtain informed consent being requested for any aspect of this study (If you are obtaining consent for part of the study and requesting a waiver for another aspect of the study select both Yes AND No)?

- [X] Yes I am requesting a waiver of consent
- [ ] No I am not requesting a waiver of consent

*Specify for what type of data the waiver is being requested?

- Prospective data collection
- [X] Secondary use of identifiable information
- [ ] Secondary use of non-identifiable information

*In accordance with Tri-Council Policy Statement 2, Article 5.5A, please confirm that ALL of the following conditions apply:

- Identifiable information is essential to the research
- The use of identifiable information without the participants’ consent is unlikely to adversely affect the welfare of individuals to whom the information relates
- The researchers will take appropriate measures to protect the privacy of individuals, and to safeguard the identifiable information
- The researchers will comply with any known preferences previously expressed by individuals about any use of their information
- It is impossible or impracticable to seek consent from individuals to whom the information relates

The researchers have obtained any other necessary permission for secondary use of information for research purposes

- [X] I confirm

*Explain why identifiable information is essential to the research:

Hospital PINs will be used to identify patients. PINs will only be collected on the master list.
*Explain why not obtaining consent is unlikely to adversely affect the welfare of individuals to whom the information relates:

As this is a retrospective chart review and the data from these patients has already been collected and all the medical procedures have been done, not obtaining consent will not affect the welfare of the participants.

*Explain what measures will be taken to protect the privacy of individuals, and to safeguard the identifiable information:

The master linking list which contains the hospital PINs will be stored separately from the study data and encrypted. Study data will be stored on the REDCap database system and research PACS server and will be password protected. The REDCap database is a Lawson-approved Electronic Data Capture Platform. The research PACS server has been approved by Western’s Cyber Security and Business Services.

*Explain why it is impossible or impracticable to obtain consent:

It is impossible and/or impractical to conduct this research without prior consent as many of our patients may not be followed up in London (i.e. followed up in their hometown) as well as it is likely patients may have passed away due to the disease being studied.

13.1 *(For patient orientated research studies.) Do you plan now or in the future to link your study data to the large healthcare databases held at the Institute for Clinical Evaluative Sciences (ICES)? For example, this would allow you to follow patients passively life-long, determine their healthcare costs, assess how similar your patients are compared to Ontario citizens, and help identify control groups.

- Yes
- No
- N/A

13.2 *Are you collecting personal identifiers for this study?

- Yes
- No

13.3
13.3 *Identify any personal identifiers collected for this study. Select all that apply.

- Full Name
- Initials
- Ontario Health Card Number
- Address
- Full Postal Code
- Partial Postal Code
- Telephone Number
- Email Address
- Family Physican or other care provider names
- Full Date of Birth
- Partial Date of Birth
- Full Date of Death
- Partial Date of Death
- Sex
- Gender
- Age
- Medical Device Identifier
- Hospital Patient Identification Number (PIN)
- Full Face Photograph
- Voice/Audio Recording
- Race
- Ethnicity
- Other

*Explain and justify full date of birth and if it will be stored on paper or electronically

The date of birth is required to calculate the patient's age at time of treatment. The date of birth data will be stored electronically.

*Explain and justify full date of death and if it will be stored on paper or electronically

The date of death is required to calculate the overall survival for a patient (date of treatment to date of death). The date of death data will be stored electronically.

*Explain and justify sex and if it will be stored on paper or electronically

The patient's sex will be recorded as piece of clinical data used in the statistical analysis of this study. The sex data will be stored electronically.
*Explain and justify hospital PIN and if it will be stored on paper or electronically

The hospital PIN (both LHSC PIN and LRCP PIN) will be recorded to allow for the look-up of patient charts and imaging data, as required by this study. The hospital PIN data will be stored electronically in the master list.

13.4

13.4 *Will there be a master list linking identifiers/identifiable information (e.g., name, contact information) to the unique participant code (e.g., study number, pseudonym)?

- Yes
- No

*Who will have access to the master list?

The PI and study team will have access to the master list

13.5

13.5 *Where will information collected as part of this study be stored (applies to both paper copy and electronic copy)? (select all that apply)

- University or Hospital network drive
- University or Hospital local hard-drive
- Office/Lab of PI or Research team member on Institutional Property
- Laptop
- Memory Stick
- Cloud Storage
- Off-site
- Other

*Specify Other:

Lawson REDCap Database

13.6
13.6 *Indicate the measures in place to protect the confidentiality and security of any study data including Personal Information (PI) or Personal Health Information (PHI) that is accessed, collected and used (select all that apply):

- Access to study data and/or medical records will be limited to authorized personnel
- Access to electronic data will, at least, be password protected (if not password protected AND encrypted)
- Electronic data will be stored on a Western, hospital or other institutional server with firewalls and other security and back-up measures in place
- Study data stored on external hard drive, laptop(s) and/or portable device(s) will be encrypted
- Paper copies of study data will be stored in locked filing cabinets in a secure location
- A master log with identifiers will be stored separately from the study data
- Other
*Specify Other

The electronic data that is stored on Ward Lab computers and servers will be password protected, but not encrypted. These computers and servers are physically secured within LHSC premises, preventing unauthorized physical access to computer hard-drives.

"Study data stored on external hard drive, laptop(s) and/or portable device(s) will be encrypted" is left unchecked since no data will be placed on external hard drive, laptop(s) and/or portable device(s)

"Paper copies of study data will be stored in locked filing cabinets in a secure location" is left unchecked since no paper copies of study data will be produced

13.7

13.7 Describe where study data/database, source data (including completed surveys), and Letters of Information and Consent, whether electronic or paper, will be kept:

Clinical data will be collected and stored within Lawson's REDCap Database system and then exported for analysis without patient identifiers onto a computer hard drive within the LRCP/Victoria Hospital computer network.

Imaging data will be stored on a server hard drive within the LRCP/Victoria Hospital computer network.

13.10

13.10 *Are you transporting materials (paper, devices and/or media) that include Personal Information (PI) and/or Personal Health Information (PHI) between sites? (See Confidentiality and Data Security guidelines)

- Yes
- No

13.11

13.11 *Will you be sending/sharing data off-site for this study?

- Yes
- No
13.12 *Who will have access to the identifiable data?

The PI, study team, Western’s Research Ethics Board and Lawson’s Quality Assurance and Education Team will have access to the identifiable data.

13.13 *How long will you retain identifiable data?

- 7 years as per UWO policy
- 15 years as per Lawson policy
- 25 years as per Health Canada policy
- Other

13.14 *How will you destroy the identifiable data after this period (if applicable)?

After 15 years, all identifiable data will be permanently destroyed/deleted as per Lawson data destruction policies.

13.15 *Will you link the locally collected data with any other datasets, databases or registries (e.g., health registries, Statistics Canada)?

- Yes
- No

13.16 *Is the purpose of this study to establish a registry/database?

- Yes
- No
*Will Personal Identifiers (PI) be stored in the registry/database?

- Yes

*What identifiers will be stored?

- Full date of birth
- Full date of death
- Sex

*Who maintains the registry/database?

Dr. Aaron Ward

*Where is the registry/database located?

- Lawson REDCap database system
- Research PACS server within the LHSC computer network (local hard drive)

13.17

13.17 *Indicate the extent the study participant is able to withdraw their study data from the research study and any limitations on the withdrawal

N/A

14.1

14.1 *Is this study funded?

- Yes
- No

14.2

14.2 *How is the study funded?

- Industry
- Internal Grant (departmental/faculty, VP, IRF/SRF, etc.)
- External Grant (Tri-Council (e.g., CIHR, SSHRC, NSERC, NCE), government, charitable foundation, etc.)
- Other

*Specify Internal Funder(s):
14.3

14.3 *Are there any (or will there be) research funds held in an account at Western or Lawson?

☒ Lawson
☒ Western University
☒ No

14.4

14.4 *What is the status of funding from this source?

☒ Obtained
☒ Awarded but not received

14.5

14.5 Indicate what compensation, if any, will be provided to participants and include a justification for compensation. If this question does not apply, indicate "not applicable" or "N/A".

N/A
14.10 Attach an itemized study budget. The budget should reflect all costs to complete the study (e.g., REB fees for industry sponsored studies, database extraction, student payments, participant reimbursements, etc.).

<table>
<thead>
<tr>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>Study budget</td>
</tr>
</tbody>
</table>

16.1

16.1 *Will the PI or Co-Investigator(s) or anyone connected to them through their interpersonal relationship (including their partners, family members, or their former or current professional associates) receive any personal financial benefit in connection with this study?

- [ ] Yes
- [ ] No

16.2

16.2 *Will the PI or Co-Investigator(s) or anyone connected to them through their interpersonal relationships (including their family members, friends, or their former or current professional associates) receive any personal (financial or otherwise) benefits including patent or intellectual property rights, royalty income, employment, share ownership, stock options, etc?

- [ ] Yes
- [ ] No

16.3

16.3 *Is the PI or Co-Investigator(s) aware of any other community relationships, academic interests, financial partnerships, or economic interests (e.g., spin-off companies in which researchers have stakes or private contract research outside of the academic realm) or any other incentives that may compromise their integrity, independence or ethical duties in the conduct of the research?

- [ ] Yes
- [ ] No

16.4
16.4 * Is the PI to Co-Investigator(s) aware of any institutional conflicts of interest (financial or non-financial) that may have an impact on the research?

☐ Yes
☐ No

16.5 * Does the PI or Co-Investigator(s) or anyone connected to them through their interpersonal relationships (including their family members, friends, or their former or current professional associates) have any proprietary interest in the product under study or in any entity that is sponsoring or otherwise supporting the conduct of the study?

☐ Yes
☐ No

16.6 * Will or does the PI or Co-Investigator(s) or anyone connected to them through their interpersonal relationships (including their family members, friends, or their former or current professional associates) have any association or connection with an entity that is sponsoring or otherwise interested in the outcome of the study? (e.g., consultant, advisor, board member, employee, director, etc.)

☐ Yes
☐ No

16.7 * Are you or your institution the sponsor of this investigator-initiated/sponsored study?

☐ Yes
☐ No

*Describe any real, potential, or perceived conflict of interest

None

*Describe the proposed management plan to mitigate the conflict of interest:

N/A
16.8 * Are there any other real, potential or perceived conflict of interest to declare to the REB?

☐ Yes
☐ No

18.1

Although the REB requests that you delete previous version documents and replace them with updated, revised documents please DO NOT delete any of the response letters. They can all stay attached. Ensure you have different version date and/or number for each response letter.

18.1 * Upload the Response Letter, listing all REB recommendations/questions/comments and an explicit response to each:

<table>
<thead>
<tr>
<th>Documents</th>
<th>Type</th>
<th>Document Name</th>
<th>File Name</th>
<th>Version Date</th>
<th>Version</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>REB Response Letter</td>
<td>REB Response Letter</td>
<td>REB Response Letter.pdf</td>
<td>09/Feb/2021</td>
<td>v1</td>
<td>150.2 KB</td>
<td></td>
</tr>
</tbody>
</table>

18.2

18.2 If changes have been made to a previously submitted consent/assent form(s) at the request of the REB, upload track-changes versions of all proposed consent and/or assent form (e.g. screening, main, optional), if applicable:

18.3

18.3 If changes have been made to a previously submitted study instruments/stimuli (e.g., survey, questionnaire, interview guide, focus group guide, observation guide, etc.) at the request of the REB, upload the track-changes version(s):

<table>
<thead>
<tr>
<th>Documents</th>
<th>Type</th>
<th>Document Name</th>
<th>File Name</th>
<th>Version Date</th>
<th>Version</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tracked Changes Document</td>
<td>REB REDCap Instruments</td>
<td>REB REDCap Instruments.docx</td>
<td>09/Feb/2021</td>
<td>v2.0</td>
<td>499.8 KB</td>
<td></td>
</tr>
<tr>
<td>Tracked Changes Document</td>
<td>REB Research Plan</td>
<td>REB Research Plan.docx</td>
<td>09/Feb/2021</td>
<td>v4.0</td>
<td>31.6 KB</td>
<td></td>
</tr>
</tbody>
</table>

18.4

18.4 If changes have been made to a previously submitted protocol, research plan, research outline please upload the track-changes version(s):
18.5

Please provide any additional comments for the REB to consider (if applicable): 


19.1

*Confirm that all study team members have received a certificate for completion of human research ethics training through one of the following (select ALL that apply):

- Tri-Council Policy Statement (TCPS2) Core Tutorial
- Collaborative Institutional Training Initiative (CITI Program)
- Other
The Principal Investigator may choose to sign off electronically on all re-submissions (i.e., response to REB recommendations) or he/she may delegate this task to another qualified individual. NOTE: The PI is still fully responsibility for the scientific and ethical conduct of the study at this institution.

- I attest that this application as submitted is in compliance with the TCPS2 (2nd edition of Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans); AND, if applicable, with the provisions of the Ontario Personal Health Information Protection Act and its applicable Regulations; AND, with all other applicable laws, regulations or guidelines (e.g., if applicable, Food and Drugs Act and applicable Regulations; International Conference on Harmonization Guidance E6: Good Clinical Practice);
- I attest that, to the best of my knowledge, the information in this application is complete, current and accurate;
- I attest that this application contains the current and complete protocol, including, if applicable, any sub-studies;
- I acknowledge that I am responsible for promptly reporting any of the following to the REB:
  - modifications or amendments, such as changes in PI, changes in Co-investigator (if applicable), specific required changes to the Letter of Information/consent form, etc.;
  - all local reportable events that meet the REB reporting criteria, including but not limited to local unexpected, serious adverse events (SAEs), privacy breaches, protocol deviations and any new information that may adversely affect the safety of the participants or significantly affect the conduct of the study;
  - progress report (renewal/continuing review form), annually or as often as requested by the REB;
  - study completion or termination;
- I certify that REB approval and all external and local institutional approvals will be obtained before the study will commence;
- I certify that the research team will adhere to the protocol and consent form as approved by the REB unless to eliminate an immediate safety hazard to participants and in accordance with any conditions placed on the REB approval;
- I certify that all information provided in this application represents an accurate description of the conduct of the study.
- I have made efforts to ensure that the research intent, purpose, and impact of this study will be free from bias or discrimination in accordance with the Canadian Charter of Rights and Freedoms.

Privacy and Security Acknowledgement:

- On behalf of all members of my research team, I recognize the importance of maintaining the confidentiality of personal health information (PHI)/Personal Information (PI) and the privacy of individuals with respect to that information;
- I will ensure that the PHI/PI is used only as necessary, to fulfill the specific study objectives and related study questions described in the application approved by the REB. This includes all conditions and restrictions imposed by the REB and the institution in which the study is being conducted, governing the use, security, disclosure, return or disposal of the study participants’ personal information;
- I agree to take any further steps required by the REB or the institution to ensure that the confidentiality and security of the PHI/PI is maintained in accordance with the Personal Health Information Protection Act (PHIPA) and/or Freedom of Information Protection of Privacy Act (FIPPA), its accompanying regulations, and the Tri-Council Policy Statement

Signed: This form was signed by Aaron Ward on 09/Feb/2021 17:45

Figure I.2: Study application to the Western University Health Sciences Research Ethics Board (30 pages). The referenced “Research Plan” document is provided in figure I.3. All other referenced documents are not included as they contain details that are not relevant in the context of this thesis. Sensitive information redacted.
Retrospective Outcome Analysis of Brain Metastasis Patients Treated with Stereotactic Radiosurgery/Radiotherapy (SRS/SRT)

David DeVries\textsuperscript{1,2}, Andrew Leung\textsuperscript{3,4}, Joanna Laba\textsuperscript{5,6}, George Hajdok\textsuperscript{1,5}, Aaron Ward\textsuperscript{1,2,6}

\textsuperscript{1}Department of Medical Biophysics, Western University
\textsuperscript{2}Gerald C. Baines Centre, London Regional Cancer Program
\textsuperscript{3}Department of Medical Imaging, Western University
\textsuperscript{4}Department of Medical Imaging, London Health Sciences Centre
\textsuperscript{5}Department of Radiation Oncology, London Regional Cancer Program
\textsuperscript{6}Department of Oncology, Western University

Introduction:
Various cancers are well known to spread, or metastasize, to the brain to form lesions that are known as brain metastases. Brain metastases are often treated with radiation in the form of stereotactic radiosurgery (SRS) or stereotactic radiotherapy (SRT), where high doses of radiation are delivered to the metastases in either 1-3 sessions (SRS) or 5 sessions (SRT). When choosing to employ SRS/SRT, and when designing SRS/SRT treatment plans, clinicians are faced with multiple decisions. This study aims to develop a robust machine learning model to predict various outcomes of an SRS/SRT treatment based on pre-treatment input data, with the goal of providing clinicians with more information during treatment decision making. Previous literature has shown statistical correlations between the qualitative appearance of brain metastasis in medical imaging and treatment outcomes [1-3], with other studies showing promising initial machine learning results based on clinical and imaging data [4-6]. With this study, we hope to provide the most comprehensive machine learning study in the field, studying both SRS and SRT populations from the same centre, as well as incorporating the most diverse set of readily available medical imaging types. We have also identified a large subset of patients specifically with cancer that started in the lungs (primary lung cancer) before metastasizing. We intend to produce novel analysis that investigates only this more homogeneous cohort as opposed to the typical variety of patient primary cancers used in previous studies. It is hypothesized that this analysis of only lung cancer patients will produce more accurate outcome predictions for these patients that represent the majority of brain metastasis patients treated with SRS/SRT.
Methods:

Summary:
The study design will consist of a standard machine learning design, in which a database of patients is created. For each patient in the database, a set of data points available pre-treatment, or “inputs”, will be collected, both from clinical and imaging records (see “Variables to be Collected” below). Each patient will also have a series of known post-treatment outcomes, or “outputs”, collected from clinical records. The patients will then be split into two groups, with one group used to “train” a machine learning model. This training is done by providing the model with both the inputs and outputs for the patient, after which the model algorithmically uncovers possible relationships between the two. The trained model is then “tested” with the second group of patients by providing the model only with the pre-treatment inputs from this second group. The model then produces a post-treatment output, or prediction, based on the input, which can then be validated against the known treatment outcomes for the patients. In this way the accuracy of the model can be evaluated for possible clinical use.

Sites:
The data for this study will be collected at the London Regional Cancer Program (Department of Radiation Oncology) and Victoria Hospital (Department of Medical Imaging) sites of LHSC.

Number of Participants:
Approximately 350 patients have been identified to be part of the study population. For this machine learning study, a strict sample size calculation does not apply. In general, the larger the study population used for a machine learning study, the more representative of the true population the study population will be, leading to more accurate and generalizable results. For this study, the 350 patient figure provided represents an approximate number of all the brain metastasis patients treated with SRS or SRT at the London Regional Cancer Program that meet the inclusion and exclusion criteria provided below. This maximum study population also provides a large enough sample size that patients treated with SRS and SRT can be analyzed separately. It is also anticipated that enough brain metastasis patients with only primary lung cancer will be accrued in the study population to allow for these patients to also be analyzed separately.

Retrospective Start/End Dates:
January 1, 2014 – January 12, 2021
**Patient Inclusion/Exclusion Criteria:**

**Inclusion Criteria:**
- Patients with a primary cancer that is not located within the brain or central nervous system
- Patients with one or more brain metastases (pathologic confirmation not required, but probable primary cancer source is required)
- Patient brain metastases treated with either SRS or SRT

**Exclusion Criteria:**
- Patient SRS or SRT treatment occurred before January 1, 2014
- Patient did not receive MRI and CT brain imaging before their SRS/SRT treatment
- Patient date of death is unknown

**Variables to be Collected:**
The following variables will be collected for each patient in the study. All non-imaging variables will be captured within Lawson’s REDCap database system. Imaging data will be stored on a research PACS (Picture Archiving and Communication System) computer server located within LHSC’s computer network.

**Personal Identifiers:**
- Hospital patient identification number (LRCP and LHSC PINs)
- Patient date of birth
- Patient date of death
- Patient sex

**Non-Identifying Information**
- Brain imaging data
  - X-ray CT scans (*pre-treatment*)
  - Multi-sequence MRI scans (*pre and post-treatment*)
  - Brain radiation therapy contouring data (*pre-treatment*)
  - Brain radiation therapy dose planning volumes (*pre-treatment*)
- Lung imaging data (*if applicable*)
  - X-ray CT scans (*pre-treatment*)
  - PET scans (*pre-treatment*)
  - Lung radiation therapy contouring data (*pre-treatment*)
- Lung radiation therapy dose planning volumes (pre-treatment)

- Primary cancer diagnosis
  - Date
  - Site
  - Stage
  - Histological classification
  - Pathology marker status

- Systemic therapy (possibly multiple per patient)
  - Start date
  - Therapy agent delivered

- Brain radiation therapy (possibly multiple per patient)
  - Date delivered
  - Treatment intent
  - Lesion intact vs. post-operative surgical bed
  - Prescribed dose
  - Number of fractions

- Lung radiation therapy (if applicable; possibly multiple per patient)
  - Date delivered
  - Target
  - Treatment intent
  - Prescribed dose
  - Number of fractions

- Brain radiology follow-up (possibly multiple per patient)
  - Scan date
  - Lesion treatment response score (RANO-BM)
  - Presence of pseudo-progression
  - Presence of radio-necrosis
  - Locations of new lesions (if applicable)

- Salvage treatment (if applicable)
  - Treatment date
  - Treatment type
Objectives:
This study’s objective is to reliably predict the following post-treatment outcome endpoints for SRS and SRT patients:
1. Overall survival
2. Intracranial local recurrence
3. Intracranial distant progression
4. Time to salvage treatment
5. Radio-necrosis occurrence

Statistical Analysis:
For each of the endpoints above, a variety of machine learning experiments will be performed, including the following techniques:
- Radiomics feature extraction from imaging data
- Feature selection algorithms
- Conventional machine learning models such as support vector machines and random decision forests
- Deep learning machine learning models such as conventional neural networks and convolutional neural networks

Conclusion:
Through the collection of the above variables, and the analysis of them, this study aims to produce machine learning models that can reliably predict five specific outcomes when treating a brain metastasis patient with SRS/SRT. If successfully, this study will provide clinicians with a tool to provide enhanced information to them during treatment decision making. In particular, the information this tool could provide is designed to aid in deciding whether to treat a brain metastasis more aggressively or conservatively. Treating more aggressively can provide enhanced tumour control, while treating more conservatively can spare healthy tissues from adverse side-effects and provide the opportunity for re-treatment if required. With treatments
guided in this way, it is hoped that one day treatments can be optimally tailored to patients to offer them the best possible treatment outcomes.

References


Figure I.3: Research plan submitted to the Western University Health Sciences Research Ethics Board (6 pages).
Copyright permissions

Figure 1.6 and chapters 2, 3, and 4 are the reproduced materials presented in this thesis and therefore require copyright permissions from their original publishers.

Figure 1.6 was originally published in the journal *Cancers* from the Multidisciplinary Digital Publishing Institute (MDPI). As shown below, the figure can be reproduced under MDPI’s open-access policy and use of the Creative Commons BY Attribution 4.0 International License.

Chapter 2 was originally published in the journal *Scientific Reports*. As shown below, the chapter can be reproduced under *Scientific Reports*’s open-access policy and use of the Creative Commons BY Attribution 4.0 International License.

Chapter 3 was originally published in the journal *Neuro-Oncology Advances*. As shown below, the chapter can be reproduced under *Neuro-Oncology Advances*’s open-access policy and use of the Creative Commons BY Attribution 4.0 International License.

Chapter 4 was under review for publication in the journal *Scientific Reports* at the time of thesis submission, and so copyright permission from the journal is anticipated to be required. As shown below, the chapter can be reproduced under *Scientific Reports*’s open-access policy and use of the Creative Commons BY Attribution 4.0 International License.
Permissions

No special permission is required to reuse all or part of article published by MDPI, including figures and tables. For articles published under an open access Creative Common CC BY license, any part of the article may be reused without permission provided that the original article is clearly cited. Reuse of an article does not imply endorsement by the authors or MDPI.

MDPI’s open-access copyright permission policy. Taken from: https://www.mdpi.com/openaccess.

License agreement and author copyright

*Scientific Reports* does not require authors to assign copyright of their published original research papers to the journal. Articles are published under a [CC BY license](https://creativecommons.org/licenses/by/4.0/) (Creative Commons Attribution 4.0 International License). The CC BY License allows for maximum dissemination and re-use of open access materials and is preferred by many research funding bodies. Under this license, users are free to share (copy, distribute and transmit) and remix (adapt) the contribution including for commercial purposes, providing they attribute the contribution in the manner specified by the author or licensor ([read full legal code](https://creativecommons.org/licenses/by/4.0/legalcode)).

Visit our open research site for more information about [Creative Commons licensing](https://creativecommons.org/licenses/by/4.0/).

*Scientific Reports*’s open-access copyright permission policy. Taken from: https://www.nature.com/srep/journal-policies/editorial-policies#license-agreement.

*Neuro-Oncology Advances* is a fully open access journal, and all articles are published in the journal under an open access licence immediately upon publication. You will need to pay an open access charge to publish under an open access licence.

*Neuro-Oncology Advances*’s open-access copyright permission policy. Taken from: https://academic.oup.com/noa/pages/General_Instructions#License.
David A. DeVries, MSc

PhD Curriculum Vitæ

CAMPEP PhD Candidate

Revised: July 1, 2023

Education

Sep 2018 – July 2023  **PhD Medical Biophysics (CAMPEP)**
Western University, Schulich School of Medicine and Dentistry
Thesis: *Predicting brain metastasis response to stereotactic radiosurgery using magnetic resonance imaging radiomics and machine learning*
Supervisors: Dr. Aaron Ward and Dr. George Hajdok

Sep 2016 – Aug 2018  **MSc Physics**
Confferred Oct 2018  Queen’s University, Faculty of Arts and Science
Thesis: *Application and development of algebraic reconstruction algorithms for high dose rate brachytherapy gel dosimetry with optical computed tomography readout*
Supervisor: Dr. John Schreiner

Sep 2011 – Apr 2016  **BCS Hons. – Co-op, Joint Honours Computer Science & Physics**
Confferred Jun 2016  University of Waterloo, Faculty of Mathematics
Graduated with Distinction - Dean's Honours List

Academic & Employment History

Sep 2018 - Present  **Research student (as part of PhD program)**
Department of Medical Biophysics
Western University, Schulich School of Medicine and Dentistry
Baines Centre for Translational Cancer Research/
Lawson Health Research Institute
London Regional Cancer Program/London Health Sciences Centre

May 2019 – Aug 2019  **Quality assurance trainee (as part of PhD program)**
Department of Radiation Oncology
London Regional Cancer Program/London Health Sciences Centre

Teaching Experience

Jan 2020 – May 2020  **Course developer**
Course name: Biomedical Applications of Neural Networks
Western University, Department of Medical Biophysics
Awards & Honours

Jun 2023  
**Poster Competition Finalist**  
(also listed under “Local and Provincial Conference Abstracts”)  
London Health Research Day  
1 of top 8 poster presentations

Jun 2023  
**Poster Presentation Award**  
(also listed under “Local and Provincial Conference Abstracts”)  
London Oncology Research and Education Day  
1 of top 10 poster presentations

Jun 2023  
**Podium Presentation Certificate of Merit**  
(also listed under “Local and Provincial Conference Abstracts”)  
London Imaging Discovery Day

Feb 2022  
**Student Conference Support Winner**  
SPIE Medical Imaging Conference  
$700 travel award

Sep 2021  
**Alan C. Groom Award**  
Western University, Department of Medical Biophysics  
Awarded to the senior student that presents the most effective departmental seminar  
$1000

Jun 2021  
**Young Investigators Symposium Finalist**  
(also listed under “Conference Proceedings”)  
COMP Annual Scientific Meeting  
1 of 10 finalists chosen for podium presentations

Jun 2021  
**Best Podium Presentation**  
(also listed under “Local and Provincial Conference Abstracts”)  
London Oncology Research and Education Day

Mar 2021  
**Rising Stars Poster Award**  
(also listed under “Local and Provincial Conference Abstracts”)  
OICR Translational Research Conference  
$100 award

Sep 2019 – Dec 2022  
**Postgraduate Scholarship – Doctoral Program**  
Natural Sciences and Engineering Research Council of Canada  
$63,000 stipend award

Sep 2019  
**Ontario Graduate Scholarship**  
Western University  
$15,000 stipend award (declined due to awarding of another scholarship)

Sep 2018 – Aug 2019  
**Ontario Graduate Scholarship**  
Western University  
$15,000 stipend award

Sep 2018  
**Selected Vanier Scholarship Applicant**  
Western University  
Selected to put forward my application for the Vanier scholarship
Supervision & Mentorship

Research Trainee Supervision
2021 – 2022

**Primary supervisor (co-supervisor: Dr. Aaron Ward)**

Faris Khan (*4th* year undergraduate thesis project)

Western University, Department of Medical Biophysics

Role: provided extensive training, mentorship and scientific oversight for Faris’s thesis project on a novel method for brain metastasis prognostics

Student Mentorship

2020 – 2021

**1st year graduate student mentor**

Jermiah Joseph

Western University, Department of Medical Biophysics

Role: provided academic and personal mentorship during 1st year

Invited Talks & Media Appearances


Publications

**Articles in Peer-Reviewed Journals**


**Articles Under Review for Peer-Reviewed Journals**


**Conference Proceedings Papers**

*Awards earned for an abstract/presentation are listed here in bold as well as under “Awards & Honours”*


**National and International Conferences Abstracts**

*Notable presentations are highlighted in bold*


**Local and Provincial Conference Abstracts**

*Awards earned for an abstract/presentation are listed here in bold as well as under “Awards & Honours”*


Abbreviations Used

- AAPM American Association of Physicists in Medicine
- ASTRO American Society for Radiation Oncology
- CAMPEP Commission on Accreditation of Medical Physics Education Programs
- CAMRT Canadian Association of Medical Radiation Technologists
- CARO Canadian Association of Radiation Oncology
- COMP Canadian Organization of Medical Physicists
- OICR Ontario Institute for Cancer Research
- SPIE Society of Photo-Optical Instrumentation Engineers